Scientometrics
DOI 10.1007/s11192-016-2235-4

# Citation regression analysis of computer science publications in different ranking categories and subfields

**Yifan Qian[1,2] · Wenge Rong[1,3] · Nan Jiang[3] · Jie Tang[4] · Zhang Xiong[1,3]**

Received: 1 August 2016


**Abstract** A number of bibliometric studies point out that the role of conference publications in computer science differs from that in other traditional fields. Thus, it is interesting to identify the relative status of journal and conference publications in different subfields of computer science based on the citation rates categorised by the China Computer Federation (CCF) classifications and venue types. In this research, we construct a dataset containing over 100,000 papers recommended by the CCF catalogue and their citation information. We also investigate some other factors that often influence a paper's citation rate. An experimental study shows that the relative status of journals and conferences varies greatly in different subfields of computer science, and the impact of different publication levels varies according to the citation rate. We also verify that the classification of a publication, number of authors, maximum h-index of all authors of a paper, and average number of papers published by a publication have different effects on the citation rate, although the citation rate may have a different degree of correlation with these factors.

**Keywords** Citation rate · Computer science · Influence factor · Multiple regression

## Introduction

Conference papers in computer science have a higher status than in other disciplines (Freyne et al. 2010). Considering that the rate of technical innovation is fast and researchers need to report their results in a timely manner, conferences are more

✉ Wenge Rong
w.rong@buaa.edu.cn

[1] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

[2] Sino-French Engineer School, Beihang University, Beijing 100191, China

[3] School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[4] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Published online: 09 January 2017

suitable than journals. This is because the period of review for conferences is normally shorter than that for journals (Shamir 2010), which is effective for a young and fast-growing discipline (Fortnow 2009). As a result, although the purpose of conferences as a forum for scientists is to discuss their research ideas and share their work with each other, the computer science community also publishes peer-reviewed papers as conference proceedings. The vast majority of peer-reviewed publications are communicated in the form of conference papers, and conference proceedings have become the primary channel of research communication in computer science. However, in most other scientific disciplines, research results are reported in the form of peer-reviewed papers published in journals (Vardi 2009).

To better understand the importance of journals and conferences in the area of computer science, several different researchers and organizations have tried to rank journals and/or conferences according to their own experience and understanding.[1,2] Later, some authorities released their ranking results. For example, the Computing Research and Education Association of Australasia (CORE)[3] started to provide rankings for journals and conferences, and these have become important for academic evaluation. Similarly, the China Computer Federation (CCF)[4] has also developed a ranking system for journals and conferences in computer science with three classifications in ten different subfields.

Moreover, measuring different journals and conferences has become a challenging task. One longstanding way of evaluating academic performance is through publication output using citation data (Thelwall and Wilson 2014). In fact, the IF is calculated by the number of citations within the ISI dataset (Garfield 2006). However, there is also an essential challenge in such ranking systems: they do not take into account the place of publications of these citation papers, thereby making them insufficient for the ranking of publications (Zhu et al. 2015). This problem of ignoring the category of citation has attracted a lot of attention, and some improvements have been developed for this bibliometric challenge. For example, as an alternative to the IF, the SCImago Journal Rank (SJR indicator)[5] accounts for both the number of citations received by a journal and the importance or prestige of the journals containing such citations (Falagas et al. 2008; Butler 2008).

Besides the consideration of citation categories in the ranking system, it is also notable that the performance of citation data varies greatly in different areas (Crespo et al. 2014; Marx and Bornmann 2014). For instance, Bornmann et al. (2012) pointed out the chance of a paper being cited is strongly related to the different subfields of chemistry. Crespo et al. (2014) studied the impact of differences in citation practices and argued that the number of citations received by an article depends on the field to which it belongs. As the citation data of a paper are subfield-specific in chemistry and other disciplines, it is reasonable that this phenomenon may also exist in computer science.

The two elements mentioned above, i.e., citation category and research area, could probably affect the ranking of journals and conferences. Thus, in this study, we use conference and journal ranking metrics to investigate how quality and research area, along with other factors, affect the citation performance of academic papers. It is expected that the results will provide an insight for future studies on academic performance evaluation.

[1] http://www3.ntu.edu.sg/home/assourav/crank.html.

[2] http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html.

[3] http://www.core.edu.au/.

[4] http://www.ccf.org.cn/sites/ccf/paiming.jsp.

[5] http://www.scimagojr.com/.

The remainder of this paper is organised as follows. In section "Related work", we review some related concepts and background information regarding academic evaluation. Section "Dataset" introduces the dataset for this study and analyses the preliminary citation rate. In section "Factors influencing citation counts", we explore the factors that affect the citation rate of a paper, examine whether they have the same level influencing citation rates and explain our conclusions.

## Related work

In terms of journal and conference publications, academic evaluation has become an essential topic in bibliometric studies of computer science (Eckmann et al. 2011). The different roles of journals and conferences are frequently debated in the literature.

Chen and Konstan (2010) pointed out that computing researchers are right to view conferences as an important archival venue and use the acceptance rate as an indicator of future impact. With two means of evaluating the citations (the h5 metric and average citations per paper), Vrettas and Sanderson (2015) indicated that the computer science discipline values conferences as a publication venue more highly than any other academic field. Rahm and Thor (2005) analysed the citation frequencies of two main conference databases (SIGMOD and VLDB) and three journal databases (TODS, VLDB Journal, Sigmod Record) over a period of 10 years. They found that the conference papers had a larger average number of citations than journal papers.

However, some other researchers believe that journal publications generally enjoy a higher status than conference publications. Freyne et al. (2010) concluded that the impact of computer science in top-ranking conference papers matches that of papers in middle-ranking journals, and is only slightly beyond the impact of papers in journals in the bottom half of the Thompson Reuters rankings in terms of citations in Google Scholar. Similarly, Franceschet (2010) stated that although computer scientists publish more in conference proceedings than in archival journals, the impact of journal publications is significantly higher than that of conference papers.

From a bibliographic perspective, measuring the quality of academic research and the performance of publications has also been debated. The most commonly used indicator is citation data (Thelwall and Wilson 2014). Bensman et al. (2010) employed the citation rate to evaluate the impact per paper from the perspective of the annual average number of times it is cited. Although these citation-based indicators are commonly used to help research evaluations, there are ongoing controversies about their value, because they cannot accurately reflect the citation category (Thelwall and Fairclough 2015). To solve this problem, citations need to be classified based on their category. For example, Freyne et al. (2010) focused on 15 conferences and 15 journals, including first-, second-, and third-tier venues roughly in line with ISI rankings, to investigate the importance of citation rate. Similarly, Zhu et al. (2015) asked the authors of the citing papers themselves to identify the most influential references, and compared the results with independent annotations.

Inspired by the debate about journal and conference publications in computer science and previous research into the distinction between citation categories, we conducted a study into the quality of different citations with respect to venue type (journal and conference) and other factors including the classification of publications, type of publication, annual average number of papers published by the publication, number of authors, and maximum h-index of all authors of a paper.

## Dataset

### Dataset configuration

The CCF, established in 1956, is one of the largest national academic organisations in China. In 2012, it released a catalogue including ten subfields of important international journals and conferences in the field of computer science (1. Computer systems and high-performance computing; 2. Computer networks; 3. Network and information security; 4. Software engineering/software/programming language; 5. Databases, data mining, and information retrieval; 6. Theoretical computer science; 7. Computer graphics and multimedia; 8. Artificial intelligence and pattern recognition; 9. Human–computer interaction and ubiquitous computing, and 10. Miscellaneous). In this catalogue, journals and conferences are further divided into three different classifications, i.e., A, B, C, according to reputation. Classification A refers to a handful of top international journals and conferences. Following this, classification B refers to internationally famous journals and conferences which have significant academic influence. Finally, classification C refers to important journals and conferences recognised in international academic circles. CCF conference papers are referred to full papers or regular papers, i.e., all the other forms of conference papers (Short paper/Poster/Demo paper/Technical brief/Summary) are not included. In 2014, CCF slightly revised the list and changed the ranking of certain journals and conferences; the present list can be found on the CCF website.[6] In determining the catalogue of rankings, CCF took into account the quality of journals and conferences as well as the broad balance between the different areas. Obviously, the number and quality of journals and conferences are inherently variable, and the catalogue can only be updated to reflect changes occasionally. It is important to point out that this catalogue is a recommendation list that CCF considers worthy of publications by researchers in the field of computer science. In this research, we will use CCF's recommendation list as the guideline for constructing the dataset.

Initially, 102,887 papers published from 2010–2012 in the first nine subfields of the CCF list (excluding Miscellaneous) were retrieved from AMiner.[7] Actually based on papers' titles and their publication venues in AMiner dataset, we first distinguished CCF papers, and only kept full papers/regular papers for conferences. In general, this AMiner dataset (Tang et al. 2008) includes paper information, paper citation, author information, and author collaborations. It consists of four files: (1) AMiner-Paper.rar, which includes 2,092,356 papers and 8,024,869 citations; (2) AMiner-Author.zip, with details of 1,712,433 authors; (3) AMiner-Coauthor.zip, containing 4,258,615 collaboration relationships; and (4) AMiner-Author2-Paper.zip, which includes the relationship between author ID and paper ID. We downloaded this dataset in early June 2015. As the AMiner dataset has a full range of computer science papers and related author information, we designed our database based on this dataset. The papers selected for our dataset were published in 201 journals and 261 conferences listed on the CCF website, which contains a total of 236 journals and 303 conferences. Thus, our dataset covers more than 85% of publication venues in these nine subfields. Table 1 summarizes the dataset's coverage of publications for the nine CCF subfields.

Secondly, we further crawled the citation information of each paper in our dataset to determine its overall citation count and identify corresponding cited papers from Google

[6] http://www.ccf.org.cn/sites/ccf/paiming.jsp.

[7] https://cn.aminer.org/aminernetwork.

**Table 1** Dataset coverage of publications for nine CCF subfields

| Subfields | Journal number (CCF number) | | | Journal proportion | | | Conference number (CCF number) | | | Conference proportion | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | A (%) | B (%) | C (%) | A | B | C | A (%) | B (%) | C (%) |
| 1 | 3 (3) | 11 (11) | 9 (10) | 100 | 100 | 90.0 | 5 (5) | 24 (26) | 23 (26) | 100 | 92.3 | 88.5 |
| 2 | 3 (3) | 6 (6) | 8 (10) | 100 | 100 | 80.0 | 3 (3) | 10 (11) | 10 (17) | 100 | 90.9 | 58.8 |
| 3 | 3 (3) | 3 (4) | 5 (8) | 100 | 75.0 | 78.6 | 5 (5) | 11 (12) | 16 (20) | 100 | 91.7 | 80.0 |
| 4 | 3 (3) | 12 (13) | 6 (8) | 100 | 92.3 | 75.0 | 7 (7) | 21 (21) | 21 (22) | 100 | 100 | 95.5 |
| 5 | 4 (4) | 12 (13) | 11 (14) | 100 | 92.3 | 78.6 | 5 (5) | 9 (11) | 13 (13) | 100 | 81.8 | 100 |
| 6 | 2 (2) | 13 (13) | 8 (12) | 100 | 100 | 66.7 | 3 (3) | 7 (7) | 7 (10) | 100 | 100 | 70.0 |
| 7 | 3 (3) | 8 (10) | 9 (12) | 100 | 80.0 | 75.0 | 2 (3) | 8 (11) | 7 (10) | 66.7 | 72.7 | 70.0 |
| 8 | 4 (4) | 18 (20) | 29 (37) | 100 | 90.0 | 78.4 | 4 (5) | 7 (13) | 15 (17) | 80.0 | 53.8 | 88.2 |
| 9 | 2 (2) | 3 (4) | 3 (4) | 100 | 75.0 | 75.0 | 2 (2) | 5 (6) | 11 (12) | 100 | 83.3 | 91.7 |
| Total | 27 (27) | 86 (94) | 88 (115) | 100 | 91.4 | 76.5 | 36 (38) | 102 (118) | 123 (147) | 94.7 | 86.4 | 83.7 |

Subfield: 1—Computer systems and high-performance computing; 2—Computer networks; 3—Network and information security; 4—Software engineering/software/programming language; 5—Databases, data mining, and content retrieval; 6—Theoretical computer science; 7—Computer graphics and multimedia; 8—Artificial intelligence and pattern recognition; 9—Human–computer interaction and ubiquitous computing

**Table 2** Citation distribution from different publication years in dataset

| Publication year | Citation count | A | | B | | C | |
|---|---|---|---|---|---|---|---|
| | | Jour | Conf | Jour | Conf | Jour | Conf |
| 2010 | 0 | 76 | 160 | 273 | 491 | 453 | 929 |
| | 1–1000 | 2235 | 2888 | 6829 | 6560 | 6608 | 6717 |
| | 1000+ | 6 | 4 | 1 | 4 | 1 | 1 |
| 2011 | 0 | 58 | 219 | 287 | 549 | 436 | 913 |
| | 1–1000 | 2452 | 4042 | 6182 | 6404 | 7238 | 6065 |
| | 1000+ | 6 | 1 | 3 | 0 | 2 | 0 |
| 2012 | 0 | 54 | 264 | 343 | 606 | 605 | 1262 |
| | 1–1000 | 2066 | 3505 | 5676 | 6643 | 7357 | 5411 |
| | 1000+ | 0 | 1 | 1 | 0 | 0 | 0 |

Scholar up to the end of August 2015. These citations were distinguished into different CCF classifications/types or non-CCF papers according to the list of papers in CCF venues published from 2010 to 2015 in DBLP.[8] For conferences we only referred to proceedings, which is consistent with CCF catalogue definition for conference papers. Due to the heavy workload of distinguishing citations, we can not guarantee that the accuracy of this process can be 100%. However, we can ensure that the precision of this process can reach 95%. The citation distribution of the papers in the dataset is presented in Table 2. In this study, a paper's citation count cannot be greater than 1000, because the largest number of citation papers returned by Google Scholar is 1000 and we cannot perfectly count the distribution of citation categories for those not in this list. From Table 2, it is clear that about 7.6% of papers have never been cited, and there are only 31 papers whose citation count is greater than 1000. Papers with citation counts of 1–1000 occupy over 92% of the 102,887 papers.

## Preliminary citation rate analysis

To better understand the dataset, some basic variables and related symbols are defined in Table 3. Based on these basic variables, we can now define some fundamental concepts used in this research.

(1) *Citation count* Given a paper $p \in PaperSet(y, c, s, t)$, where $p$ was published in year $y$, and with the subfield $s$, type $t$, and classification $c$, its citation count is defined as $CC(p)$. Furthermore, its citation papers can be further identified as to whether they come from the CCF list. As a result, the citation count can be further defined as:

$$\begin{aligned} CC(p) = {} & CC\_AJ(p) + CC\_AC(p) \\ & + CC\_BJ(p) + CC\_BC(p) \\ & + CC\_CJ(p) + CC\_CC(p) \\ & + CC\_NONCCF(p) \end{aligned} \tag{1}$$

where $CC\_AJ(p)$ indicates paper $p$'s citation count from CCF recommended A journals. All other variables represent the citation count from CCF recommended B and C journals, A, B, and C conferences, and non-CCF-listed venues.

[8] http://dblp.uni-trier.de/db/.

**Table 3** Symbols of dataset

| Symbol | Description |
|---|---|
| $p$ | A publication |
| $s$ | The subfield of a given publication, $s \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ |
| $t$ | The type of a given publication, $t \in \{journal, conference\}$ |
| $c$ | The classification of a given publication, $c \in \{A, B, C\}$ |
| $y$ | The year a publication was published; in this paper, $y \in \{2010, 2011, 2012\}$ |
| $CitationYear(p)$ | The publication year set of citation papers in a publication, $y \leq CitationYear(p) \leq 2015$ |
| $PublishedTime(p)$ | How long since the publication was published; in this paper, it is calculated as $2015 - y$ |
| $PaperSet(y, c, s, t)$ | The paper set, where $s$ is the subfield, $t$ is the venue type, $c$ is the domain classification, and $y$ is the publication year |
| $n(y, c, s, t)$ | The number of papers in paper set PaperSet(y,c,s,t) |

(2) *Citation rate (CR)* Citation rate is the annual average number of times that a paper has been cited since it was published (Bornmann et al. 2012; Bensman et al. 2010). In this study, this metric is employed to evaluate the impact per paper from the perspective of the annual average number of times it is cited. Based on the citation count, $CR(p)$ is defined as follows:

$$CR(p) = \frac{CC(p)}{PublishedTime(p)} \tag{2}$$

Similarly, CR can be divided into CCF listed categories. For example, paper $p$'s citation rate within CCF A journals can be defined as:

$$CR\_AJ(p) = \frac{CC\_AJ(p)}{PublishedTime(p)} \tag{3}$$

In our study, we define CR as the total citation rate of a paper as calculated by Eq. 2. As mentioned above, the category of citation papers can be distinguished. Therefore, CR can be divided into seven parts: (1) CR_AJ; (2) CR_AC; (3) CR_BJ; (4) CR_BC; (5) CR_CJ; (6) CR_CC; and (7) CR_NONCCF. These represent the different classifications and different types of citation papers according to the CCF classifications. For example, A in CR_AJ denotes classification A and J denotes journals; the first C in CR_CC denotes classification C and the second C denotes conferences.

(3) Besides the evaluation of an individual paper's CR, we also investigate the geometric mean citation rate for a certain category. The geometric mean is based on the arithmetic mean of the natural log of the data, and is more appropriate than the basic arithmetic mean for highly skewed data, such as citation data, because it is less affected by a few large values (Zitt 2012). As the citation data contains zero values, we add 1 to the citation rate to ensure that the log of the data can be calculated (Fairclough and Thelwall 2015). Under this condition, the geometric mean citation rate for a category $\overline{CR}(y, c, s, t)$ is defined as follows:

$$\overline{CR}(y, c, s, t) = \left( \prod_{p \in PaperSet(y,c,s,t)} (CR(p) + 1) \right)^{\frac{1}{n(y,c,s,t)}} \tag{4}$$

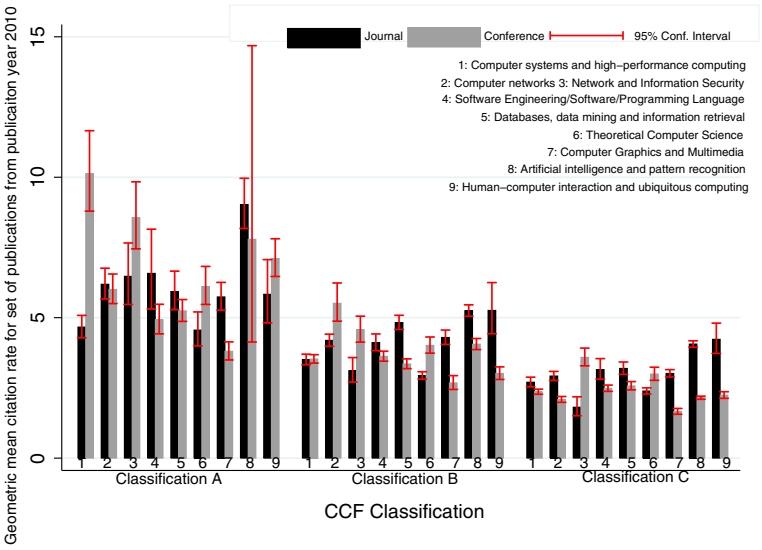


**Fig. 1** Geometric mean citation rate for CCF papers published in 2010, grouped by CCF classification, subfield, and venue type

This equation can be rewritten in the form of the natural log of the citation data as follows:

$$\overline{CR}(y,c,s,t) = e^{\frac{1}{n(y,c,s,t)} \sum_{p \in PaperSet(y,c,s,t)} ln(CR(p)+1)} \tag{5}$$

Similarly, the citations also come from different CCF classification venues. For example, for a certain category, the geometric mean citation rate from CCF A journals can be defined as:

$$\overline{CR}\_AJ(y,c,s,t) = \left( \prod_{p \in PaperSet(y,c,s,t)} (CR\_AJ(p)+1) \right)^{\frac{1}{n(y,c,s,t)}} \tag{6}$$

As defined by Eq. 4, $\overline{CR}(y,c,s,t)$ represents a geometric mean citation rate for a specific set of papers, where $y$ belongs to $\{2010, 2011, 2012\}$, $c$ belongs to $\{A, B, C\}$, $s$ belongs to $\{1, 2, \ldots, 9\}$, and $t$ belongs to $\{journal, conference\}$. Summary statistics for $\overline{CR}(y,c,s,t)$ with a 95% confidence interval for sets of publications from 2010, 2011, and 2012 are depicted in Figs. 1, 2 and 3.

Figures 1, 2 and 3 each contain 54 bars distributed into three big groups labelled Classification A, Classification B, and Classification C. In each group, the labels 1–9 represent the nine subfields. Each subfield has two bars representing different venue types (i.e. journal in black and conference in grey). For example, the first bar on the left of Fig. 1 represents $\overline{CR}(2010, A, 1, journal)$, namely the geometric mean citation rate for CCF papers published in 2010, grouped by CCF classification A, subfield 1, and venue type journal.

After investigating the details of every subfield from Figs. 1, 2 and 3, three inequalities can be derived:

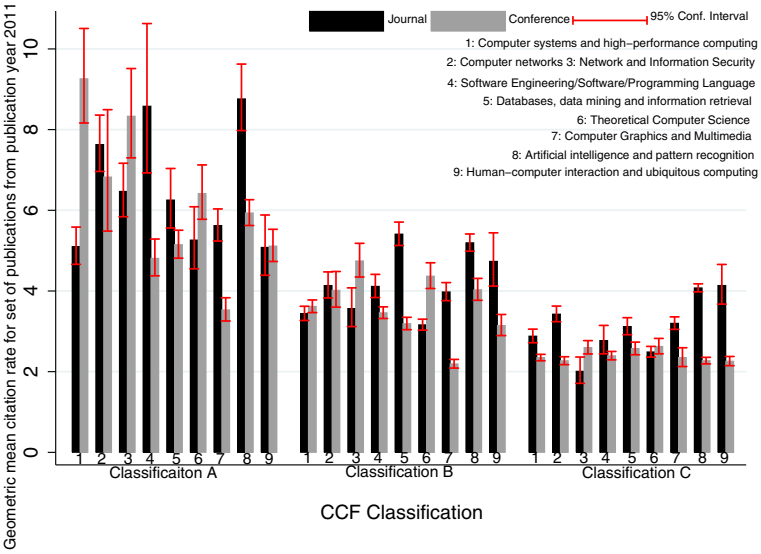


**Fig. 2** Geometric mean citation rate for CCF papers published in 2011, grouped by CCF classification, subfield, and venue type

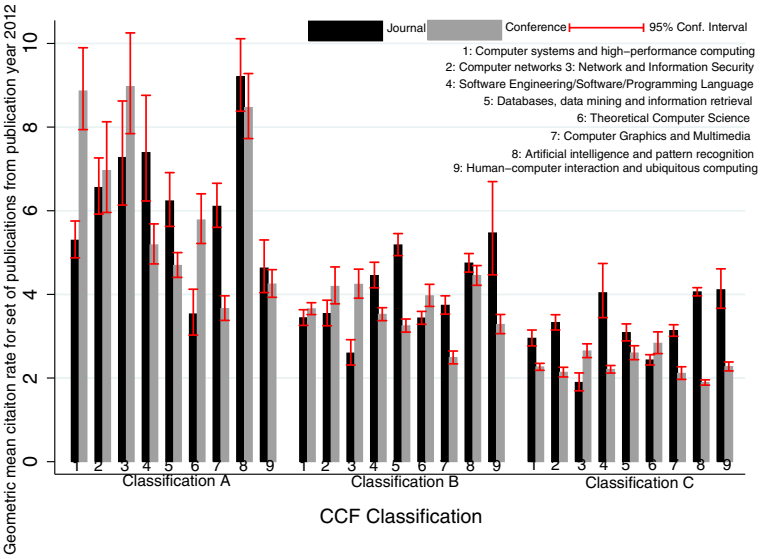


**Fig. 3** Geometric mean citation rate for CCF papers published in 2012, grouped by CCF classification, subfield, and venue type

$$\forall y \in \{2010, 2011, 2012\}, \forall s \in \{1, 3, 6\}, \\ \overline{CR}(y, A, s, conference) > \overline{CR}(y, A, s, journal) \tag{7}$$

$$\forall y \in \{2010, 2011, 2012\}, \forall s \in \{1, 3, 6\}, \\ \overline{CR}(y, B, s, conference) > \overline{CR}(y, B, s, journal) \tag{8}$$

$$\forall y \in \{2010, 2011, 2012\}, \forall s \in \{3, 6\}, \\ \overline{CR}(y, C, s, conference) > \overline{CR}(y, C, s, journal) \tag{9}$$

There are three computer science subfields for which the higher classification (A and B) conferences have more general impact than higher classification journals from the perspective of the geometric citation mean for sets of publications from different publication years.

Besides the overall geometric mean citation rate, we further investigated the difference in terms of geometric citation rate from different venue types and classifications. Similar to Figs. 1, 2 and 3, a breakdown of the geometric mean citation rate from different CCF venue types and classifications is presented in Fig. 4a–f, i.e., CR_AJ, CR_AC, CR_BJ, CR_BC, CR_CJ, CR_CC. Note that we do not include papers with citation counts greater than 1000—these 31 papers were neglected as we can only retrieve 1000 citation papers from Google Scholar. Thus, we cannot determine the full distribution of citation rates for these 31 papers.

From Fig. 4a, it is apparent that $\overline{CR}\_AJ(y, c, s, t)$ decreases sharply from c=A to c=C. From Fig. 4b, we can see that $\overline{CR}\_AC(y, c, s, conference)$ is universally greater than $\overline{CR}\_AC(y, c, s, journal)$. From Fig. 4c, d, the overall trend for the decrease from $\overline{CR}\_BJ(y, A, s, t)$ to $\overline{CR}\_BJ(y, B, s, t)$ and $\overline{CR}\_BC(y, A, s, t)$ to $\overline{CR}\_BC(y, B, s, t)$ is similar to that in Fig. 4a, b, although the gap between them is reduced. From Fig. 4e, f, the performance of $\overline{CR}\_CJ(y, C, s, t)$ and $\overline{CR}\_CC(y, C, s, t)$ appears to be much better than for the previous cases.

## Factors influencing citation counts

### Factor description

Bibliometric studies published in recent years have revealed the associations among a number of factors concerning paper citation rates (Bornmann et al. 2012; Tahamtan et al. 2016). The citation rate of a paper is influenced by various "extrinsic" factors not directly related to the content or quality of the paper (Onodera and Yoshikane 2015; Smolinsky 2016).

*Subfield*

Crespo et al. (2014) researched the impact of subfields in citation practices, and argued that the number of citations received by an article depends on the field to which it belongs. Bornmann et al. (2012) also proved that the chance of a paper being cited is strongly related to the different subfields of chemistry . Therefore, in this study, it is reasonable to assume that the performance of citation data will also vary in different subfields in the domain of computer science.

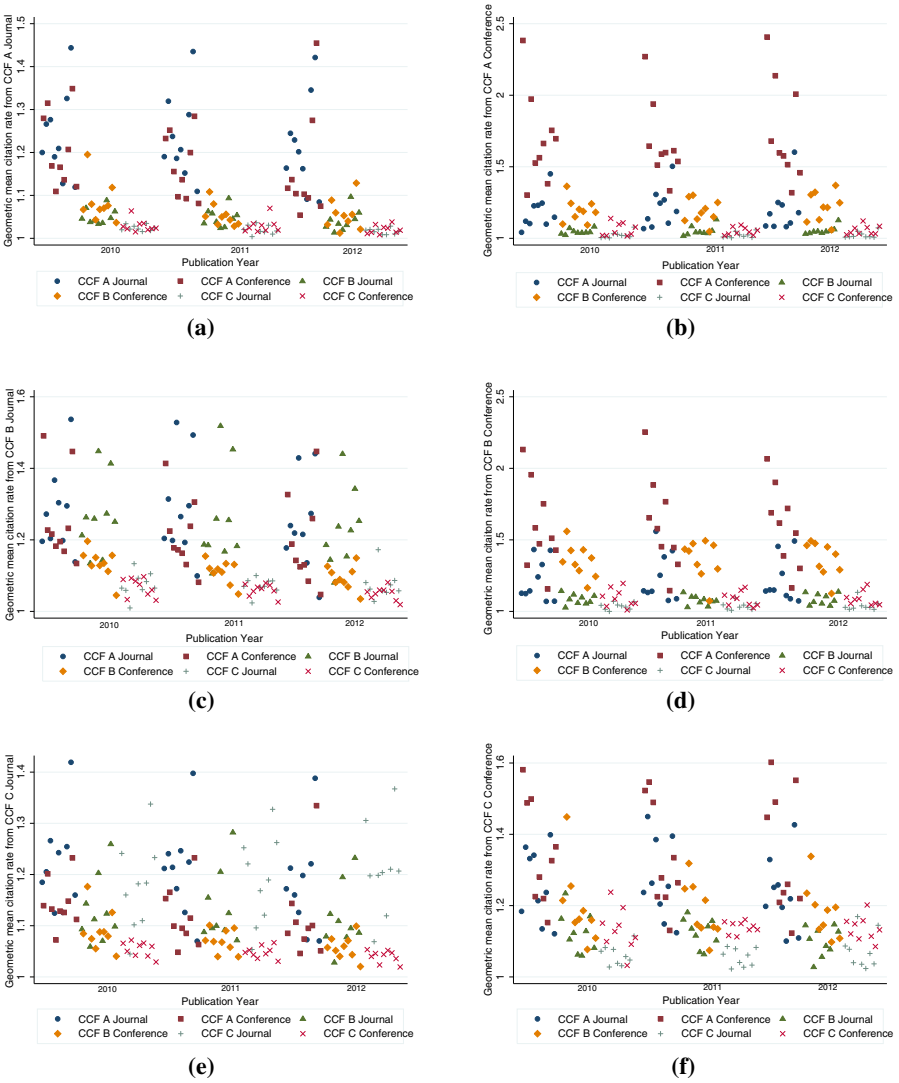


**Fig. 4** Geometric mean citation rate from different CCF classifications and venue types for sets of publications from publication years 2010, 2011, and 2012. **a** Geometric mean citation rate from CCF Journal A for CCF papers published in 2010, 2011, and 2012 grouped by CCF classification, subfield, and venue type. **b** Geometric mean citation rate from CCF Conference A for CCF papers published in 2010, 2011, and 2012 grouped by CCF classification, subfield, and venue type. **c** Geometric mean citation rate from CCF Journal B for CCF papers published in 2010, 2011, and 2012 grouped by CCF classification, subfield, and venue type. **d** Geometric mean citation rate from CCF Conference B for CCF papers published in 2010, 2011, and 2012 grouped by CCF classification, subfield, and venue type. **e** Geometric mean citation rate from CCF Journal C for CCF papers published in 2010, 2011, and 2012 grouped by CCF classification, subfield, and venue type. **f** Geometric mean citation rate from CCF Conference C for CCF papers published in 2010, 2011, and 2012 grouped by CCF classification, subfield, and venue type

*Type of publication*

Freyne et al. (2010) used a large-scale experiment covering 8,764 journal and conference papers to highlight how leading conferences compare favourably to mid-ranking journals, demonstrating that conference publications enjoy greater status in computer science than in other disciplines. Franceschet (2010) gave a bibliometric view of the publishing frequency and impact of conference proceedings compared to archival journal publication , insisting that meetings in the computer field hold special status because they have the advantage of offering scholars the opportunity to present and discuss their paper with peers. Therefore, it will be interesting to investigate the difference between journals and conferences in computer science from a citation perspective.

*Classification of publications*

The publications listed in the high-ranking classification can receive more attention from scholars in academic circles (Beel and Gipp 2010). Moreover, the quality of papers published in high-ranking classifications should be guaranteed and more strictly selected. When scholars cite papers to support their own study, they prefer to cite papers published in high-ranking classifications to make their paper more convincing. As a result, it will be interesting to study how the classification affects the overall citation data.

*Annual average number of papers published of the publication*

The IF reflects the average number of citations of articles recently published in a journal (Seglen 1994), and can itself attract citations to articles in the publication (Van Dalen and Henkens 2005). As the IF depends on the number of papers published by a publication, it is reasonable to argue that this factor influences the citation rate—when a publication venue publishes more papers, more scholars are automatically associated with the publication, which will expand its academic circle and ensure the paper is more widely known. However, in the digital age, papers are no longer tied to their respective journals, and can be passed among scholars electronically. Hence, papers can now be read and cited based on their own merits, independently of the journals physical availability, reputation, or IF (Lozano et al. 2012). Therefore, it is argued that the annual average number of papers published by a publication could affect the impact of this publication.

*Number of authors*

Some researchers have proposed three points associated with a positive association between the number of authors and the citation rates of papers (Leimu and Koricheva 2005; Peng and Zhu 2012; Rigby 2013), whereas other studies have demonstrated that the ability of the number of authors to predict the citation impact of articles is weak or insignificant (Walters 2006; Bornmann et al. 2012). As there are conflicting conclusions from different fields, it will be very interesting to verify this effect in the computer field.

*Maximum h-index of all authors of a paper*

There have been many discussions about the halo effect on scientific impact, suggesting that articles written by authors with high h-index values attract more citations than those

written by others (Onodera and Yoshikane 2015). The reputation of a scholar is normally positively correlated with his/her h-index. A scholar's h-index is defined as having h papers that have each been cited in other papers at least h times. The higher the h-index of an author, the better reputation or the higher achievement level the author has. As a result, it is meaningful to explore the association between citation rates and the highest h-index of the authors of a co-authored paper. If there is only one author, the highest h-index is that of the author. If the maximum h-index of all authors is very high, it indicates that an authoritative scholar is the (co-)author of this paper.

The above six factors of a publication are denoted as (1) subfield, (2) type, (3) classification, (4) avgPubCount, (5) author_number, and (6) author_max_h_index. In this study, we performed a multiple regression analysis to reveal the factors that exert the strongest effect on a certain outcome.

## Regression analysis

### *Convert continuous variables to categorical variables*

To study the impact on citation rate of different levels of avgPubCount, author_number, and author_max_h_index, we must classify these factors into different categories, namely cat_avgPubCount, cat_author_number, and cat_author_max_h_index. For cat_avgPubCount, we categorize papers into ten groups on the basis of the average publication count of the venue where the paper is published. For cat_author_max_h_index, we do the same thing on the basis of the maximum author h-index of all the authors of the paper. The bounds between categories are determined by the accumulation of papers in one category. Every category accounts for approximately 10% of all papers. Regarding author_number, as most papers have fewer than six authors, we categorised the papers into six groups denoting 1, 2, 3, 4, 5, and more than five authors. The results of this conversion are presented in Table 4.

### *Regression model selection*

Our outcome variables are count data, and the normal regression models for this kind of outcome variable are the Poisson regression model (PRM) or negative binominal regression model (NBRM) (Cameron and Trivedi 2013). As the outcome variables for PRM and NBRM must be non-negative integers, we cannot directly use the citation rate as the outcome variable. However, PRM and NBRM may also be appropriate for rate data, where the rate is a count of events occurring to a particular unit of observation divided by some measure of that unit's exposure (Dalgaard 2008). For example, biologists may count the number of tree species in a forest, and the rate would be the number of species per square kilometer. More generally, event rates can be calculated as the number of events per unit time, which allows the observation window to vary for each unit. In these examples, exposure is the unit area, person-years, or unit time. In our study, the citation rate (citation count per year) is an integer variable, and the exposure can be set as $(2015 - y)$, where $y \in \{2010, 2011, 2012\}$ is the year of publication. To facilitate the following, we call this variable the time, where time = $(2015 - y)$. Therefore, PRM and NBRM can be used to research the citation rate.

Poisson regression is often used for modelling count data, and there are a number of extensions that are useful for count models. NBRM is considered as a generalization of PRM, as it has the same mean structure as PRM and an extra parameter to model the over-

**Table 4** Results of variable conversion

| cat_avgPubCount | avgPubCount | Freq | Percent | Cumulation |
|---|---|---|---|---|
| 1 | (0, 40] | 12,483 | 12.14 | 12.14 |
| 2 | (40, 55] | 8510 | 8.27 | 20.41 |
| 3 | (55, 80] | 12,663 | 12.31 | 32.72 |
| 4 | (80, 100] | 9179 | 8.92 | 41.64 |
| 5 | (100, 130] | 8171 | 7.94 | 49.58 |
| 6 | (130, 160] | 10,811 | 10.51 | 60.09 |
| 7 | (160, 215] | 10,549 | 10.26 | 70.35 |
| 8 | (215, 300] | 10,669 | 10.37 | 80.72 |
| 9 | (300, 400] | 10,291 | 10.01 | 90.73 |
| 10 | $(400, +\infty]$ | 9530 | 9.27 | 100.00 |

| cat_author_max_h_index | author_max_h_index | Freq | Percent | Cumulation |
|---|---|---|---|---|
| 1 | [0, 1] | 11,497 | 11.18 | 11.18 |
| 2 | (1, 2] | 7129 | 6.93 | 18.11 |
| 3 | (2, 4] | 15,259 | 14.84 | 32.95 |
| 4 | (4, 5] | 7334 | 7.13 | 40.08 |
| 5 | (5, 7] | 13,079 | 12.72 | 52.80 |
| 6 | (7, 9] | 10,603 | 10.31 | 63.11 |
| 7 | (9, 11] | 8613 | 8.37 | 71.48 |
| 8 | (11, 15] | 11,678 | 11.35 | 82.82 |
| 9 | (15, 20] | 8679 | 8.44 | 91.27 |
| 10 | $(20, +\infty]$ | 8985 | 8.73 | 100.00 |

| cat_author_number | author_number | Freq | Percent | Cumulation |
|---|---|---|---|---|
| 1 | (0, 1] | 9585 | 9.32 | 9.32 |
| 2 | (1, 2] | 26,913 | 26.17 | 35.49 |
| 3 | (2, 3] | 29,965 | 29.13 | 64.62 |
| 4 | (3, 4] | 19,963 | 19.41 | 84.03 |
| 5 | (4, 5] | 9468 | 9.21 | 93.24 |
| 6 | $(5, +\infty)$ | 6962 | 6.76 | 100.00 |

dispersion whereby the conditional variance of the dependent variable exceeds the conditional mean (Long and Freese 2006). Therefore, NBRM can be used for over-dispersed count data. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for NBRM are likely to be narrower than those for PRM (Berk and MacDonald 2008). If over-dispersion is present, estimates from the PRM are inefficient with standard errors that are biased downward, even if the model includes the correct variables. Accordingly, it is important to test for over-dispersion. Because the NBRM reduces to the PRM when $\alpha = 0$ ($\alpha$ is known as the dispersion parameter), we can test for over-dispersion by testing $H0 : \alpha = 0$. To test this hypothesis, Stata provides a likelihood-ratio test that is listed after the estimates of the parameters for the routine "nbreg". Thus, we performed this test for six citation rates from different CCF classification and venue

types for sets of publications from 2010, 2011, and 2012. The results show that $\alpha$ is significantly different from 0. Clearly, over-dispersion is a problem, and the NBRM is preferred.

From the above, there is good reason to use the NBRM to deal with our data. In Stata, we can directly use the nbreg command below to construct the NBRM and apply "listcoef, help percent" to show the percentage change in the expected count of the outcome variable (in this example, CR_AJ and publication year 2010) when the categorical variable changes from the base to another category (Bruin 2006). [There is another point to explain here: in Stata, to treat a variable as a categorical variable, we need to add **i.** in front of the variable name (StataCorp 2005)]

```
. nbreg CR_AJ i.category i.type i.classification
  i.category i.cat_avgPubCount i.cat_author_number
  i.cat_author_max_h_index if publicationYear == 2010,
  exposure(time)

. listcoef, help percent
```

With this method, we can deal with different citation rates (CR, CR_AJ, CR_AC, CR_BJ, CR_BC, CR_CJ, CR_CC) as outcome variables for sets of publications from 2010, 2011, and 2012 separately. The results are presented in Tables 5, 6 and 7.

**Table 5** NBRM: Percentage change in expected citation rates compared with base for the set of publications from publication year 2010

| Factors | CR | CR_AJ | CR_AC | CR_BJ | CR_BC | CR_CJ | CR_CC |
|---|---|---|---|---|---|---|---|
| *Computer science subfield* | | | | | | | |
| Computer systems and high-performance computing | Base | | | | | | |
| Computer networks | 19.0 | 64.2 | 38.9 | −3.5 | −45.8 | 10.0 | 53.3 |
| Network and information security | 34.3 | 1.1 | 77.6 | −15.1 | −25.8 | −3.2 | 7.1 |
| Software engineering/software/programming language | −2.1 | −43.7 | 26.7 | 5.5 | −22.7 | −28.8 | −27.8 |
| Databases, data mining, and information retrieval | 15.6 | −20.4 | 125.5 | 11.9 | −23.0 | 11.2 | −36.8 |
| Theoretical computer science | −22.2 | −27.1 | 101.1 | −29.9 | 1.8 | −31.9 | −44.3 |
| Computer graphics and multimedia | −1.8 | 61.7 | 54.9 | −5.9 | −51.3 | 7.4 | −31.9 |
| Artificial intelligence and pattern recognition | 38.4 | 40.3 | 212.0 | 19.4 | 4.1 | 64.2 | −10.1 |
| Human–computer interaction and ubiquitous computing | 7.0 | −39.6 | 149.8 | −52.4 | −36.9 | −32.9 | −29.4 |
| *Type of publication* | | | | | | | |
| Journal | Base | | | | | | |
| Conference | −38.3 | 1.1 | 208.2 | −50.3 | 191.4 | −64.5 | 32.4 |
| *Classification of publication* | | | | | | | |
| A | Base | | | | | | |
| B | −36.2 | −66.5 | −70.2 | −8.0 | −41.7 | −31.2 | −36.9 |
| C | −58.5 | −85.1 | −89.8 | −59.9 | −81.6 | −13.6 | −59.8 |

**Table 5** continued

| Factors | CR | CR_AJ | CR_AC | CR_BJ | CR_BC | CR_CJ | CR_CC |
|---|---|---|---|---|---|---|---|
| *Annual average number of papers published by the publication* | | | | | | | |
| (0, 40] | Base | | | | | | |
| (40, 55] | 2.5 | −28.3 | −25.3 | 1.0 | −20.6 | 18.9 | −15.7 |
| (55, 80] | −10.6 | −25.6 | −37.4 | −11.2 | −20.3 | 9.3 | −14.1 |
| (80, 100] | −17.6 | −26.9 | −49.4 | 2.0 | −44.4 | 9.9 | −20.2 |
| (100, 130] | −9.0 | 0.2 | −37.1 | 2.5 | −25.7 | 10.5 | −2.2 |
| (130, 160] | −8.4 | −31.1 | −50.1 | −8.8 | −26.0 | 21.7 | −13.8 |
| (160, 215] | −18.8 | −38.3 | −65.1 | 26.0 | −44.1 | 25.9 | 3.2 |
| (215, 300] | −21.1 | −34.1 | −65.1 | −24.6 | −65.2 | −9.4 | −27.5 |
| (300, 400] | −10.6 | −20.9 | −58.8 | 37.2 | −65.0 | 73.4 | −15.4 |
| 400+ | −5.9 | −38.6 | −74.3 | −10.2 | −75.9 | 26.9 | −22.3 |
| *Number of authors* | | | | | | | |
| 1 | Base | | | | | | |
| 2 | 0.3 | 15.3 | −14.6 | 0.0 | −4.7 | −8.0 | 16.3 |
| 3 | 6.8 | 14.1 | −15.6 | 2.8 | −7.3 | 1.1 | 19.7 |
| 4 | 9.0 | 22.1 | −12.3 | 7.2 | −4.6 | 5.1 | 23.3 |
| 5 | 6.8 | 18.7 | −20.3 | 2.9 | −10.9 | 4.0 | 17.9 |
| 5+ | 33.1 | 25.2 | −1.4 | 15.6 | 6.8 | 20.0 | 51.8 |
| *Maximum h-index of all authors of a paper* | | | | | | | |
| 0–1 | Base | | | | | | |
| 2 | 26.5 | 52.2 | 18.0 | 33.2 | 38.6 | 44.9 | 43.1 |
| 3–4 | 48.1 | 141.9 | 100.2 | 89.9 | 108.4 | 83.9 | 98.3 |
| 5 | 60.8 | 212.7 | 172.1 | 139.1 | 173.2 | 109.6 | 149.5 |
| 6–7 | 74.4 | 239.7 | 291.2 | 166.1 | 275.4 | 129.1 | 158.4 |
| 8–9 | 103.5 | 358.5 | 445.3 | 202.2 | 415.3 | 186.6 | 225.9 |
| 10–11 | 122.1 | 411.1 | 522.3 | 247.4 | 469.8 | 184.6 | 280.2 |
| 12–15 | 160.5 | 548.6 | 778.1 | 287.9 | 652.3 | 221.8 | 320.7 |
| 16–20 | 198.0 | 647.4 | 1021.9 | 349.2 | 868.2 | 270.6 | 386.2 |
| 20+ | 297.6 | 840.6 | 1594.5 | 411.9 | 1207.1 | 384.2 | 503.1 |

## Results

### *Factor 1: Subfield in CS*

Taking subfield 1 (Computer systems and high-performance computing) as the base, we can calculate the percentage change in expected citation rates if the subfield changes to another while holding all other variables constant. The results in Tables 5, 6 and 7 are in accordance with our assumption. For example, compared with the base, a paper in subfield 7 (Computer graphics and multimedia) increases the CR_AJ by 61.7, 107.0, and 159.9% for publication years 2010, 2011, and 2012. Moreover, a paper in subfield 4 (Software engineering/software/programming language) decreases the CR_AJ by 43.7, 36.5, and 38.3%. Moreover, subfield 8 (Artificial intelligence and pattern recognition) always comes out among the top of all nine fields for different citation rates. Publications in subfield 8 are

**Table 6** NBRM: Percentage change in expected citation rates compared with base for the set of publications from publication year 2011

| Factors | CR | CR_AJ | CR_AC | CR_BJ | CR_BC | CR_CJ | CR_CC |
|---|---|---|---|---|---|---|---|
| *Computer science subfield* | | | | | | | |
| Computer systems and high-performance computing | Base | | | | | | |
| Computer networks | 22.4 | 78.0 | 50.5 | −6.6 | −30.4 | 18.2 | 22.0 |
| Network and information security | 30.0 | 21.8 | 82.4 | −24.9 | −25.5 | −5.3 | −6.6 |
| Software engineering/software/programming language | −3.9 | −36.5 | 12.6 | −2.9 | −20.9 | −36.3 | −38.2 |
| Databases, data mining, and information retrieval | 22.1 | −15.2 | 116.2 | 15.6 | −34.9 | 9.7 | −37.7 |
| Theoretical computer science | −10.9 | −21.3 | 90.8 | −17.8 | 1.7 | −22.8 | −35.8 |
| Computer graphics and multimedia | −2.3 | 107.0 | 55.8 | −11.8 | −58.5 | 4.4 | −21.3 |
| Artificial intelligence and pattern recognition | 52.9 | 65.6 | 206.3 | 40.9 | 18.9 | 87.7 | 1.7 |
| Human–computer interaction and ubiquitous computing | 2.4 | −38.0 | 150.3 | −62.1 | −45.4 | −32.3 | −32.4 |
| *Type of publication* | | | | | | | |
| Journal | Base | | | | | | |
| Conference | -37.0 | 1.9 | 224.8 | -47.2 | 236.4 | -65.9 | 60.8 |
| *Classification of publication* | | | | | | | |
| A | Base | | | | | | |
| B | −39.6 | −78.4 | −78.2 | −22.0 | −49.7 | −34.3 | −48.6 |
| C | −59.0 | −89.7 | −92.9 | −70.6 | −86.0 | −19.4 | −69.7 |
| *Annual average number of papers published by the publication* | | | | | | | |
| (0, 40] | Base | | | | | | |
| (40, 55] | −5.2 | −23.9 | −25.6 | −2.2 | −32.4 | −4.2 | −22.6 |
| (55, 80] | −9.8 | −29.5 | −34.2 | −6.4 | −29.7 | −6.0 | −12.6 |
| (80, 100] | −15.7 | −23.7 | −59.0 | 15.1 | −34.7 | 1.1 | −28.4 |
| (100, 130] | −18.8 | −18.3 | −57.2 | −0.2 | −37.9 | −4.6 | −15.8 |
| (130, 160] | −18.6 | −37.7 | −60.4 | −20.0 | −46.1 | 4.3 | −32.4 |
| (160, 215] | −13.9 | −33.9 | −60.7 | 34.0 | −47.0 | 25.8 | 0.1 |
| (215, 300] | −26.6 | −47.0 | −75.3 | −16.0 | −65.1 | −22.0 | −22.1 |
| (300, 400] | −19.8 | −48.7 | −74.4 | 33.8 | −66.2 | 53.1 | −40.0 |
| 400+ | −7.6 | −42.9 | −65.6 | −14.0 | −57.4 | 27.3 | −48.0 |
| *Number of authors* | | | | | | | |
| 1 | Base | | | | | | |
| 2 | 16.2 | 37.3 | 31.5 | 23.1 | -1.2 | 10.5 | 56.4 |
| 3 | 24.7 | 53.9 | 31.7 | 36.8 | 5.3 | 26.8 | 78.3 |
| 4 | 34.1 | 59.7 | 43.6 | 50.4 | 9.8 | 42.2 | 82.2 |
| 5 | 43.1 | 76.0 | 45.6 | 57.0 | 10.6 | 47.8 | 114.2 |
| 5+ | 54.1 | 53.8 | 46.3 | 54.8 | 29.7 | 53.6 | 123.1 |
| *Maximum h-index of all authors of a paper* | | | | | | | |
| 0–1 | Base | | | | | | |
| 2 | 22.6 | 10.8 | 30.6 | 25.0 | 37.4 | 33.0 | 39.9 |
| 3–4 | 34.8 | 47.0 | 63.5 | 44.7 | 53.1 | 54.6 | 57.7 |

**Table 6** continued

| Factors | CR | CR_AJ | CR_AC | CR_BJ | CR_BC | CR_CJ | CR_CC |
|---|---|---|---|---|---|---|---|
| 5 | 46.4 | 80.8 | 92.4 | 73.5 | 79.6 | 74.1 | 90.5 |
| 6–7 | 46.8 | 81.9 | 128.3 | 79.6 | 123.8 | 69.7 | 89.9 |
| 8–9 | 64.8 | 131.4 | 176.1 | 90.4 | 156.6 | 85.4 | 122.9 |
| 10–11 | 78.4 | 123.2 | 246.3 | 110.0 | 200.7 | 90.3 | 132.4 |
| 12–15 | 110.4 | 168.8 | 370.8 | 137.8 | 263.0 | 92.5 | 160.9 |
| 16–20 | 142.9 | 214.5 | 476.5 | 168.4 | 345.1 | 138.1 | 191.0 |
| 20+ | 206.1 | 254.1 | 712.0 | 194.3 | 474.9 | 190.0 | 241.1 |

more frequently cited than those in other subfields. However, when we consider subfield 7, the differences between various citation rates are very small. Subfield 7 for CR_AJ takes the first or second place for publication years 2010, 2011, and 2012, whereas subfield 7 for other citation rates only achieves a medium ranking.

This leads us to the conclusion that the impact on different citation rates varies within the same subfield. Crespo et al. (2014) reported that citation data depends on the field to which it belongs. Bornmann et al. 2012 also proved that the chance of a paper being cited is strongly related to the different subfields of chemistry. Our result is consistent with their results.

*Factor 2: Type of venue publication*

In Tables 5, 6 and 7, we regard journals as the base factor. Holding all other variables constant, for CR_AC, CR_BC, CR_CC, a paper published in a conference could significantly increase these citation rates. This result tells us that someone who wants to publish a paper in a conference publication, especially CCF classification A, may cite more papers published in conference publications. Similar to this result, a paper published in a conference could decrease CR_BJ, CR_CJ. However, for CR_AJ, a paper published in a conference can increase the citation rate by 1.1, 1.9, and 6.2%, respectively, for publication years 2010, 2011, and 2012. This reveals that the impact of conference publications on papers in top-ranking journals is slightly greater than that of journal publications.

Rahm and Thor (2005) found that the conference papers had a larger average number of citations than journal papers using two main conference databases (SIGMOD and VLDB) and three journal databases (TODS, VLDB Journal, Sigmod Record) over a period of 10 years. This result is different from our result for CR where journal paper attracts more citation rate when controlling other variables constant. But we can also see that their result is consistent with our results for CR_AJ, CR_AC, CR_BC and CR_CC. As we mentioned before, some researchers believe that journal publications generally enjoy a higher status than conference publications (Freyne et al. 2010; Franceschet 2010). Overall, we can conclude that journal and conference relative status varies for the CR in different categories.

*Factor 3: Classification of publications of CCF*

We take classification A as the base. As expected, Tables 5, 6 and 7 indicate that the higher the classification of a publication, the higher the citation rate. However, there is a

**Table 7** NBRM: Percentage change in expected citation rates compared with base for the set of publications from publication year 2012

| Factors | CR | CR_AJ | CR_AC | CR_BJ | CR_BC | CR_CJ | CR_CC |
|---|---|---|---|---|---|---|---|
| *Computer science subfield* | | | | | | | |
| Computer systems and high-performance computing | Base | | | | | | |
| Computer networks | 16.8 | 76.9 | 55.3 | −10.9 | −47.6 | 9.2 | 10.7 |
| Network and information security | 16.9 | 40.2 | 72.3 | −25.2 | −22.8 | −21.1 | −18.8 |
| Software engineering/software/programming language | −4.7 | −38.3 | 8.3 | 1.3 | −25.4 | −30.0 | −34.1 |
| Databases, data mining, and information retrieval | 6.1 | −14.6 | 77.8 | 5.2 | −27.9 | 2.3 | −45.2 |
| Theoretical computer science | −15.8 | −5.6 | 59.8 | −33.7 | 3.4 | −32.3 | −34.3 |
| Computer graphics and multimedia | −11.7 | 159.9 | 8.7 | 1.4 | −61.2 | 0.1 | −35.4 |
| Artificial intelligence and pattern recognition | 38.4 | 154.6 | 274.8 | 28.8 | −0.1 | 66.7 | 6.1 |
| Human–computer interaction and ubiquitous computing | −1.6 | −27.6 | 111.0 | −67.4 | −45.4 | −48.3 | −33.9 |
| *Type of publication* | | | | | | | |
| Journal | Base | | | | | | |
| Conference | −34.9 | 6.2 | 277.0 | −50.9 | 206.7 | −68.4 | 70.3 |
| *Classification of publication* | | | | | | | |
| A | Base | | | | | | |
| B | −43.8 | −73.6 | −74.9 | −27.0 | −52.5 | −43.3 | −51.0 |
| C | −63.1 | −91.1 | −93.3 | −71.2 | −85.3 | −17.1 | −69.0 |
| *Annual average number of papers published by the publication* | | | | | | | |
| (0, 40] | Base | | | | | | |
| (40, 55] | −11.4 | −37.3 | −37.7 | −7.0 | −17.5 | −7.4 | −22.4 |
| (55, 80] | −19.3 | −32.1 | −41.9 | 2.6 | −8.5 | −1.4 | −9.3 |
| (80, 100] | −17.8 | −24.2 | −44.5 | 9.7 | −40.6 | 2.0 | −23.1 |
| (100, 130] | −26.5 | −33.3 | −58.4 | −5.4 | −29.5 | −8.6 | −21.8 |
| (130, 160] | −16.3 | −22.6 | −59.4 | −10.6 | −39.2 | 13.7 | −26.6 |
| (160, 215] | −21.3 | −24.5 | −61.9 | 23.9 | −37.9 | 24.9 | −12.0 |
| (215, 300] | −32.3 | −26.9 | −68.4 | −12.7 | −68.1 | −22.4 | −31.7 |
| (300, 400] | −18.0 | −22.8 | −65.4 | 53.8 | −63.9 | 66.1 | −35.7 |
| 400+ | 0.5 | −17.4 | −58.6 | 6.3 | −60.0 | 61.4 | −32.9 |
| *Number of authors* | | | | | | | |
| 1 | Base | | | | | | |
| 2 | −0.5 | −18.8 | −2.2 | −27.9 | 27.7 | −18.9 | −8.4 |
| 3 | 7.8 | −12.8 | 12.3 | −21.5 | 27.5 | −10.8 | 2.6 |
| 4 | 11.1 | −7.0 | 16.8 | −22.9 | 31.0 | −4.7 | 12.3 |
| 5 | 9.6 | 1.0 | 18.7 | −23.9 | 37.1 | −11.3 | 17.1 |
| 5+ | 37.1 | 0.8 | 38.3 | −18.8 | 57.7 | 5.7 | 34.5 |
| *Maximum h-index of all authors of a paper* | | | | | | | |
| 0–1 | Base | | | | | | |
| 2 | 32.2 | 13.0 | 41.2 | 27.3 | 56.5 | 31.4 | 44.9 |
| 3–4 | 35.2 | 20.6 | 30.8 | 41.6 | 79.1 | 42.8 | 40.2 |

**Table 7** continued

| Factors | CR | CR_AJ | CR_AC | CR_BJ | CR_BC | CR_CJ | CR_CC |
|---|---|---|---|---|---|---|---|
| 5 | 38.3 | 35.3 | 57.7 | 42.7 | 124.6 | 43.3 | 58.5 |
| 6–7 | 53.0 | 64.4 | 82.7 | 81.2 | 172.9 | 60.1 | 79.0 |
| 8–9 | 64.1 | 108.3 | 118.2 | 95.2 | 254.5 | 78.6 | 110.3 |
| 10–11 | 77.2 | 120.7 | 178.6 | 98.8 | 280.9 | 64.9 | 123.7 |
| 12–15 | 99.6 | 149.5 | 214.0 | 137.7 | 416.4 | 87.3 | 136.2 |
| 16–20 | 132.0 | 180.4 | 318.2 | 172.6 | 501.7 | 95.9 | 189.8 |
| 20+ | 186.8 | 219.6 | 492.9 | 203.6 | 700.0 | 133.6 | 216.5 |

notable exception in that CR_CJ for classification C is higher than classification B. This means that, from classification B to C, with all other variables constant, CR_CJ is expected to increase.

It has been proved that the reputation for place of publications is one of the most strongly influencing factors (Didegah and Thelwall 2013; Peng and Zhu 2012). Our result for classification of publications of CCF is general consistent with their findings.

*Factor 4: Annual average number of papers published by the publication*

As mentioned in section "Regression analysis", the continuous variable avgPubCount has been converted into the categorical variable cat_avgPubCount. The results in Tables 5, 6 and 7 indicate that the increment in the annual average number of papers published by a publication does not effectively lead to any augmentation in citation rates.

For CR, we observe that a paper in cat_avgPubCount 1 ($\text{avgPubCount} \leq 40$) and in cat_avgPubCount 10 ($\text{avgPubCount} > 400$) can have higher CR than the other sets (cat_avgPubCount 2–8). For CR_AJ and CR_AC, overall, with the increase in avgPubCount, CR_AJ and CR_AC decrease, but there is a local increase in some cases. The percentage change varies between 40% for CR_AJ and 80% for CR_AC. For CR_BC and CR_CC, they have similar situations with CR_AJ and CR_AC. However, for CR_BJ and CR_CJ, overall, with the increase in avgPubCount, CR_BJ and CR_BC increase and there is also a local decrease in some cases. The percentage change for CR_CC is more obvious than CR_BC.

*Factor 5: Number of authors*

Contrary to what was expected from section "Factor description" for this factor, the performance varies with different citation rates and different publication years. We do see a general increase in citation rate with the number of authors once a paper has at least two authors. Indeed, there are specific cases where the citation rate decreases when the number of authors increases.

Based on the results in Tables 5, 6 and 7, for publication year 2010, overall, CR, CR_AJ, CR_BJ, CR_CJ and CR_CC increase with the increase of number of authors. CR_AC and CR_BC decreases with the increase of number of authors. For publication 2011, overall, all the CR in different categories increase with increase of number of authors, especially CR_CC grows the most. For publication 2012, all the CR in different categories also increase with increase of number of authors once a paper has at least two

authors. Thus the contribution of number of authors for CR in different categories and different publication years is distinct. From our experiment, the number of authors to predict the citation impact is really weak.

Many studies have reported a positive correlation between the number of authors and the citation rates of articles (Aksnes 2003; Leimu and Koricheva 2005; Fanelli 2013; Rigby 2013). However, some studies demonstrated that the ability of the number of authors to predict the citation impact of articles is weak (Bornmann and Daniel 2008; Van Dalen and Henkens 2001) which is consistent with our results.

*Factor 6: Maximum h-index of all authors of a paper*

From Tables 5, 6 and 7 for cat_author_max_h_index, the results verify our assumption that the greater the reputation of the authors of a paper, the higher the citation rate will be. This conclusion holds for all citation rates. For a base of author_max_h_index = 0 or 1, indicating that the authors may be early career researchers, the citation rates are usually very low compared to other papers for which author_max_h_index is relatively high.

Many previous studies have already demonstrated that h-index is significant predictor for citation rates (Wang et al. 2011, 2012; He 2009). Our result for this factor is highly consistent with their studies.

## Conclusions

In this paper, we have presented a bibliometric study of citation rates from different CCF classifications and venue types (CR, CR_AJ, CR_AC, CR_BJ, CR_BC, CR_CJ, CR_CC) for sets of publications from 2010, 2011, and 2012 in the field of computer science. We applied negative binomial regression models to study the effect of various factors on these citation rates. Contrary to previous studies, we did not base our approach on the total citation count of a paper. Instead, we divided the citation rate into different categories, which gave us a wider perspective. This made the bibliometric study more meaningful and profound. In addition, our dataset included nine subfields of computer science, making this a rigorous overall examination of this field.

To identify the impact on different citation rates in different subfields of computer science, we examined six factors: (1) subfield; (2) type of publication; (3) classification of publications in CCF; (4) the annual average number of papers published by a publication; (5) the number of authors; and (6) the maximum h-index of all authors. With our NBRM results, we can not only answer the two questions posed in section "Introduction" (Q1: relative status of journal and conference publications in computer science and Q2: whether the performance of citation data varies greatly in different subfields of computer science), but also clarify the effect of four factors on the citation data. A detailed analysis of these six factors was presented in section "Results".

To summarize: (1) for Q1, a conference publication's impact is greater than that of a journal publication when taking into account conference citation rates (CR_AC, CR_BC, CR_CC). Similarly, journal publications have a greater impact than conference publications when taking into account journal citation rates (CR_BJ, CR_CJ), although this is not true for CR_AJ. Therefore, the relative status of the journal and conference depends on what kind of citation rate we use as a measure. However, we can still agree that conferences enjoy a very high status in computer science, as the impact on conferences of all

classifications and journals of classification A is better than that of journals. (2) Regarding Q2, as expected, citation rate data varies greatly in different subfields of computer science. One subfield can be better than other subfields for one kind of citation rate, but worse than other subfields for other kinds of citation rate. We also noticed that subfield 8 (Artificial intelligence and pattern recognition) has a stable and strong effect on different kinds of citation rates compared with other subfields. Subfields 4 (Software engineering/software/programming language), 5 (Databases, data mining, and content retrieval), and 7 (Computer graphics and multimedia) also exhibit stable performance on different kinds of citation rates, but always rank at the medium level of all nine subfields. The other subfields fluctuate regarding the choice of citation rate. Besides the subfield and type, we also compared the base of four categorical factors while fixing all other independent variables.

Some scholars have indicated that the scientific publications have different citation lifecycle (Wang et al. 2013), which may have different distribution and citation style. We have attempted to conduct a citation regression analysis of computer science publications in different ranking categories and subfields in an objective way. Considering the subfield, venue and publication type (i.e., journal vs conference), it is deserved to study in detail the citation trend in computer science area. Our work opens several interesting new directions for future work. It is possible, for example, to consider other factors that determine the citation rate, such as the number of references in a paper and the distribution of quality of references; the ability of a paper and a pdf link to be found in Google Scholar; and author affiliations. It would also be interesting to study the cross-citation rate among different subfields to identify cross-subfield collaboration in computer science.

**Acknowledgements** This work was partially supported by the State Key Laboratory of Software Development Environment of China (No. SKLSDE-2017ZX-15), the National Social Science Foundation of China (No. 13&ZD190), an Royal Society-Newton Advanced Fellowship Award, and the Fundamental Research Funds for the Central Universities.

## References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159–170.

Beel, J., & Gipp, B. (2010). Academic search engine spam and google scholar's resilience against it. *Journal of Electronic Publishing*, *13*(3). http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0013.305.

Bensman, S. J., Smolinsky, L. J., & Pudovkin, A. I. (2010). Mean citation rate per article in mathematics journals: Differences from the scientific model. *Journal of the American Society for Information Science and Technology*, *61*(7), 1440–1463.

Berk, R., & MacDonald, J. M. (2008). Overdispersion and poisson regression. *Journal of Quantitative Criminology*, *24*(3), 269–284.

Bornmann, L., & Daniel, H. D. (2008). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by angewandte chemie international edition, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology*, *59*(11), 1841–1852.

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*(1), 11–18.

Bruin, J. (2006). *Newtest: Command to compute new test*. Los Angeles: UCLA: Academic Technology Services, Statistical Consulting Group.

Butler, D. (2008). Free journal-ranking tool enters citation market. *Nature News*, *451*(7174), 6–6.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge: Cambridge University Press.

Chen, J., & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM*, *53*(6), 79–83.

Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2014). The effect on citation inequality of differences in citation practices at the web of science subject category level. *Journal of the Association for Information Science and Technology*, *65*(6), 1244–1256.

Dalgaard, P. (2008). Rates and poisson regression. In *Introductory statistics with R* (pp. 259–274). Berlin: Springer.

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064.

Eckmann, M., Rocha, A., & Wainer, J. (2011). Relationship between high-quality journals and conferences in computer vision. *Scientometrics*, *90*(2), 617–630.

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics*, *9*(4), 895–906.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of scimago journal rank indicator with journal impact factor. *The FASEB Journal*, *22*(8), 2623–2628.

Fanelli, D. (2013). Positive results receive more citations, but only in some disciplines. *Scientometrics*, *94*(2), 701–709.

Fortnow, L. (2009). Viewpoint time for computer science to grow up. *Communications of the ACM*, *52*(8), 33–35.

Franceschet, M. (2010). The role of conference publications in CS. *Communications of the ACM*, *53*(12), 129–132.

Freyne, J., Coyle, L., Smyth, B., & Cunningham, P. (2010). Relative status of journal and conference publications in computer science. *Communications of the ACM*, *53*(11), 124–132.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama*, *295*(1), 90–93.

He, Z. L. (2009). International collaboration does not have greater epistemic authority. *Journal of the American Society for Information Science and Technology*, *60*(10), 2151–2164.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. College Station: Stata Press.

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*, *63*(11), 2140–2145.

Marx, W., & Bornmann, L. (2014). On the causes of subject-specific citation rates in web of science. *Scientometrics*, *102*(2), 1823–1827.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, *66*(4), 739–764.

Peng, T. Q., & Zhu, J. J. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. *Journal of the American Society for Information Science and Technology*, *63*(9), 1789–1803.

Rahm, E., & Thor, A. (2005). Citation analysis of database publications. *ACM Sigmod Record*, *34*(4), 48–53.

Rigby, J. (2013). Looking for the impact of peer review: Does count of funding acknowledgements really predict research impact? *Scientometrics*, *94*(1), 57–73.

Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, *45*(1), 1–11.

Shamir, L. (2010). The effect of conference proceedings on the scholarly communication in computer science and engineering. *Scholarly and Research Communication*, *1*(2). http://www.srconline.ca/index.php/src/article/viewFile/25/43.

Smolinsky, L. (2016). Expected number of citations and the crown indicator. *Journal of Informetrics*, *10*(1), 43–47.

StataCorp (2005). *Stata Statistical Software Release 9*. College Station: StataCorp LP.

Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, pp 990–998.

Thelwall, M., & Fairclough, R. (2015). The influence of time and discipline on the magnitude of correlations between citation counts and quality scores. *Journal of Informetrics*, *9*(3), 529–541.

Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, *8*(4), 963–971.

Van Dalen, H., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics*, *50*(3), 455–482.
Van Dalen, H. P., & Kn, H. (2005). Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics*, *64*(2), 209–233.
Vardi, M. Y. (2009). Conferences vs. journals in computing research. *Communications of the ACM*, *52*(5), 5.
Vrettas, G., & Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology*, *66*(12), 2674–2684.
Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics*, *69*(3), 499–510.
Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.
Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. *Scientometrics*, *87*(3), 695–706.
Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics*, *93*(3), 635–644.
Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427.
Zitt, M. (2012). The journal impact factor: Angel, devil, or scapegoat? A comment on JK Vanclay's article 2011. *Scientometrics*, *92*(2), 485–503.