Comparative Study of the Performance of Different Field Normalization Methods Across Different Field Classification Schemes: Postprint

Authors: Ren Yuanqiu, Wang Xing, Zheng Qinqin

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] To investigate the influence of different disciplinary classification schemes on the effectiveness of field normalization methods and to compare the performance of various normalization approaches. [Method/Process] Under the Web of Science disciplinary classification scheme, a comparative study was conducted on the effectiveness of three commonly used normalization methods: the mean-ratio method, median-ratio method, and Z-score method. By varying the granularity of disciplinary classification schemes, an empirical test was performed on the sensitivity of these three normalization methods under the Essential Science Indicators (ESI) and Organisation for Economic Cooperation and Development (OECD) disciplinary classification schemes. [Results/Conclusion] The results demonstrate that employing different disciplinary classification schemes did not substantially affect the effectiveness of the normalization methods, with their performance remaining essentially stable. From the perspective of CCDF citation distribution curves, the shapes of CCDF curves processed by the three normalization methods were significantly more compact compared to those of the original citation CCDF curves, and the citation distribution patterns remained largely similar across the three methods after changing disciplinary classification schemes of different granularities. Further quantitative examination using the top z% method reveals that the effects of the three normalization methods remained essentially unchanged after altering disciplinary classification schemes of different granularities, exhibiting the following patterns: when selecting papers below the global top 30%, the normalization effects of the mean-ratio method and Z-score method, though slightly different, both outperformed the median-ratio method; when selecting papers in the top 30%-40% range, the Z-score method demonstrated more prominent advantages; when selecting papers above the top 40%, the median-ratio method exhibited significantly superior performance compared to the other two methods.

Full Text

A Comparative Study of Field Normalization Method Effects Under Different Discipline Classification Schemes

Ren Yuanqiu[1], Wang Xing[1], Zheng Qinqin[2]
[1]School of Information, Shanxi University of Finance and Economics, Taiyuan 030006
[2]Urban Science Management Systems Application Consulting Co., Ltd., Shanghai 200120

Abstract: [Purpose/Significance] This study investigates the impact of different discipline classification schemes on field normalization method effects and compares the effectiveness of various field normalization methods. [Method/Process] Under the Web of Science classification scheme, we compared three commonly used normalization methods: the mean-based method, median-based method, and Z-score method. We then altered the granularity of discipline classification schemes to empirically test the sensitivity of these three methods under the Essential Science Indicators (ESI) and Organization for Economic Co-operation and Development (OECD) classification schemes. [Results/Conclusion] Results show that using different discipline classification schemes does not substantially affect the effectiveness of the normalization methods; their effects remain essentially unchanged. From the CCDF citation distribution curves, the shapes of curves processed by the three normalization methods are significantly more convergent than the original citation distribution curves, and the citation distributions remain largely similar after changing classification schemes. Combining this with quantitative examination using the top-z% method reveals that the effects of the three normalization methods remain basically unchanged across different classification granularities, following a clear pattern: when selecting papers below the global top 30%, both the mean-based and Z-score methods outperform the median method, though with slight differences between them; in the top 30%-40% range, the Z-score method shows notable advantages; and above top 40%, the median method demonstrates significantly superior performance compared to the other two.

Keywords: discipline classification scheme; field normalization; normalization effect; citation distribution

---

Citation counts are a crucial metric for measuring research impact. However, due to differences in citation practices and evolutionary patterns across disciplinary fields, raw citation counts cannot be directly compared across disciplines. For instance, biomedical fields experience rapid paper turnover and high publication volumes, resulting in higher citation counts, whereas mathematics and geology have longer research cycles and require more time to reach peak citation levels. To eliminate these disciplinary differences and enable fair evaluation of research impact across fields, scholars typically employ field normalization meth-

ods to mathematically transform citation counts, producing standardized scores that approximate a common distribution for cross-disciplinary comparison.

Numerous normalization methods have been proposed based on different theoretical foundations, including the mean-based method, median-based method, Z-score method, percentile method, citing-side normalization method, reverse engineering method, and commutative-based normalization method. However, most of these were developed based on subjective experience and intuition, leaving room for improvement toward the ideal state where normalized citation distributions become approximately identical across fields. Consequently, measuring and comparing normalization method effectiveness remains an active research area. Some scholars argue that since citation distributions are skewed, the median better represents central tendency than the mean, suggesting the median method outperforms the mean method. L. Bornmann et al. advocate using distribution position rather than simple parameters for evaluation, proposing percentile methods for research impact assessment. Zhang Zhihui et al. compared mean-based and Z-score methods, finding Z-score advantages in distribution tails but inferior performance in other regions, particularly low-citation portions, leading them to propose an optimal linear normalization method. Nevertheless, no consensus exists on which method best serves scientific research evaluation.

Normalization effectiveness also depends on various latent factors, particularly the specific discipline classification scheme employed. Discipline classification itself is complex, with different countries, regions, and databases maintaining distinct systems. The Centre for Science and Technology Studies (CWTS) at Leiden University uses Journal Citation Reports (JCR) subject categories as reference standards for citation normalization in its Crown Indicator. Clarivate' s Essential Science Indicators (ESI) database divides research into 22 fields and provides statistical analysis and ranking at national, institutional, journal, and author levels. These journal-based classification schemes have limitations, as interdisciplinary journals may be assigned to single categories. For example, *Computers & Mathematics with Applications* is classified under both "Mathematics, Applied" and "Computer Science, Interdisciplinary Applications" in Web of Science but assigned solely to "Mathematics" in ESI. Leydesdorff and Bornmann further note that Web of Science categories were developed for information retrieval rather than citation analysis, containing substantial overlap and inadequately handling interdisciplinary journals.

Some scholars have constructed custom classification systems. J. Ruiz-Castillo and L. Waltman developed a paper-level system with 5,119 research fields, while C. Collianer et al. proposed an "Item-oriented Approach" that extracts nouns and adjectives from titles and abstracts, lemmatizes them, and compares performance with algorithmically constructed clustering systems. However, in practical research evaluation, papers may be assigned to different fields depending on the classification scheme, potentially affecting normalization effectiveness. Studies have examined this sensitivity: M. Zitt et al. investigated how different

granularity levels affect the mean-based method, finding it sensitive to field granularity. J. Adams et al. validated these findings using UK research institutions, while W. Glänzel et al. analyzed European universities and proposed that 60 fields represent an optimal granularity for institutional evaluation. A. Perianes-Rodríguez and J. Ruiz-Castillo further suggested that higher-granularity classification systems generally exhibit better normalization performance, though these studies focused only on the mean-based method, lacking systematic comparison of multiple approaches.

This study addresses two research questions: (1) Under the widely used Web of Science classification scheme, systematically compare the effectiveness of mean-based, median-based, and Z-score methods, summarizing their citation distribution characteristics and relative performance; (2) By varying classification scheme granularity, test the sensitivity of these three methods under ESI and OECD schemes to explore how different classification systems influence normalization effects. This investigation ensures fair and accurate comparison of normalization effects, promotes method maturation, and enriches normalization research. Particularly under China' s "Double First-Class" university initiative, selecting appropriate classification schemes and normalization methods forms the foundation for effective research evaluation, enabling objective assessment of Chinese universities' research impact and their gap with world-class institutions.

2. Data and Methods

This study obtains citation data from the InCites database under three classification schemes—Web of Science, ESI, and OECD—and applies mean-based, median-based, and Z-score normalization methods. By comparing effectiveness across schemes, we examine relationships between classification systems and normalization outcomes, analyze reasons for different effects, and summarize usage patterns. The research design is illustrated in Figure 1 [Figure 1: see original paper].

2.1 Selection of Discipline Classification Schemes

We selected Web of Science, ESI, and OECD classification schemes based on three considerations: (1) All three are widely used in research evaluation, originate from the same citation system (Clarivate' s InCites database), ensuring data comparability and avoiding cross-system differences (e.g., between Google Scholar, Scopus, and Web of Science); (2) The schemes represent different granularities—255 categories in Web of Science, 45 in OECD, and 22 in ESI—providing sufficient distinction; (3) They are all accessible within InCites, enabling consistent data extraction.

2.2 Citation Data Collection

From InCites, we downloaded all 2013 articles under each classification scheme: 1,495,337 papers from Web of Science, 1,362,619 from ESI, and 1,495,258 from OECD. The citation window spans 2013–2019, providing six years of citation accumulation for stable, reliable data while avoiding short-window biases.

2.3 Field Normalization Methods

We compared three representative and operational methods: mean-based, median-based, and Z-score normalization.

The mean-based method calculates standardized scores as the ratio of raw citation count to the mean citation count within the same field:

$$m = \frac{c}{\mu}$$

where m is the standardized score, c is raw citation count, and μ is the mean citation count.

The median-based method similarly uses the ratio to the field median:

$$m = \frac{c}{M}$$

The Z-score method incorporates both central tendency and dispersion:

$$z = \frac{c - \mu}{\sigma}$$

where σ is the standard deviation of citation counts in the field.

2.4 Normalization Effectiveness Testing

We assessed effectiveness by examining whether normalized citation distributions approximate a common distribution, using two established methods: CCDF (Complementary Cumulative Distribution Function) plots and the top-z% method, widely adopted in prior research. CCDF plots display standardized score distributions across fields—greater overlap indicates better normalization. The top-z% method sorts all papers by standardized score, selects the global top z%, calculates each field' s actual proportion within this subset, and compares it to the expected proportion z%. More uniform distributions (smaller deviations from z%) indicate better normalization.

For a given paper, the CCDF value represents the proportion of papers in its field with citation counts greater than or equal to its own. For example, if field M has the citation distribution shown in Table 1 , plotting points (0,1.0), (1,0.4),

and (2,0.1) and connecting them yields the CCDF curve (Figure 2 [Figure 2: see original paper]).

The top-z% method uses mean deviation (MD) to quantify effectiveness:

$$MD = \frac{1}{n}\sum_{i=1}^{n}|p_i - p|$$

where p_i is the actual proportion from field i, p is the expected proportion z%, and n is the number of fields. Smaller MD values indicate better normalization, with MD = 0 representing ideal performance.

2.5 Statistical Analysis

We compiled citation counts across the three classification schemes, calculated descriptive statistics (means, medians, standard deviations), and used Python to generate distribution plots for effectiveness testing.

3. Results and Discussion

3.1 Comparison of Three Normalization Methods Under Web of Science

Python-generated CCDF plots for Web of Science (Figure 3 [Figure 3: see original paper]) show that normalized distributions are substantially more convergent than original citation distributions. Although complete ideal convergence is not achieved, normalization effects are evident, particularly in high-citation tails where curves show strong overlap. The median method performs well in low-citation regions above CCDF = 0.5, while Z-score shows advantages in the "waist" region near (0, 0.4).

To examine subtle differences across citation levels, we applied the top-z% method at 12 thresholds: z = 1%, 2%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%. Table 2 presents MD values.

For papers below the top 5%, Z-score performs best. Between top 5%-20%, Z-score is slightly inferior to the mean method, but both outperform the median method. In the top 30%-40% range, Z-score shows clear advantages, corresponding to the "waist" region in CCDF plots. Above top 40%, the median method becomes superior, with particularly strong performance above top 50%, consistent with CCDF patterns where curves begin converging around CCDF = 0.5.

3.2 Comparison Under ESI Classification Scheme

To test whether these patterns hold across classification granularities, we examined ESI scheme performance. Figure 4 [Figure 4: see original paper] shows that normalized distributions are again more convergent than original distributions.

Compared to Web of Science, Z-score' s "waist" effect is more pronounced, and the median method shows convergence trends beginning at CCDF = 0.4.

Table 3 presents top-z% results. The mean method shows smallest MD values below top 30%, while Z-score and median methods also perform well (MD < 0.01) at top 1%-2%. At top 30%, median and Z-score methods begin outperforming the mean method. Above top 40%, the median method' s superiority becomes pronounced, even appearing at top 30% in ESI. While minor differences exist compared to Web of Science, overall patterns remain consistent across classification schemes.

3.3 Comparison Under OECD Classification Scheme

OECD scheme results (Figure 5 [Figure 5: see original paper]) again show normalized distributions are more convergent than original distributions, with patterns similar to Web of Science and ESI. Table 4 confirms that Z-score performs well in top 1% and top 30%-40% ranges, while the median method shows advantages above top 40%, consistent with previous findings. No substantial changes occur across classification schemes.

3.4 Discussion

Although normalization effectiveness must be implemented within specific classification schemes, changing classification granularity does not substantially alter method performance. CCDF curves after normalization are clearly more convergent than original distributions, and these patterns persist across schemes. Quantitative top-z% analysis confirms that method effects remain basically unchanged, following consistent patterns: for papers below top 30%, mean and Z-score methods outperform the median method; in the top 30%-40% range, Z-score shows particular advantages; above top 40%, the median method is significantly superior.

These patterns likely stem from method characteristics and citation distribution properties. The mean-based method uses the average to reflect basic distribution features but cannot capture dispersion or positional differences. The median method, based on positional representation, improves representativeness for the majority of low-to-medium citation papers but lacks sensitivity to extreme values, limiting high-citation performance. Z-score' s advantage in the top 30%-40% range corresponds to the proportion of papers exceeding the mean in most fields.

These findings have practical implications for evaluating Chinese universities' research impact. Different normalization methods can be applied to different citation levels: mean or Z-score for high-citation papers, median for low-citation papers. Aggregating standardized scores can effectively identify high-performing institutions and guide healthy research development.

Limitations include: (1) We examined only three commonly used classification

schemes from InCites; future research should include more granularities; (2) We did not examine Scopus or Google Scholar schemes due to non-source paper citation differences—addressing these differences represents a future direction. Additionally, incorporating newly proposed normalization methods would provide more comprehensive guidance for "Double First-Class" university evaluation.

References

[1] Zhang Zhihui, Cheng Ying, Liu Niancai. Optimizing linear field normalization methods and their impact on research evaluation results: Ranking 39 "985 Project" universities by paper quality[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(3): 300-312.

[2] Radicchi F, Fortunato S, Castellano C. Universality of citation distributions: toward an objective measure of scientific impact[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(45): 17268-17272.

[3] Waltman L, van Eck N J, van Leeuwen T N, et al. Towards a new crown indicator: an empirical analysis[J]. Scientometrics, 2011, 87(3): 467-481.

[4] Waltman L, van Eck N J, van Leeuwen T N, et al. Towards a new crown indicator: some theoretical considerations[J]. Journal of informetrics, 2011, 5(1): 37-47.

[5] Wang X, Zhang Z. Improving the reliability of short-term citation impact indicators by taking into account the correlation between short- and long-term citation impact[J]. Journal of informetrics, 2020, 14(2): 101019.

[6] Leydesdorff L, Opthof T. Remaining problems with "New Crown Indicator" (MNCS) of the CWTS[J]. Journal of informetrics, 2011, 5(1): 224-225.

[7] Vaccario G, Medo M, Wider N, et al. Quantifying and suppressing ranking bias in a large citation network[J]. Journal of informetrics, 2017, 11(3): 766-782.

[8] Bornmann L. How to analyze percentile impact data meaningfully in bibliometrics: the statistical analysis of distributions, percentile rank classes and top-cited papers[J]. Journal of the American Society for Information Science and Technology, 2013, 64(3): 587-595.

[9] Leydesdorff L, Opthof T. Normalization at the field level: fractional counting of citations[J]. Journal of informetrics, 2010, 4(4): 644-646.

[10] Zitt M, Small H. Modifying the journal impact factor by fractional citation weighting: the audience factor[J]. Journal of the American Society for Information Science and Technology, 2008, 59(11): 1856-1860.

[11] Radicchi F, Castellano C. A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions[J]. PLOS ONE, 2012, 7(3): e33833.

[12] Crespo J A, Li Y, Ruiz-Castillo J. The measurement of the effect on citation inequality of differences in citation practices across scientific fields[J]. Journal of informetrics, 2013, 7(1): 228-230.

[13] Bornmann L. Toward an ideal method of measuring research performance: some comments to the Opthof and Leydesdorff (2010) paper[J]. Journal of informetrics, 2011, 5(1): 224-225.

[14] Bornmann L, Mutz R. Further steps toward an ideal method of measuring citation performance: the avoidance of citation (ratio) averages in field-normalization[J]. Journal of informetrics, 2011, 5(1): 228-230.

[15] Calver M C, Bradley J S. Should we use the mean citations per paper to summarize a journal' s impact or to rank journals in the same field?[J]. Scientometrics, 2009, 81(3): 611-615.

[16] Bornmann L, Mutz R, Neuhaus C, et al. Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results[J]. Ethics in Science and Environmental Politics, 2008, 8(1): 93-102.

[17] Zhang Zhihui. Research on linear field normalization methods for paper impact[D]. Shanghai: Shanghai Jiao Tong University, 2015.

[18] Chen Shiji, Shi Liwen, Zuo Wenge. Identification methods and empirical analysis of potential disciplines in research institutions: a case study of China Agricultural University[J]. Journal of Intelligence, 2012, 31(2): 43-49.

[19] Van Raan A F J. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments[J]. Assessment theory and practice, 2003, 1(12): 20-29.

[20] Wang Yingxin, Huang Delong, Liu Dehong. ESI indicator principles and calculation[J]. Library and Information Service, 2006, 50(9): 73-75.

[21] Hu Z, Tian W, Xu S, et al. Four pitfalls in normalizing citation indicators: an investigation of ESI' s selection of highly cited papers[J]. Journal of informetrics, 2018, 12(4): 1133-1145.

[22] Leydesdorff L, Bornmann L. The operationalization of "fields"as WoS subject categories (WCs) in evaluative bibliometrics: the cases of "library and information science" and "science & technology studies" [J]. Journal of the Association for Information Science and Technology, 2016, 67(3): 707-714.

[23] Ruiz-Castillo J, Waltman L. Field-normalized citation impact indicators using algorithmically constructed classification systems of science[J]. Journal of informetrics, 2015, 9(1): 102-117.

[24] Collianer C. A novel approach to citation normalization: a similarity-based method for creating reference sets[J]. Journal of the Association for Information Science and Technology, 2015, 66(3): 489-500.

[25] Collianer C, Ahlgren P. Comparison of publication-level approaches to ex-post citation normalization[J]. Scientometrics, 2019, 120(1): 283-300.

[26] Zitt M, Ramanana-Rahary S, Bassecoulard E. Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalization[J]. Scientometrics, 2005, 63(2): 373-401.

[27] Adams J, Gurney K, Jackson L. Calibrating the zoom-atest of Zitt' s hypothesis[J]. Scientometrics, 2008, 75(1): 81-95.

[28] Glänzel W, Thijs B, Schubert A, et al. Subfield-specific normalized relative indicators and a new generation of relational charts: methodological foundations illustrated on the assessment of institutional research performance[J]. Scientometrics, 2009, 78(1): 165-188.

[29] Perianes-Rodríguez A, Ruiz-Castillo J. A comparison of the Web of Science and publication-level classification systems of science[J]. Journal of informetrics, 2017, 11(1): 32-45.

[30] Bar-Ilan J. Which h-index?—a comparison of WoS, Scopus and Google Scholar[J]. Scientometrics, 2013, 74(2): 257-271.

[31] Wang J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3): 851-872.

[32] Bornmann L, Daniel H D. Universality of citation distributions—a validation of Radicchi et al.' s relative indicator cf=c/c0 at the micro level using data from chemistry[J]. Journal of the American Society for Information Science and Technology, 2009, 60(8): 1664-1670.

[33] Radicchi F, Castellano C. Rescaling citations of publications in physics[J]. Physical review e, 2011, 83(4): 046116.

[34] Waltman L, van Eck N J, van Raan A F J. Universality of citation distributions revisited[J]. Journal of the American Society for Information Science and Technology, 2012, 63(1): 72-77.

[35] Zhang Z, Cheng Y, Liu N C. Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories[J]. Scientometrics, 2014, 101(3): 1679-1695.

[36] Zhang Z, Cheng Y, Liu N C. Improving the normalization effect of mean-based method from the perspective of optimization: optimization-based linear methods and their performance[J]. Scientometrics, 2015, 102(1): 587-607.

Author Contributions:

Wang Xing: Overall research design, experimental design, paper guidance, writing and revision;

Ren Yuanqiu: Data collection and processing, data analysis, experimental design, paper writing and revision;

Zheng Qinqin: Data processing and analysis.