Using the Coefficient of Variation to Identify Sleeping Beauty Literature in Postprint Research

Authors: Tang Jie, Zeng Jingjing

Date: 2023-04-01T16:02:48+00:00

Abstract

[Objective/Significance] This study reviews existing methods for identifying sleeping beauty papers, systematically examines their respective advantages and disadvantages, and proposes an improved identification approach that balances accuracy with operational simplicity.

[Method/Process] Building upon the well-established Bcp index identification method, this research adopts its core principle of leveraging the "degree of dispersion" in citation curves for identification. By introducing the statistical concept of "coefficient of variation" and applying it to differentiate among various citation curve patterns, we propose the PCV index as a novel approach for identifying sleeping beauty papers.

[Results/Conclusion] The empirical results demonstrate that the PCV index enables relatively simple and accurate identification of sleeping beauty papers, while exhibiting low dependence on total citation counts.

Full Text

Preamble

Identifying Sleeping Beauty Papers Using the Coefficient of Variation

Tang Jie[1],[2],[3], Zeng Jingjing[1],[2]

[1]Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000
[2]Lanzhou Information Center, Chinese Academy of Sciences, Lanzhou 730000
[3]University of Chinese Academy of Sciences, Beijing 100049

Abstract: [Purpose/Significance] This paper reviews existing methods for identifying sleeping beauty papers, summarizes their respective strengths and weak-

nesses, and proposes an improved identification method that balances accuracy and operational simplicity. [Method/Process] Building upon the well-developed Bcp index identification method and drawing on its core concept of utilizing the "dispersion degree" of citation curves, this study introduces the statistical concept of "coefficient of variation" to distinguish different types of citation curves, thereby proposing the PCV index for identifying sleeping beauty papers. [Result/Conclusion] The identification results demonstrate that the PCV index can effectively and accurately identify sleeping beauty papers with relatively low dependence on total citation counts.

Keywords: Sleeping beauty paper; Citation curve; Coefficient of variation
Classification Number: G252
DOI: 10.13266/j.issn.0252-3116.2021.06.010

The lifecycle and aging patterns of scientific literature constitute an important component of science communication research. Generally, papers are cited by other works within a few years after publication, gradually reaching a citation peak before declining until they are no longer cited [1]. However, scholars have discovered a category of papers that receive few citations initially but experience a sudden surge after a dormant period. The metrologist A.F.J. van Raan [2] termed these "sleeping beauties in science" and proposed three metrics—sleeping duration, sleep depth, and awakening intensity—to characterize them, after which this phenomenon became subject to quantitative and standardized investigation.

The essence of the "sleeping beauty" phenomenon lies in its content representing transformative or forward-looking research [3]. Identifying sleeping beauty papers helps improve scientific evaluation systems, encourages innovative research, and furthers understanding of scientific information flow mechanisms while discovering potential innovation points. This endows sleeping beauty papers with significant research value and makes their identification an important research topic in library and information science.

Existing identification methods can be categorized into three types [4]: (1) Curve fitting methods, which use mathematical expressions or appropriate curve types to fit the annual citation distribution of individual papers for identification [1, 5]; (2) Subjective indicator methods, which set indicators and manually determine thresholds to judge whether a paper qualifies as a sleeping beauty [2, 6-7]; and (3) Objective indicator methods, which use metric values to measure the degree to which a paper can be considered a sleeping beauty, thereby eliminating the arbitrariness of subjective threshold setting [8-15].

Previous surveys reveal that when identifying sleeping beauty papers within specific disciplines, subjective indicator methods are used far more frequently than other approaches due to their operational simplicity and speed. However, they require artificial threshold setting, exhibit strong subjectivity, and often lead to incomplete identification [10]. In contrast, objective indicator methods and curve fitting methods avoid arbitrariness in defining identification standards

and yield more accurate results, though their calculation processes are more complex. Therefore, this study attempts to further explore sleeping beauty identification methods that balance accuracy and operational simplicity.

2 Methodology

In recent years, identification methods for sleeping beauty papers have evolved from subjective to objective indicators [16], and J. Li and F.Y. Ye [17] have noted that threshold setting should be avoided when identifying sleeping beauty papers. Against this backdrop, this paper draws on the concept of the Bcp index [12] to propose a new identification method.

2.1 Bcp Index

The Bcp index (Formula 1) represents a mature objective identification method developed through refinements of the B index [10] and SBc index [11]. As illustrated in Figure 1 [Figure 1: see original paper], the reference line l connects the point (0, C_0) representing citations in the publication year to the point (t , 1) where the cumulative percentage of annual citations reaches 1. In Formula (1), (1-C_0)/t represents the slope of reference line l. For any t < t , the difference between l and C is calculated, and these differences are summed from t=0 to t=t .

$$Bcp = \sum_{t=0}^{t_m} \left(\frac{1 - C_0}{t_m} \cdot t + C_0 - C_t \right) \quad \text{Formula (1)}$$

The core concept of the Bcp index identification method lies in calculating the dispersion degree of citation curves. Sleeping beauty papers exhibit the key characteristic of low initial citations followed by sudden high citations later, resulting in high dispersion in their cumulative citation curves [18]. Therefore, this study introduces the statistical concept of "coefficient of variation" to explore different citation curve types and their dispersion degrees for sleeping beauty identification.

2.2 Coefficient of Variation

The coefficient of variation (CV), also known as the relative standard deviation, is a normalized measure of data distribution dispersion [19], defined as the ratio of standard deviation to mean. The calculation formula is:

$$CV = \frac{\sigma}{\mu} \quad \text{Formula (2)}$$

where σ represents standard deviation and represents mean.

As a dimensionless quantity, the coefficient of variation eliminates differences between numerical values at different scales during calculation, providing strong comparability between data. This ensures that when constructing a citation curve research framework, the simulated citation quantities set for different curve types will not affect research results, allowing focus to remain on citation curve morphology.

2.3 Coefficient of Variation for Different Citation Curves

Citation curves, also known as citation patterns, citation histories, or citation lifecycles [1], graphically describe the temporal distribution of citation counts. Numerous scholars have summarized citation curve types [1, 4, 20-25] (see Table 1).

Table 1 reveals that classification standards in relevant research emphasize different aspects. Some studies categorize citation curve types based on absolute citation counts, while others distinguish curve morphology based on relative citation counts. Considering that sleeping beauty papers themselves are highly cited [6, 26-29] and to comprehensively understand the dispersion degrees of different citation curve types, this study adopts the following research framework after summarizing the citation curve types in Table 1: (1) Classic type: Papers conforming to general literature aging patterns, accumulating most citations early and reaching a citation peak, after which citations gradually decline; (2) Flash-in-the-pan type: Papers that quickly reach a citation peak after publication, followed by rapid citation decline; (3) Exponential growth type: Papers with continuously increasing annual citations after publication, also known as genius papers, which are relatively rare [24]; (4) Sleeping beauty type: Papers with low initial citations that surge after a dormant period, also called delayed recognition type; (5) Multi-peak type: Papers with multiple peaks in citation history, also known as fluctuating type.

After establishing the research framework, this study simulates the five citation curves. Assuming five papers published in 2000 with 300 total citations by 2019, annual citation curves (Figure 2 [Figure 2: see original paper]) and cumulative citation curves (Figure 3 [Figure 3: see original paper]) are drawn according to each type's characteristics, with their coefficients of variation calculated (see Table 2).

The test results in Table 2 show that for annual citation curves, sleeping beauty and flash-in-the-pan types exhibit higher CV values than other types. However, flash-in-the-pan papers show significantly lower cumulative citation CV values. For cumulative citation curves, both sleeping beauty and exponential growth types have high CV values. This indicates that relying solely on one type of citation curve CV may confuse sleeping beauty papers with exponential growth or flash-in-the-pan papers. To identify sleeping beauty papers more accurately, we must capture their characteristic of having high CV values for both curve types. Therefore, this study uses the product of the two curve coefficients of

variation to measure the degree to which a paper can be considered a sleeping beauty, terming this metric the Product of Coefficients of Variation index (PCV index), calculated as:

$$PCV = CV_{yearly} \times CV_{cumulative} \quad \text{Formula (3)}$$

where CV_{yearly} represents the coefficient of variation for the annual citation curve and CV_{cumulative} represents that for the cumulative citation curve.

3 Empirical Study

3.1 Data Source Selection

This study uses the Web of Science (WoS) Core Collection as its data source, selecting papers with WoS category "Information Science & Library Science." To ensure a citation window of 15-25 years, publication years are limited to 1995-2004, with document type "Article," yielding 23,913 papers.

Since sleeping beauty papers are themselves highly cited [1, 26-29], this provides guidance for initial data screening. This study applies Price's Law (Formula 4) to identify highly cited papers [30]:

$$N = 0.749 \times \sqrt{n_{max}} \quad \text{Formula (4)}$$

where N represents the minimum citation count for highly cited papers and n_{max} represents the highest citation count in the collection. For 1995-2004, the most cited paper in this field received 8,606 citations, yielding N = 80.29. Ultimately, 1,098 papers with citation frequencies $ $81 were selected.

3.2 Identification Results

The citation data were processed and papers numbered using a combination of the last two digits of publication year and the paper's rank by citation count within that year (e.g., 95-1 represents the most cited paper published in 1995). PCV values were then calculated, with the distribution shown in Figure 4 [Figure 4: see original paper].

Figure 4 shows that the majority of papers have PCV values concentrated between 0.25 and 1.00, with a small portion below 0.25 or above 1.00. To validate the proposed method and following previous research [5], the TOP10 papers by PCV value were selected for further evaluation. Table 3 provides information on these TOP10 papers.

3.3 Effectiveness Evaluation

3.3.1 Validity of the PCV Index To verify whether papers identified by the PCV index exhibit sleeping beauty characteristics, validity testing is required.

Current validity testing for sleeping beauty identification methods 主要包括 two approaches: (1) citation curve effect analysis, and (2) comparison of identification result overlap rates with other methods. This study examines both aspects to test the PCV index method's validity.

(1) Citation Curve Effect Analysis. Observing citation curve morphology provides a simple and intuitive way to assess method validity. Figure 5 [Figure 5: see original paper] shows the citation curves for the TOP10 papers by PCV value.

Examination of these curves reveals that papers ranked 1-6 and 8, 10 (IDs: 02-99, 01-103, 02-61, 00-21, 99-22, 96-59, 01-123, 01-32) clearly demonstrate the pattern of low initial citations followed by sudden high citations, with all eight experiencing noticeable sleeping periods. Papers ranked 7 and 9 (IDs: 95-48, 99-35) show brief fluctuations during their sleeping periods, but their average annual citations remain low due to small increases and short duration. Overall, the TOP10 papers all exhibit fundamental sleeping beauty characteristics.

(2) Comparison of Identification Result Overlap Rates. Given the conceptual similarity between PCV and Bcp indices, their identification results were compared. Table 4 lists the TOP10 papers under both frameworks and each paper's ranking in the alternative framework.

Table 4 shows five overlapping papers between the two TOP10 lists, yielding a 50% overlap rate. Table 5 summarizes overlap rates from different identification methods in previous studies.

As Table 5 indicates, overlap rates vary significantly across methods, ranging from 0% to 75%. Scholars note this variation relates to method characteristics and sleeping beauty curve morphology [16]. With a 50% overlap rate in the TOP10 comparison, this study's results are relatively high compared to previous findings, suggesting the PCV index method is effective.

3.3.2 Differences Between PCV and Bcp Indices To further explore differences between PCV and Bcp indices in identifying sleeping beauty papers, the ten non-overlapping TOP10 papers from both frameworks were analyzed using six metrics drawn from relevant studies [2, 12]: (1) Publication duration: years from publication to 2019 (using 2019 as the cutoff since 2020 data remains incomplete); (2) Total citations: cumulative citations through 2019; (3) Average annual citations: ratio of total citations to publication duration; (4) Citation peak: maximum annual citation count; (5) Sleeping duration: years with average annual citations between 0-2, following van Raan's definition [2]; (6) Awakening intensity: average annual citations in the four years after sleep ends. Statistics are presented in Table 6 , with independent samples t-test results in Table 7 .

The t-test results reveal significant differences between papers identified by PCV and Bcp indices across publication duration, total citations, average annual ci-

tations, and citation peak. Papers ranking higher by Bcp index exceed those ranking higher by PCV index on all four metrics, indicating Bcp index is more sensitive to older, highly cited papers. In contrast, PCV index better identifies younger sleeping beauty papers. Additionally, sleeping durations differ significantly: Bcp-identified papers average 4.40 years versus 12.40 years for PCV-identified papers, suggesting PCV index identifies papers with more pronounced "sleeping" characteristics. No significant difference appears in awakening intensity.

In summary, the PCV index method serves as an effective complement to sleeping beauty identification systems.

4 Conclusions and Discussion

This study reviewed existing sleeping beauty identification methods, summarized their advantages and disadvantages, and proposed the PCV index to expand the identification methodology system. Drawing on the Bcp index's core concept of measuring citation curve dispersion, the PCV index considers both annual and cumulative citation curves, further reducing dependence on total citation counts. This enables more flexible identification of younger papers exhibiting sleeping beauty characteristics. Comparison with Bcp index results also reveals that PCV-identified sleeping beauty papers have longer sleep durations. Moreover, PCV index calculation, based on the coefficient of variation, is computationally simple. In conclusion, the PCV index represents an effective, flexible, and easy-to-operate method for identifying sleeping beauty papers.

However, the PCV index method has limitations. First, as an objective indicator method, it shares the common defect of being unable to absolutely demarcate boundaries between sleeping beauty and other paper types [10]. Second, while this study used highly cited papers as its data source, the coefficient of variation's properties and effectiveness evaluation results both demonstrate PCV index's minimal dependence on total citation counts. Consequently, Bcp index identifies papers with higher total citations and greater impact. Future research could address this by imposing stricter total citation thresholds during data screening. Finally, this study raises additional questions requiring further discussion.

4.1 Impact of Disciplinary Characteristics on Identification Effectiveness

Using highly cited papers in library and information science as samples, this study analyzed the TOP10 papers, whose PCV values range from 1.02 to 1.82—relatively dispersed values. As values decrease, papers exhibit weaker "sleeping beauty" characteristics. This phenomenon relates to sample characteristics: library and information science is not a field prone to top-tier sleeping beauty papers, and the relatively short citation window may also affect identification results. Furthermore, objective identification methods avoid numerical restric-

tions on citations during specific periods to eliminate subjectivity in threshold setting, but this also creates ambiguous boundaries between sleeping beauty and other papers, necessitating artificial demarcation during screening. This study selected TOP10 as its criterion, but whether this standard applies to other disciplines given disciplinary differences and variations in sleeping beauty prevalence between humanities/social sciences and natural sciences requires verification.

4.2 Expanded Applications of Citation Curve Coefficients of Variation

Calculations for different citation curve types reveal that besides sleeping beauty papers, other special paper types exhibit distinctive CV values. For example, flash-in-the-pan papers have high initial citations that drop sharply, yielding high annual citation CV values. However, due to technological replacement or topic shifts, these papers are quickly forgotten [33], resulting in insufficient later citation growth and early stabilization of total citations, ultimately producing lower cumulative citation CV values. Future research could exploit these characteristics for identification purposes.

4.3 Differences Between PCV and Bcp Indices

Given conceptual similarities between PCV and Bcp indices, this study compared their identification overlap rates, achieving 50% overlap within the target range. Referencing previous overlap rate comparisons, this value is relatively high though slightly below expectations, warranting further investigation. Examining their calculations reveals that to avoid dependence on citation magnitude, Bcp index converts the annual citation curve's vertical axis to "cumulative percentage of annual citations." However, because its calculation involves summing distances from citation curve points to the reference line (see Formula (1)), the method is more sensitive to older papers, partially explaining why Bcp-identified papers have significantly longer publication durations. In contrast, PCV index comprises two citation curve coefficients of variation, focusing on "dispersion degree" and enabling identification of papers with more fluctuating citation curves. In summary, the two methods share conceptual similarities but exhibit objective differences.

4.4 Optimization of Sleeping Beauty Identification Methods

The identification methodology system continues expanding. However, different methods' various entry points and emphases, despite capturing sleeping beauty papers' basic characteristics, lead to persistent result variations. Moreover, scholars note that different stages of sleeping beauty citation curves are influenced by different factors, making their morphology diverse and complex [17]. Consequently, achieving perfect precision and recall is challenging. While objective indicator methods have gradually replaced subjective methods as the mainstream, practice reveals that artificial selection of TOPN boundaries remains necessary, demonstrating the need to combine objective indicators with

subjective judgment. Recent scholars have suggested combining existing methods to improve identification accuracy and comprehensiveness through mutual constraint and supplementation [17], but how to select methods for combination requires practical exploration based on each method's characteristics. Additionally, according to most scholars' practical preferences and considering Moore's Law and Zipf's Principle of Least Effort, method optimization should address not only accuracy but also operational simplicity. Future research should attempt to balance both aspects in sleeping beauty identification.

References

[1] Li Jiang, Jiang Mingli, Li Yuting. A framework for citation curve analysis: Evidence from Nobel laureates' citation curves [J]. Journal of Library Science in China, 2014, 40(2): 41-50.
[2] Raan AFJ van. Sleeping beauties in science [J]. Scientometrics, 2004, 59(3): 467-472.
[3] Du Jian, Sun Yinan, Zhang Yang, et al. Bibliometric characteristics of transformative research and early identification methods [J]. Bulletin of National Natural Science Foundation of China, 2019, 33(1): 88-98.
[4] Song Chengyu, Li Xiuxia, Liu Liming. Sleeping beauty paper identification based on citation curve derivatives [J]. Information and Documentation Services, 2019, 40(3): 33-38.
[5] Song Chengyu, Li Xiuxia, Xie Ruixia, et al. Sleeping beauty paper identification based on quadratic function curve fitting [J]. Journal of Intelligence, 2018, 37(6): 119-123+207.
[6] Garfield E. Delayed recognition in scientific discovery: Citation frequency analysis aids search for case histories [J]. Current Contents, 1989, 12(23): 154-160.
[7] Costas R, Leeuwen TN van, Raan AFJ van. Is scientific literature subject to a 'sell-by-date'? A general methodology to analyze the 'durability' of scientific documents [J]. Journal of the American Society for Information Science and Technology, 2010, 61(2): 329-339.
[8] Wang J. Citation time window choice for research impact evaluation [J]. Scientometrics, 2013, 94(3): 851-872.
[9] Li J, Shi D, Zhao SX, et al. A study of the "heartbeat spectra" for sleeping beauties in science [J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(24): 7426-7431.
[10] Ke Q, Ferrara E, Radicchi F, et al. Defining and identifying sleeping beauties in science [J]. Proceedings of the National Academy of Sciences, 2015, 112(24): 7426-7431.
[11] Peruzzo F. Sleeping beauties and the citation dynamics in the network of scientific papers [EB/OL]. [2019-09-20] http://tesi.cab.unipd.it/50039/1/Peruzzo_{Fabio}.pdf.
[12] Du Jian, Wuyishan. A new parameter-free index for identifying sleeping beauties—Validation based on sleeping beauties in Science and Nature [J]. Information Studies: Theory & Application, 2017, 40(2): 19-25.
[13] Teixeira AAC, Vieira PC, Abreu AP. Sleeping beauties and their princes

in innovation studies [J]. Scientometrics, 2017, 110(2): 541-580.
[14] Bornmann L, Ye YA, Ye FY. Identifying "hot papers" and papers with "delayed recognition" in large-scale datasets by using dynamically normalized citation impact scores [J]. Scientometrics, 2018, 116(2): 655-674.
[15] Ye FY, Bornmann L. "Smart girls" versus "sleeping beauties" in the sciences: The identification of instant and delayed recognition by using the citation angle [J]. Journal of the Association for Information Science and Technology, 2018, 69(3): 359-367.
[16] Zong Zhangjian. Research progress on sleeping beauty paper identification methods [J]. Library and Information Service, 2019, 63(16): 132-142.
[17] Li J, Ye FY. Distinguishing sleeping beauties in science [J]. Scientometrics, 2016, 108(2): 821-828.
[18] Du Jian. Research on identification methods and awakening mechanisms of sleeping beauty papers [D]. Nanjing: Nanjing University, 2017.
[19] Wang Wensen. Coefficient of variation—A simple yet useful statistical indicator for measuring dispersion [J]. China Statistics, 2007, (6): 41-42.
[20] Avramescu A. Actuality and obsolescence of scientific literature [J]. Journal of the American Society for Information Science, 1979, 30(5): 296-303.
[21] Aversa ES. Citation patterns of highly cited papers and their relationship to literature aging—A study of the working literature [J]. Scientometrics, 1985, 7(3/6): 383-389.
[22] Cano V, Lind NC. Citation life cycles of ten citation classics [J]. Scientometrics, 1991, 22(2): 297-312.
[23] Qu Wenjian, Hu Zhiwei, Zhou Xiaoyu. Citation curve characteristics analysis of highly cited papers in library and information science [J]. Journal of Intelligence, 2017, 36(8): 138-143.
[24] Li Lingying, Min Chao, Sun Jianjun. Quantification and distribution of citation peaks [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(7): 697-708.
[25] Xiong Zequan, Duan Yufeng. Research on citation patterns of Chinese academic journal papers—Taking library and information science papers from 2006-2008 as examples [J]. Library and Information Service, 2019, 63(8): 107-115.
[26] Glänzel W, Schlemmer B, Thijs B. Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon [J]. Scientometrics, 2003, 58(3): 571-586.
[27] Zong ZJ, Liu XZ, Fang H. Sleeping beauties with no prince based on the co-citation criterion [J]. Scientometrics, 2018, 117(3): 1841-1852.
[28] Du Jian, Wuyishan. Important characteristics, predictive clues, and policy implications of sleeping beauty papers [J]. Studies in Science of Science, 2018, 36(11): 1938-1945.
[29] Guo Fei, Yan Xiaoyan. Analysis and improvement of sleeping beauty paper identification methods [J]. Library and Information Service, 2016, 60(8): 93-98.
[30] Zhong Zhen. Distinguishing Research Front from Research Frontier from citation structure differences between highly cited and zero-cited papers [J]. Library and Information Service, 2015, 59(8): 87-96.

[31] Du Jian, Wuyishan. Research on identification methods for sleeping beauty and prince papers [J]. Library and Information Service, 2015, 59(19): 84-92.
[32] Li Xiuxia, Shao Zuoyun, Liu Chao. Sleeping beauty paper identification in library and information science based on K-value algorithm [J]. Library and Information Service, 2017, 61(21): 114-122.
[33] Li Jiang. Review of "sleeping beauties" and "flash-in-the-pan" phenomena in science [J]. Journal of Academic Libraries, 2016, 34(3): 38-43.

Author Contributions: Tang Jie: research design and planning, data collection and analysis, manuscript writing; Zeng Jingjing: research design and planning, manuscript revision and improvement.

Identifying Sleeping Beauties in Science by Coefficient of Variation

Tang Jie[1,2,3], *Zeng Jingjing*[1,2]

[1]Northwest Institute of Eco-Environment and Resources, CAS, Lanzhou 730000
[2]Lanzhou Information Center, Chinese Academy of Sciences, Lanzhou 730000
[3]University of Chinese Academy of Sciences, Beijing 100049

Abstract: [Purpose/significance] This paper aims to review existing identification methods of sleeping beauties in science, discuss strengths and weaknesses of different kinds of methods, and put forward a brand-new method for identifying sleeping papers. [Method/process] This study is based on the Bcp index, which is a well-developed and accurate method for identifying sleeping beauties in science. Through referring to the core idea of using the "dispersion degree" of citation curve for identification, the concept of "coefficient of variation" in statistics is introduced to the new method. Then the PCV index is proposed to identify various citation curves, sleeping beauties in particular. [Result/conclusion] As shown in the results, PCV index can effectively identify the sleeping beauties literature. In addition, compared to the Bcp index, the new method has the advantages of simplicity and accuracy, and further reduces the dependence on the total number of citations.

Keywords: sleeping beauty; citation curve; coefficient of variation

Note: Figure translations are in progress. See original paper for figures.