# A Study on Temporal Effects on Patent Citation Frequency (Postprint)

**Authors:** Luo Wenxin, Zhao Yajuan

**Date:** 2023-07-26T00:00:00+00:00

## Abstract

[Purpose/Significance] Investigating the influence of temporal factors on patent citation frequency can mitigate the constraints imposed by temporal factors on technology evaluation activities, thereby enhancing the accuracy and reliability of such evaluations. [Method/Process] U.S. patent data from 1975 to 2017 were collected to conduct a correction study on patent citation frequency using the fixed-effects method. Patents were categorized by publication year and technical field, with within-group means and six top quantiles selected as citation frequency benchmarks. The baseline of citation frequency at the current time point and its historical temporal variations were statistically analyzed. A neural network model was constructed to fit the temporal evolution patterns of the baseline and predict future baseline values at subsequent statistical time points. [Results/Conclusion] Temporal disparities between patent publication years and statistical years preclude direct comparison of patent citation frequencies. This study establishes patent citation frequency baselines based on different technical fields, publication years, and statistical years, providing a reference framework for patent evaluation.

## Full Text

## Preamble

**Time Impact Study of Patent Cited Frequency**

Luo Wenxin[1],[2], Zhao Yajuan[1],[2]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

## Abstract

[**Purpose/Significance**] Investigating the influence of temporal factors on patent cited frequency can reduce the constraints imposed by time on technology evaluation activities and improve the accuracy and reliability of assessments. [**Method/Process**] This study collected U.S. patent data from 1975 to 2017 and conducted a revision study of patent cited frequency based on the fixed effects method. Patents were grouped by publication year and technical field, with within-group means and six TOP quantiles selected as benchmarks for cited frequency. The baseline values at the current time point and their historical temporal variations were statistically analyzed. A neural network model was established to fit the temporal variation patterns of the baselines and predict benchmarks for future statistical time points. [**Result/Conclusion**] The temporal difference between patent publication year and statistical year makes direct comparison of patent cited frequencies impossible. This paper establishes baseline values for patent cited frequency based on different technical fields, publication years, and statistical years, providing a reference for patent evaluation.

**Keywords:** patent cited frequency; time; fixed effect; temporal variation; benchmark

## Introduction

Patent citation analysis constitutes an important component of patent analysis, holding significant value for identifying key patents, exploring technological development trajectories, and evaluating patent value. This analysis relies on a series of patent citation metrics, such as citation count and cited frequency. In practical applications, the rational use of patent citation metrics requires further investigation due to the diverse ways in which patent citations are generated and the rich implications they reveal. This paper focuses on the temporal impact on patent citation metrics, specifically examining cited frequency as the primary indicator.

Patent citation metrics comprise two basic indicators and a series of derived indicators. The two fundamental metrics are citation count and cited frequency. Citation count can be further subdivided into patent citation count and non-patent citation count, from which indicators such as technological originality and technological universality are derived. Cited frequency, in turn, gives rise to metrics including citation index, science linkage, and technological strength. Between these two basic indicators, cited frequency holds greater practical significance, which is why this study concentrates on this particular metric within patent citation measurement.

The influence of temporal factors on citation metrics has long attracted scholarly attention both domestically and internationally, with most research focusing on paper citation metrics. In the domain of paper citation studies, the widely recognized ESI (Essential Science Indicators) system has been established, providing

citation metric benchmarks based on different fields and timeframes to eliminate the influence of temporal and disciplinary factors on evaluation indicators. However, similar practices are lacking in the patent field, where benchmarks based on different fields and timeframes have yet to be developed, leaving patent citation metrics without standardized references.

B. H. Hall first addressed the temporal impact issue of patent cited frequency in 2001 through research on U.S. patent citation data from 1975 to 1999. Hall identified two key problems: First, any patent's cited frequency only reflects citations up to the statistical time point; second, patent citation is influenced by the patent examination system, and variations in examination systems across different periods create differences in citation opportunities. Subsequent empirical studies found similar patterns in countries beyond the United States. Early research remained largely descriptive until 2014, when Wan Xiaoli, inspired by Hall's work, provided a detailed analysis of why patent cited frequency is temporally affected. She illustrated the "time cross-section" problem and "citation inflation" issue: the former refers to incomplete statistics due to unknown future citations, while the latter indicates that the average number of citations per patent increases annually, creating greater citation opportunities for individual patents. A. Breitzman discovered in 2015 that similar issues exist when evaluating groups of patents. Recent research continues to address these concerns, with A. B. Jaffe emphasizing in 2017 that patent cited frequency increases over time, exhibiting strong cohort effects—a concept from sociological research referring to how patents from different periods are differentially affected by time.

After reviewing relevant literature, we identified three primary ways in which time affects patent cited frequency: (1) patents published closer to the statistical time point have lower cited frequencies, making direct comparison impossible; (2) rapid advances in computer technology have improved patent examiners' search capabilities, increasing citation opportunities year by year, thus preventing direct comparison of patents published in different years over the same time interval; and (3) only citations up to the statistical time point can be calculated, making evaluation of younger patents prone to significant error. All three issues stem from differences in publication years or statistical years.

Based on this context, this study investigates the influence of temporal factors on patent cited frequency to address measurement problems caused by differences in publication and statistical years.

## 2. Data and Methods

### 2.1 Data Collection and Processing

Research data were collected from the Derwent Innovation (DI) database. U.S. patent citation information is more complete, making the United States the focus of this study. U.S. patent citation data have been computerized since 1975, so the data range was limited to all U.S. patents published between 1975 and 2017 and their citing patents.

The research dataset comprises two parts: the Base Patent Set (I) and the Citing Patent Set (II). The Base Patent Set includes all U.S. patents published from 1975 to 2017 and serves as the main dataset. The Citing Patent Set contains all patents that cite the base patents and functions as an extended dataset. The collected bibliographic fields are shown in .

The data collection and processing workflow is illustrated in [Figure 1: see original paper]. When collecting Base Patent Set data, the search was limited to the U.S. Granted Patents and U.S. Patent Applications databases with the query "DP>=(19750101) AND DP<=(20171231)" executed on January 20, 2018, yielding 11,742,361 records. For the Citing Patent Set, publication numbers of citing patents were first extracted from the Base Patent Set, deduplicated to create a publication number list, which was then used for retrieval, ultimately producing 1,995,3069 records.

Data processing included database construction, missing value handling, outlier screening, and data merging. The data merging step consolidated patent records from the Base Patent Set by application number, removing duplicates while retaining the earliest publication year and date among duplicate records and taking the union of citing patents. This consolidation was necessary because a single U.S. patent may generate multiple publication documents from application to grant, appearing as multiple records in the DI database with inconsistent citation patterns.

### 2.2 Research Approach and Methods

The research approach proceeds as follows: First, addressing the current situation where patent cited frequency is temporally affected and lacks evaluation standards, we identified the key problem as temporal differences between publication and statistical years that prevent direct comparison. Second, to eliminate this temporal effect, we reviewed correction methods for patent cited frequency and selected appropriate approaches: percentiles, relative impact indicators, and fixed effects methods. Finally, based on collected and processed patent data, we applied these selected methods to conduct temporal impact correction studies from both current time point and historical sequence perspectives, as shown in [Figure 2: see original paper].

## 3. Temporal Impact Correction of Patent Cited Frequency

Correction methods for temporal effects on citation metrics can be categorized into relative impact indicators, incorporation of temporal factors, citation window selection, percentiles, distribution curve simulation, and introduction of non-citation indicators. In the paper domain, relative impact indicators and percentile methods are commonly used across multiple disciplines and metrics, offering broad applicability, while other methods are mostly applied to specific disciplines or metrics with limited generalizability.

In the patent field, B. H. Hall calculated the mean cited frequency for patents

processed by the USPTO from 1969 to 1999 as of the end of 1999, proposing a fixed effects method for standardizing patent cited frequency metrics. Hall selected the mean cited frequency of similar patents as the benchmark and proposed relative cited frequency—dividing a patent's absolute cited frequency by the average cited frequency of patents in the same technical field and grant year. Fixed effect, originally a fundamental concept in experimental design, refers to treating the effect levels of observed factors as fixed parameters. In statistics, fixed effects models are statistical models with fixed parameters. For example, when data are grouped by several factors, within-group means can be selected as fixed effects for each subgroup.

Integrating correction methods from both paper and patent domains, this study employs the fixed effects method, selecting both mean cited frequency and TOP quantiles as benchmarks to correct for temporal effects. Assuming that variations in patent cited frequency across publication and statistical years are systematic, these variations must be eliminated when comparing different patents. Patents are grouped by publication year, statistical year, and technical field category, with within-group means and TOP quantiles selected as fixed effects for each subgroup.

The correction study includes both current time point correction and historical sequence correction, providing not only baseline values for current patent evaluation but also fitting historical variation patterns to predict future baseline values.

### 3.1 Current Time Point Baseline

With the statistical time point set at the end of 2017, baseline values were calculated by grouping patents by publication year and technical field (IPC section). The results are shown in and , where Table 2 presents within-group means and Table 3 shows TOP quantile levels. Six TOP quantiles were selected: TOP 0.01% (top 0.01%), TOP 0.10% (top 0.10%), TOP 1.00% (top 1.00%), TOP 10.00% (top 10.00%), TOP 20.00% (top 20.00%), and TOP 50.00% (top 50.00%). [Figure 3: see original paper] graphically depicts Table 2.

As of the end of 2017, the mean cited frequency for all fields and individual technical fields showed an initial increase followed by a decrease with later publication years. The decline in mean cited frequency for recently published patents reflects the truncation effect of statistical time—newer patents have not yet been fully cited by subsequently published patents. For earlier patents that are approximately free from statistical truncation effects, substantial differences in citation patterns across publication years are evident. Comparing different technical fields reveals that the peak publication years vary: Section A (Human Necessities) peaked in 1996, while Section G (Physics) peaked in 2000. The magnitude of mean cited frequency also differs across fields, with Section A showing significantly higher means than other sections, Sections G (Physics) and H (Electricity) having similar magnitudes, and Section D (Textiles; Paper)

showing the lowest means across all publication years.

For evaluating individual patents, one can locate the corresponding group based on publication year and IPC section, then divide the patent's cited frequency by the group mean from Table 2 or compare it against the TOP quantile levels in Table 3. For evaluating patent portfolios, a weighted cited frequency value can be calculated by grouping patents by publication year and IPC section, identifying the corresponding TOP quantiles from Table 3, and assigning patents to seven percentile ranges: ≥TOP 0.01%, TOP 0.01%-0.10%, TOP 0.10%-1.00%, TOP 1.00%-10.00%, TOP 10.00%-20.00%, TOP 20.00%-50.00%, and <TOP 50.00%. After calculating the proportion of patents in each range, different weights are assigned to each percentile range, and the weighted value is obtained by summing the products of range weights and patent proportions.

### 3.2 Historical Sequence Baseline and Prediction

**3.2.1 Historical Sequence Baseline** Patent cited frequency increases with the passage of statistical years. The means and TOP quantiles differ across statistical years. While previous research typically used a single statistical year, this study calculates means and TOP quantiles across different statistical years to explore the annual variation and growth patterns of patent cited frequency baselines. Based on these historical data, we fit models to predict baseline values for future statistical years.

[Figure 4: see original paper] shows the distribution of mean cited frequencies for different publication years across various statistical years. [Figure 5: see original paper] displays this distribution for the eight IPC technical fields. In both figures, the X-axis represents patent publication year, Y-axis represents statistical year, and Z-axis represents mean cited frequency. For all fields, with a fixed publication year, mean cited frequency shows an upward trend with increasing statistical year; with a fixed statistical year, mean cited frequency initially increases then decreases with later publication years. For patents published before 1985, the growth rate of mean cited frequency with statistical year was relatively flat. For patents published between 1985-1990, the growth rate accelerated noticeably. For patents published between 1990-2000, the growth rate became even faster. Without considering statistical truncation, more recently published patents exhibit faster growth rates in mean cited frequency over statistical time, likely due to rapid advances in computer technology that have improved patent examiners' search capabilities, thereby increasing citation opportunities year by year.

Comparing across fields, Sections A (Human Necessities), C (Chemistry; Metallurgy), G (Physics), and H (Electricity) show similar patterns to the overall trend. Sections B (Performing Operations; Transport) and F (Mechanical Engineering) exhibit relatively flat growth rates. Section E (Fixed Constructions) shows initially slow then accelerating growth, with a noticeable inflection point around 2005. Section D (Textiles; Paper) displays fluctuating rather than con-

sistently increasing trends.

Beyond mean cited frequency, TOP quantile distributions are shown in [Figure 6: see original paper] and [Figure 7: see original paper]. [Figure 6: see original paper] presents the TOP 1.00% quantile distribution across different publication and statistical years, with substantial variation across IPC fields. As examples, [Figure 7: see original paper] shows the TOP 1.00% quantile distributions for Sections A and B.

**3.2.2 BP Neural Network Model for Baseline Prediction** This study employs a BP neural network model to predict patent cited frequency baselines for future statistical time points. The BP (Back Propagation) neural network model, proposed by the PDP (Parallel Distributed Processing) research group, consists of input, hidden, and output layers. Training involves forward propagation of input signals and backward propagation of errors, using gradient descent to optimize connection weights and biases based on neuron errors.

**Data Preparation.** The historical baseline statistics from Section 3.2.1 serve as the fitting dataset, divided into four subsets: - Dataset A: For each statistical year from 1975-2017, calculate mean cited frequencies for all publication years (946 records) - Dataset B: Group patents by IPC section, then for each statistical year calculate mean cited frequencies for all publication years (7,568 records) - Dataset C: For each statistical year, calculate TOP quantiles for all publication years (946 records) - Dataset D: Group by IPC section, then for each statistical year calculate TOP quantiles for all publication years (7,568 records)

**Model Construction.** - Model A: Uses Dataset A with 4 input neurons (publication year, citing publication year, time interval, patent count) and 1 output neuron (mean cited frequency). Hidden layer size is determined by the principle of using the minimum number of neurons that meets precision requirements. According to formula (1), hidden neurons should range from 4-12. Testing different sizes within this range identified 12 neurons as optimal. - Model B: Uses Dataset B with 5 input neurons (IPC section, publication year, citing publication year, time interval, patent count). Output and hidden layer configurations follow Model A. - Model C: Uses Dataset C with 4 input neurons and 6 output neurons representing the six TOP quantiles. According to formula (1), hidden neurons should range from 5-13, with 13 identified as optimal. - Model D: Uses Dataset D with 5 input neurons and the same 6 output neurons as Model C. Hidden layer configuration follows Model C.

**Model Training.** GroupKFold (group k-fold validation) was used, grouping the four datasets by statistical year to ensure samples from the same group appear together in either training or test sets. With K=10, datasets were divided into 10 folds, using 9 folds for training and 1 for testing in rotation.

**Model Evaluation Metrics.** Mean Absolute Error (MAE) and Mean Squared Error (MSE) were selected for model assessment, calculated according to formulas (2) and (3) respectively.

**Model Evaluation and Prediction.** - **Model A:** With 12 hidden neurons and 100 training epochs, the model achieved stable performance. The optimal performance curve in [Figure 11: see original paper] shows strong overall fit, with the X-axis representing statistical year and Y-axis representing mean cited frequency. [Figure 12: see original paper] presents a prediction example for statistical year 2018, showing predicted mean cited frequencies for patents published 1975-2017. - **Model B:** With 12 hidden neurons and 500 training epochs, the model performed better for technical fields with smaller variation in mean cited frequency (e.g., Sections B, D) while requiring improvement for fields with larger variation (e.g., Section A). [Figure 14: see original paper] shows a prediction example for statistical year 2018 across different technical fields. - **Model C:** With 13 hidden neurons and 500 training epochs, the model's optimal performance for TOP 1.00% quantile is shown in [Figure 15: see original paper]. [Figure 16: see original paper] presents a prediction example for statistical year 2018. - **Model D:** With 12 hidden neurons and 300 training epochs, the model's performance varied by technical field, showing good fit for Sections B and C but requiring enhancement for Sections A, D, and E. [Figure 18: see original paper] provides a prediction example for statistical year 2018 across technical fields.

## Conclusion

This study reveals that temporal differences between patent publication and statistical years prevent direct comparison of patent cited frequencies. We established baseline values based on different technical fields, publication years, and statistical years, providing references for patent evaluation. BP neural network models were constructed to fit historical baseline variations, yielding optimal models for predicting future baseline values. The study provides predicted baseline values for citations counted through the end of 2018.

Key findings include: (1) With fixed publication year, mean cited frequency increases with statistical year; (2) With fixed statistical year, mean cited frequency initially increases then decreases with publication year; (3) Without considering truncation, more recently published patents show faster growth rates in mean cited frequency over time, reflecting improved examiner search capabilities due to computer technology advances.

**Innovations:** (1) While existing research focuses on paper citation metrics, this study addresses temporal effects on patent citation metrics, offering novel perspectives by adapting paper-based correction methods; (2) The study establishes patent cited frequency benchmarks across different fields and timeframes, providing standardized references previously lacking in patent analysis; (3) Unlike Hall's pioneering but dated study (through 1999), this research uses temporal sequences rather than single cross-sections to model and predict baseline variations via neural networks.

**Limitations and Future Directions:** (1) This study focuses only on cited frequency, while patent citation analysis includes many valuable metrics that

warrant future temporal impact studies; (2) The established baselines have not yet been applied in practice—future empirical research should test and validate these correction methods in real scenarios; (3) The neural network approach could be further optimized, potentially through mathematical formulas or more intuitive representations of variation patterns.

## References

[1] JAFFE AB, DERASSENFOSSE G. Patent citation data in social science research: overview and best practices[J]. Journal of the Association for Information Science & Technology, 2016, 68(6): 1360-1374.
[2] REN Shengli, WANG Baoqing, GUO Zhiming, et al. Caution needed when using journal impact factors to evaluate research 成果 [J]. Chinese Science Bulletin, 2000, 45(2): 218-222.
[3] COZZENS SE. Comparing the sciences: citation context analysis of papers from neuropharmacology and the sociology of science[J]. Social studies of science, 1985, 15(1): 127-153.
[4] HALL BH, JAFFE AB, TRAJTENBERG M. The NBER patent citations data file: lessons, insights and methodological tools[R]. Cambridge: National Bureau of Economic Research, 2001: 25-
[5] WAN Xiaoli. In-depth analysis of "cited frequency" in patent quality indicators[J]. Information Science, 2014, 32(1): 68-73.
[6] BREITZMAN A, THOMAS P. Inventor team size as a predictor of the future citation impact of patents[J]. Scientometrics, 2015, 103(2): 1-17.
[7] SCHUBERT A, BRAUN T. Relative indicators and relational charts for comparative assessment of publication output and citation impact[J]. Scientometrics, 1986, 9(5): 281-291.
[8] GLANZEL 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.
[9] CHEN Zifeng, GUAN Jiancheng. Research on the impact of international patent cooperation and citation on innovation performance[J]. Science Research Management, 2014, 35(3): 35-42.
[10] CAI Hong, WU Kai, SUN Shuncheng. Empirical study on international technology spillovers based on patent citations[J]. Management Science, 2010, 23(1): 18-26.
[11] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3): 851-872.
[12] ABRAMO G, CICERO T, D'ANGELO CA. A sensitivity analysis of research institutions' productivity rankings to the time of citation observation[J]. Journal of informetrics, 2012, 6(2): 192-201.
[13] LEYDESDORFF L, BORNMANN L, MUTZ R, et al. Turning the tables on citation analysis one more time: principles for comparing sets of documents[J]. Journal of the Association for Information Science & Technology, 2011, 62(7): 1370-1381.

[14] Centre for Science and Technology Studies, Leiden University, The Netherlands. Leiden ranking[EB/OL]. [2018-10-10]. http://www.leidenranking.com/methodology.aspx.
[15] PERNEGER TV. Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ[J]. BMJ, 2004, 329(7465): 546-547.
[16] SHEMA H, BAR-ILAN J, THELWALL M. Do blog citations correlate with a higher number of future citations? research blogs as a potential source for alternative metrics[J]. Journal of the Association for Information Science & Technology, 2014, 65(5): 1018-1027.

**Author Contributions:**
Luo Wenxin: Literature review, research implementation, and paper writing;
Zhao Yajuan: Topic selection, paper revision and approval, research guidance.

**Time Impact Study of Patent Cited Frequency**

Luo Wenxin[1],[2], Zhao Yajuan[1],[2]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

**Abstract:** [Purpose/significance] To study the influence of time factor on patent cited frequency can reduce the restriction of time factor on technology evaluation activities and improve the accuracy and reliability of evaluation. [Method/process] This paper collected U.S. patent data from 1975 to 2017, and carried out the revision study of patent cited frequency based on fixed effect method. The patents were grouped according to different publication years and different technical fields. The group mean and six TOP quantiles were selected as the benchmarks of patent cited frequency, and the baseline of the current time point and the historical time series changes of the baseline were counted. A neural network model was established to fit the timing variation of the baseline and predict the baseline of future statistical time points. [Result/conclusion] The time difference between patent publication year and statistical year makes it impossible to directly compare patent citations. This paper establishes benchmarks for patent citations based on different technical fields, different publication years and different statistical years, providing reference for patent evaluation.

**Keywords:** patent cited frequency; time; fixed effect; timing variation; benchmark

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