Acta Informatica Pragensia
2026, Volume 15, Issue 1, 22–35
https://doi.org/10.18267/j.aip.281

PRAGUE UNIVERSITY
OF ECONOMICS
AND BUSINESS

Article Open Access

Analysis of Benford's Law Conformity with Web of Science Citations of Documents

David Jiri Slosar

Institute of Information Studies and Librarianship, Faculty of Arts, Charles University, Prague, Czech Republic
Library of the Czech Academy of Sciences, Prague, Czech Republic

Corresponding author: David Jiri Slosar (davidjiri.slosar@ff.cuni.cz)

Editorial Record

First submission received:
March 8, 2025

Revisions received:
May 5, 2025
May 30, 2025
June 30, 2025

Accepted for publication:
July 6, 2025

Academic Editor:
Zdenek Smutny
Prague University of Economics
and Business, Czech Republic

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

How to cite this article:
Slosar, D. J. (2026). Analysis of Benford's Law Conformity with Web of Science Citations of Documents. *Acta Informatica Pragensia*, 15(1), 22–35. https://doi.org/10.18267/j.aip.281

Copyright:
© 2026 by the author(s). Licensee Prague University of Economics and Business, Czech Republic. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

Abstract

Background: Benford's law is a statistical phenomenon that predicts the probability of a particular digit at a particular position in a number. This law has been successfully applied in a number of areas, such as accounting. In the area of scientometrics, research has been devoted mostly to journal data.

Objective: This paper investigates the conformity of Benford's law with the citation counts of records retrieved from the Web of Science database. We evaluate the conformity levels with Benford's law in the complete dataset. We determine the effect of document type (article, proceedings paper and review), year of publication (2014–2018) and Web of Science categories (254 categories) on the level of conformity of the citation counts with Benford's law.

Methods: The dataset of this research contains over 8.47 million records. All available records from the Web of Science were downloaded, so this set is the entire population of data available at the time of download. The distributions of the first significant digits in the citation counts of these records are compared with Benford's law. Mean absolute deviation (MAD) recommended by Nigrini (2012) and sum of squared deviations (SSD) recommended by Kossovsky (2015) are used to categorize the similarity of the citation counts to Benford's law.

Results: The entire dataset of this study shows marginal conformity according to both MAD and SSD intervals (with a MAD value of 0.1257 and an SSD value of 29.9; a lower value indicates a better agreement). The review document type shows a high level of conformity, while proceedings paper shows a lower level. We found significant differences in conformity between Web of Science categories.

Conclusion: This study mapped the level of conformity of the citation counts with Benford's law in data from the Web of Science database. Further directions for possible research are suggested.

Index Terms

Benford's law; Citations; Scientometrics; Bibliometrics; Web of Science.

1 INTRODUCTION

Benford's law, also referred to as the Newcomb–Benford law or the law of anomalous numbers, is an empirical observation that applies to many real-life sets of numerical data. This statistical phenomenon reveals that the leading digits in these datasets are not uniformly distributed as one might intuitively expect. Instead, they adhere to a specific logarithmic distribution. This law is prevalent across various natural datasets, demonstrating that significant digits follow a predictable pattern rather than a random distribution (Berger & Hill, 2015). A considerable amount of research has been done on Benford's law. A comprehensive overview of such studies is available at the website Benfordonline.net (Berger et al., 2009).

Successful applications of Benford’s law can be mentioned, for example, in the fields of accounting (Nigrini, 2012), music (Bantange et al., 2023) or COVID data (Kennedy & Yam, 2020). Particularly useful in the context of Benford’s law are the monographs of Kossovsky (2015) and Nigrini (2012). They describe in detail the theory, applications, computational guidelines and many other aspects of Benford’s law.

Previous investigations utilizing partial datasets indicate that scientometric data adhere to Benford’s law. While the citation counts of scientific outputs constitute a segment of scientometric data, they have not been extensively analysed, with the exception of Gupta et al. (2024). Prior research has predominantly focused on journal citation counts, Impact Factors and analogous metrics (Campanario & Coslado, 2011; Egghe & Guns, 2012; Alves et al., 2014; Alves et al., 2016; Tošić & Vičič, 2021; Sorour et al., 2024; Bertin & Lafouge, 2025).

The main objective of this study is to fill the research gap by determining the level of conformity of the citation counts of scientific outputs with Benford’s law. For this purpose, a dataset of 8.47 million records from the Web of Science database was obtained. This large dataset is the entire population within the database. This approach allows a detailed mapping of the level of conformity of the citation counts with Benford’s law. Moreover, the whole population is involved, so the problem of under-representation of the sample is not present.

Other objectives of this study are to determine whether certain factors influence the level of conformity of the citation counts of scientific outputs with Benford’s law. Only factors that could be obtained within a reasonable time were selected (see the description of data acquisition for more details). These factors are year of publication (2014-2018), document type and Web of Science category. To address the objectives of this study, four research questions (RQ) have been formulated.

> **RQ1:** What is the level of conformity of the citation counts of articles, proceedings papers and reviews published between 2014 and 2018 and indexed in the Web of Science database?

Answering this question verifies the assumption that Benford’s law is valid for the citation counts, as has been theoretically derived by Egghe (2011).

> **RQ2:** Does the year of publication of a document have an effect on the level of conformity of the citation counts of these documents with Benford’s law?

Older outputs are expected to have had more time to be cited than more recent outputs. This phenomenon can have a major impact on the level of conformity with Benford’s law (Aksnes et al., 2019). Irregularities or sudden changes in the conformity of individual years could point to limitations in the choice of the time window, for example, too small gap between the year of publication of the scientific output and the year of data collection. These limitations have not yet been identified in previous studies.

> **RQ3:** Are there differences between document types in the level of conformity of the citation counts with Benford’s law?

Differences in citation rates can have a major impact on the level of conformity with Benford’s law. Moreover, differences in the citation frequency of certain types of documents can lead to changes in the publishing habits of scientists. For example, the document type review tends to be more heavily cited than other document types and the proportion of reviews in citation databases is growing (Miranda & Garcia-Carpintero, 2018).

> **RQ4:** Are there differences between disciplines (Web of Science categories) in the level of conformity of the citation counts with Benford’s law?

Citation patterns are field-specific (Crespo et al., 2013). These field characteristics could affect the level of conformity with Benford’s law. It is important to find possible differences in fields in order to create a tool to determine the suitability of fields for being processed by scientometric methods in the evaluation of science. So far, the discussion has been conducted only at the level of broad field groups and without the use of Benford’s law (Ochsner, 2021). By investigating the conformity of the citation counts of outputs in certain fields, we are laying the foundations for possible research in which Benford's law could be used to detect predatory journals. This detection could work on the principle of detecting deviations from conformity with Benford's law in the citation counts of outputs published in journals of relevant fields.

2 THEORETICAL BACKGROUND

2.1 Benford's law

In early research, the law of anomalous numbers was used to determine the probability of the digits 1 to 9 being in the first significant (non-zero, from the left) position (Newcomb, 1881; Benford, 1938).

The formula for calculating the probability of occurrence for the first significant digit d according to Benford's law is:

$$F_d = log_{10}(\frac{d+1}{d}),$$

where F_d is the probability value and $d \in$ D = (1, 2… 9) is a digit. The decimal logarithm here indicates that the calculation refers to the decimal system (Benford, 1938). The resulting values for each significant digit are given in Table 1.

Table 1. *Calculated values for first significant digit according to Benford's law.*

First significant digit	$F_d = \log_{10}(\frac{d+1}{d})$
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

Even the distributions of digits at other positions (second digit, etc.) or their combinations (first two, etc.) can be predicted by Benford's law in some datasets. As will be explained later, citation counts typically do not have properties that would allow us to determine the level of conformity to Benford's law on any significant digit other than the first.

2.2 State of the art of Benford's law investigation in scientometrics data

In the field of information science, Egghe (2005) derived the Zipf-Mandelbrot law (Zipf, 1949; Mandelbrot, 1965) from Lotka's law (Lotka, 1926). Subsequently, Egghe (2011) extended this work by deriving Benford's law from the Zipf-Mandelbrot law. It is widely acknowledged that both Lotka's law (Ruiz-Castillo & Costas, 2014) and Zipf's law (Solla Price, 1965; Redner, 2005) are applicable to the distribution of citations in scientific literature.

Citation data are a specific type of data that has not been sufficiently investigated in the context of Benford's law. Therefore, the present study investigates the conformity of the citation counts in a large dataset (citation data of Web of Science records with years of publication 2014 to 2018 and document types article, proceedings paper and review) with Benford's law.

In the context of citation counts, it is necessary to discuss the act of citation. There are two main streams of research that comment on the properties of citations: the normative theory (Merton, 1973) and social constructivism (Latour & Woolgar, 1979). However, it does not matter which of these paradigms is applied to citation. They both agree that for the most part, citing another scholarly document is a product of human intent. Some of the citations can be considered somewhat accidental, for example, when a scientist selects from multiple studies that are equivalent in content or cites a study that they found and does not cite those that they did not even discover in their search. However, even in cases where a scientist is forced by a publisher (in order to increase the Impact Factor) to cite other papers from the journal in which the scientist wants to publish, there is human volition present in the act of citation. Thus, it would seem that citation counts should not follow Benford’s law.

The decomposition of the act of citation into the concepts of citation and reference resolves this apparent paradox. A citation can only be received from another document/scholar. A reference can only be given to another document/scholar. Citation counts are obtained essentially at random by documents through a combination of many external factors (citation dynamics of the field, Matthew effect, content influence, etc.) and thus it is not *a priori* impossible that they follow Benford’s law. A scientist has a very limited wilful influence on how their own paper is cited. Tools for influencing the citation counts received include, for example, self-citation, paper promotion or other forms of appeal to the scientific community (Bornmann & Daniel, 2008). It is fair to question whether the use of these tools causes deviations of citation counts from Benford’s law. Furthermore, reference counts are for the most part a product of human will and therefore do not follow Benford’s law (Bertin & Lafouge, 2025). This study only examines citation counts. The effect of efforts to obtain high citation counts can be considered in future research.

There are several studies in the context of evaluating the level of conformity of scientometric indicators with Benford’s law. Campanario & Coslado (2011) investigated how the number of documents, citation counts and Impact Factor of journals in the JCR (Journal Citation Reports) database from 1998 to 2007 follow Benford’s law. This is the first published study on the subject of empirical validation of Benford’s law in scientometric data.

Egghe & Guns (2012) performed a conformity analysis with a generalized Benford’s law over the data of Campanario & Coslado (2011). Alves et al. (2014) followed the study of Campanario & Coslado (2011). They used a dataset of the number of documents, citations and Impact Factor of journals from 2007 to 2011 from the JCR and Scopus databases. This dataset was divided according to the selected journals' countries of publication and fields. Alves et al. (2016) further expanded on previous studies by adding additional journal citation indicators, years of publication and second significant digit conformity according to Benford’s law.

The abovementioned studies present findings on the validity of Benford’s law with scientometric indicators such as Impact Factor, number of papers or citation counts of scientific journals. These scientometric indicators have achieved high conformity with Benford’s law. However, they did not measure the validity of Benford’s law with citation counts of scientific papers.

By aggregating the citations of individual papers to entire journals in the above studies, the citation counts may have been made more consistent with Benford’s law, in the same way that shop receipts are. The individual prices of goods in shops do not follow Benford’s law because the value is influenced by human will (instead of a price of 107 euros, for example, the price is adjusted to 99 euros). However, values on receipts that result from combinations of prices of goods follow Benford’s law very well (Kossovsky, 2015). Thus, it is not possible to assume that citation counts of scientific papers follow Benford’s law even if journal citation counts do.

In the recent years, the discussion regarding the conformity of scientometric data with Benford's law has become more intense. Tošić & Vičič (2021) made a high-quality study on the application of Benford’s law to a scientific research collaboration network. They proposed a methodology for evaluating the maturity of a research system. The study used contribution of authors instead of citation counts.

Gupta et al. (2024) focused mainly on altmetrics, but extended their analysis to the "number of Dimensions citations". Even though only marginal attention was given to results of conformity of these Dimensions citations, it is the only contribution in the area of the citation count conformity with Benford’s law. The results show that the citation counts of results in the Dimensions database are in fairly high agreement with Benford’s law. However, this agreement was not quantified in any way and its evaluation is only possible by comparing the measured distribution against the theoretical probabilities in the graph. Another limitation of the study is that it worked primarily with downloaded records with altmetrics published in 2021 (1,787,976 records). Some of those records also had the citation counts information available. The set of records is not the entire population from the Dimensions database (although likely of a significant size).

Sorour et al. (2024) investigated conformity of journal Impact Factor of journals with Benford’s law. The dataset contained more than 12,000 journals with data from 1997 to 2021. The result is that subscription journals are more likely to be non-conforming with Benford’s law than open access journals. However, no interpretation of these results was offered.

A further extension of the investigation into journal scientometric data was provided by Bertin & Lafouge (2025). They compared data of Campanario & Coslado (2011) with a newly obtained dataset of journal data from 1997 to 2019. These data included the H-index, the cumulative number of citations received over three years, the number of references and the number of articles of journals. The data were downloaded from the Web of Science and Scopus databases. The study resulted in a categorization of the scientometric objects based on the level of conformity with Benford's law. For example, it was empirically verified that reference counts do not follow Benford's law. The reference counts are thus categorized as non-Benfordian. The result relevant for the present study is that the conformity of number of citations of journals is consistently high across the years 1997 to 2019. Thus, at least at the level of citation counts of entire journals, conformity with Benford's law appears to be stable over the observed period of more than 20 years.

As was stated in the introduction, previous studies conducted on partial datasets suggest that scientometric data follow Benford's law. Citation counts of scientific outputs are part of scientometric data but have not been specifically examined (except by Gupta et al., 2024). Previous research has concentrated on journal citation counts, their Impact Factors and similar data (Campanario & Coslado, 2011; Egghe & Guns, 2012; Alves et al., 2014; Alves et al., 2016; Tošić & Vičič, 2021; Sorour et al., 2024; Bertin & Lafouge, 2025).

3 METHODS

Our analysis of Benford's law applicability has been performed on the complete population of data from the Web of Science, Core Collection database, see Table 2. All available records with the document types article, proceedings paper and review and also the year of publication 2014 to 2018 were downloaded. These document types were chosen because they are represented in the database in large numbers or are the most cited. The queries[1] were entered into the API and records have been downloaded through the API between February 24 and March 22, 2020.

In order to reduce the size of the dataset while maintaining its completeness, the document types chosen were article, proceedings paper or review, which are the most represented and most cited document types in the Web of Science database. Similarly, the time period had to be chosen. Existing research into the conformity of scientometric indicators with Benford's law does not provide a uniform time window that is appropriate for detecting trends. The lengths of the time windows vary across studies and are not justified in any way. Thus, we chose a five-year time window, which is often used in science evaluation systems. Furthermore, data acquisition for this research was carried out during the COVID-19 pandemic. This led us to limit the time series to the pre-pandemic period, which may have affected publication and citation patterns (e.g., by reducing the number of conference papers).

Table 2. *Overview of number of documents, number of citations and average citation rate of documents in WoS database.*

Document type	Number of documents	Number of citations	Average citations
Article	6,344,455	65,157,037	10.27
Proceedings paper	975,512	1,948,473	2.00
Review	449,229	10,652,996	23.71
Letter	173,881	460,394	2.65
Meeting abstract	1,260,754	166,281	0.13
Editorial material	443,436	1,259,746	2.84
Book chapter	53,653	31,665	0.59

Note: *Data are from the InCites analytics, for the InCites Dataset + ESCI, for OECD territory, from content indexed as of 31 May 2020. Data are for the publication years 2014 to 2018. This selection of document types covered more than 95% of the records from the selected period.*

[1] An example of a query is DT = (article) AND FPY = (2018) AND WC = (biology). Tags are used in advanced search of the Web of Science database. These tags read as follows: DT = document type, FPY = final publication year and WC = Web of Science category.

A total of 28 million records were retrieved with a significant number of duplicates[2], including records with no citations. After deduplication (Table 3) using UT WOS identifier, the dataset contained more than 8.47 million records cited at least once. In case the downloaded duplicate records have different citation counts (Web of Science updates citation counts daily), the first downloaded record was retained.[3] In addition to the citation counts, document type and year of publication, information about the Web of Science categories field was also extracted (254 categories).

Table 3. *Numbers of deduplicated records in years and numbers of records for analysis, after removal of records with no citations.*

Year of publication	Number of deduplicated records [million]	Number of cited records [million]
2014	1.98	1.60
2015	2.28	1.76
2016	2.38	1.78
2017	2.47	1.75
2018	2.47	1.58

Due to the large volume of data, errors occurred during the extraction process. In some cases, empty results were retrieved when downloading data. The API interface showed activity during the data download, provided a set of records, but it contained only a fraction of the data specified in the API query. These incomplete set of records were deleted and downloaded again. Therefore, a correction data package of approximately 3 million records was downloaded between 4 and 12 April 2020. The missing records were evenly distributed across years of publication, document types and Web of Science categories. We believe that the time lag in downloading the original and correction data package is negligible for the purposes of this study and has no significant effect.

In general, any dataset for verification of the applicability of Benford's law must satisfy several general prerequisites. Benford's law is not observed in datasets with, for example: a normal distribution (IQ in population), a restricted arbitrary boundary or datasets burdened by systematic human intention (Nigrini, 2014).

These effects are not presented in the dataset of citation counts. For the citation counts obtained, we therefore cannot conclusively determine whether they have properties that would make them *a priori* unsuitable for verifying conformity with Benford's law. There are, however, additional limiting conditions that apply to the dataset comprising the citation counts and it is necessary to take these conditions into account:

1. The order for the range of citation counts in the dataset must be sufficient and the range of numbers in the dataset must be large enough. Kossovsky (2015) argued that datasets for which the difference between the maximum and minimum values at $F_{diff} > 3$, according to the formula $F_{diff} = \log(max) - \log(min)$, are sufficiently robust for reliable comparisons between the relative frequencies of the first significant digits on the one hand and the probability values determined by Benford's law on the other hand.[4] The condition $F_{diff} > 3$ is satisfied in all the analyses except the analysis of the factor of Web of Science category (the analyses are explained below).
2. The dataset size must be sufficient. Nigrini (2012) stated that a dataset should contain at least 1,000 records. If a dataset contained less than 1,000 records, it was removed from the analysis.

[2] Web of Science records can be classified into one to six Web of Science categories. When downloading data by Web of Science category, document type and year of publication, it was common for multidisciplinary records to be downloaded multiple times.

[3] We also tried to keep the last record retrieved, to see whether this procedure might affect the conformity of the citations counts with Benford's law. When the first retrieved record was retained, the complete dataset contained 35.024% of records beginning with the digit one. Keeping the last downloaded record, the complete dataset contained 35.032% of records starting with the digit one. The difference is 0.008 percentage points, which we considered insignificant.

[4] He adds that it is more appropriate to calculate $F_{diff} > 3$ in a dataset in which 10% of the highest and 10% of the lowest values have been removed, thus removing outliers.

For the citation counts of scientific output records, the possibilities to determine conformity with Benford's law are limited. To detect counts other than the first significant digits (second, first two, etc.), the citation dataset numbers need to be at least five digits in length (Nigrini, 2012). However, in the entire citation dataset obtained, approximately 68.94% of the numbers are single-digit and 30.27% of the numbers are double-digit. Only ten out of these numbers are five-digit. At the same time, 16.87% of the citation dataset is made up of records for which only one citation is present. In other words, the present paper, which deals with the citation counts of records, can evaluate only the distribution of the first significant digits from the viewpoint of Benford's law.

Another characteristic of the citation counts is that they are always integer values ranging from zero to infinity (the most cited paper in the dataset of the present study received 16,256 citations). To keep $F_{diff} > 3$, then, the lower bound of the interval is given by the citation count of 1 and the minimum upper bound is given by the citation count of 1,000. Thus, to satisfy $F_{diff} > 3$, at least one record with 1,000+ citations must exist for the given dataset.

Although it is commonly used, Kossovsky (2021) argued in great detail the appropriateness of the chi-square test in the specific datasets. Moreover, the chi-square test has an excess power problem because as the number of observations increases (Cleary & Thibodeau, 2005; Nigrini, 2012; Druica et al., 2018; Gupta et al., 2024), it becomes more susceptible to insignificant spikes, which leads to the conclusion that the data do not comply (Tošić & Vičič, 2021).

Kossovsky (2015) offered the sum of squared deviations (SSD) metric as appropriate for monitoring conformity of data with Benford's law. The thresholds to determine levels of conformity are based not on statistical theory but on empirical data.

The SSD value is calculated as

$$SSD = 10{,}000 * \sum_{d=1}^{K} (AP_d - EP_d)^2,$$

where d is the relevant digit, K is the number of digits under consideration, AP_d is the relative frequency of the relevant digit and EP_d is the expected probability of the digit according to Benford's law. Furthermore, Kossovsky (2015) made use of the 10,000 coefficient to conveniently re-scale the values.

This metric is used in the present study to evaluate conformity with Benford's law. To characterise the actual level of the citation counts conformity with Benford's law, Kossovsky (2015) offered the following thresholds for the level of conformity: perfectly Benford SSD < 2, acceptably close 2-25, marginally Benford 25-100, and non-Benford SSD > 100.

Complementarily, the mean absolute deviation (MAD) metric recommended in Nigrini (2012) will be used. Despite the considerable similarity between these two metrics (Cergueti & Maggi, 2021), both MAD and SSD are used in this study to evaluate the level of conformity of our data with Benford's law.

The formula for calculating the MAD value is

$$MAD = \frac{1}{K} * \sum_{d=1}^{K} |AP_d - EP_d|,$$

where K is the number of digits under consideration, d is the relevant digit, AP_d is the relative frequency of the relevant digit and EP_d is the expected probability of the digit according to Benford's law. Nigrini (2012) presented the following thresholds for characterising the level of conformity with Benford's law: perfectly Benford MAD < 0.006, acceptably Benford 0.006 – 0.012, marginally Benford 0.012 – 0.015, and non-Benford MAD > 0.015.

The borderline values of both MAD and SSD for characterising the levels of conformity have been determined by the authors (Nigrini, Kossovsky) experimentally based on various selected datasets. Both of these descriptive statistics lack the mathematical foundations similar to statistical tests, such as chi-square. It is not possible to perform rigorous hypothesis testing based on MAD and SSD. However, their advantage is that they are not sensitive to the size of the dataset or the distribution of the data being tested, and thus provide a convenient tool for determining conformity with Benford's law in the citation counts.

We consider the average citation rate of the dataset to be an important indicator (AVG TC calculated as *(Sum of citations of all records in dataset)/(Number of records in dataset)*). During the analysis, it was revealed that this indicator effectively reflects the intensity of citations within the datasets, which may be associated with the level of conformity of the dataset to Benford's law.

The present study performs an analysis of the complete dataset (RQ1). Additionally, three analyses are performed relating to a specific factor which could influence conformity of datasets with Benford's law. These factors are year of publication (RQ2), document type (RQ3) and Web of Science categories (RQ4). Such a procedure allows detailed mapping and identification of key influences on the citation count data conformity with Benford's law.

4 RESULTS

This section contains a detailed overview of all the analyses performed addressing research questions RQ 1-4. Their further evaluation is part of the Discussion section.

4.1 Analysis addressing complete dataset – RQ1

In this analysis, all 8.47 million records for 2014-2018, chosen document types and all Web of Science categories were examined without any division. The purpose of this procedure was to answer the first research question to what extent the citation counts of scientific outputs are in conformity with Benford's law. The results are shown in Table 4 and Figure 1.

Table 4. *Absolute counts and relative frequencies for digit occurrences of numbers at first position with respect to Benford's law; MAD and SSD results.*

Analysis – Complete dataset											
1st significant digit	1	2	3	4	5	6	7	8	9	**SSD**	**MAD**
Benford	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046		
Complete dataset absolute counts	2,965,553	1,552,916	1,034,784	767,801	605,048	492,070	409,918	344,122	29,5021		
Complete dataset relative frequencies	0.35	0.183	0.122	0.091	0.071	0.058	0.048	0.041	0.035	29.8	0.0126

Using the above-described SSD and MAD metrics, the complete dataset is identically classified into the marginally Benford level of conformity. However, for both descriptive statistics, the results are near to the limits of the acceptably close level of conformity.

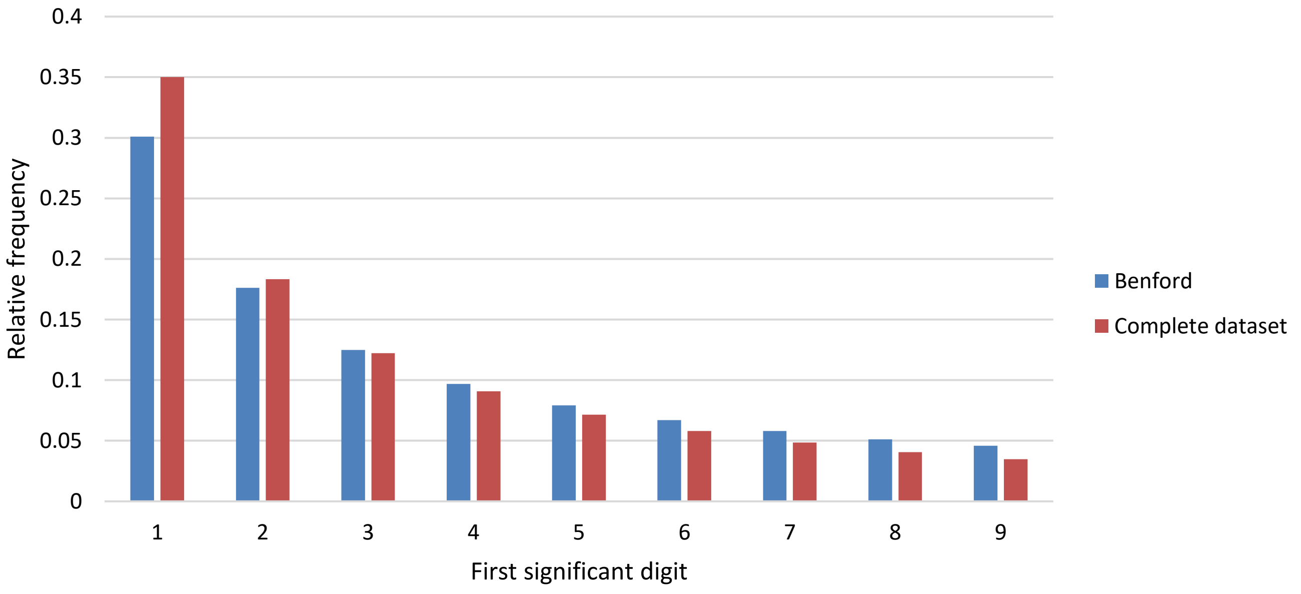

Figure 1. *Relative frequency of occurrence for first significant digit in entire dataset with respect to Benford's law.*

4.2 Analysis addressing factor of year of publication – RQ2

This analysis compares conformity of the citation counts published in individual years with Benford's law. The datasets are very robust across the breadth of orders and Kossovsky's requirement for $F_{diff} > 3$ is satisfied in all of them. The evolution in time of the conformity of citation counts with Benford's law should be evident here, see Table 5.

Table 5. *Relative frequency of occurrence for first significant digit for individual years and MAD and SSD results.*

Analysis – Factor of year of publication												
1st significant digit	1	2	3	4	5	6	7	8	9	AVG TC	MAD	SSD
Benford	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046			
2014	0.345	0.183	0.120	0.089	0.070	0.059	0.051	0.044	0.039	16.83	0.0113	23.4
2015	0.349	0.181	0.119	0.088	0.071	0.059	0.051	0.044	0.038	13.64	0.0118	27.0
2016	0.350	0.178	0.119	0.090	0.072	0.060	0.051	0.043	0.037	10.86	0.0113	27.6
2017	0.349	0.180	0.122	0.093	0.073	0.060	0.049	0.040	0.034	8.24	0.0113	27.1
2018	0.360	0.196	0.131	0.094	0.070	0.052	0.040	0.031	0.025	5.49	0.0189	53.5

The results suggest that older documents have a greater conformity (acceptably close) with Benford's law than more recent ones; the most recent year significantly deviates from Benford's law (non-Benford). The differences in the result values (except 2018) are too small to be considered significant. Additionally, the time series is too short to draw firm conclusions. We suggest that the five-year window should be extended.

4.3 Analysis addressing factor of document type – RQ3

Records of complete dataset were divided by document type, see Table 6. This analysis is used to show the differences in conformity of Benford's law between document types. Again, Kossovsky's requirement for a breadth of orders $F_{diff} > 3$ is satisfied.

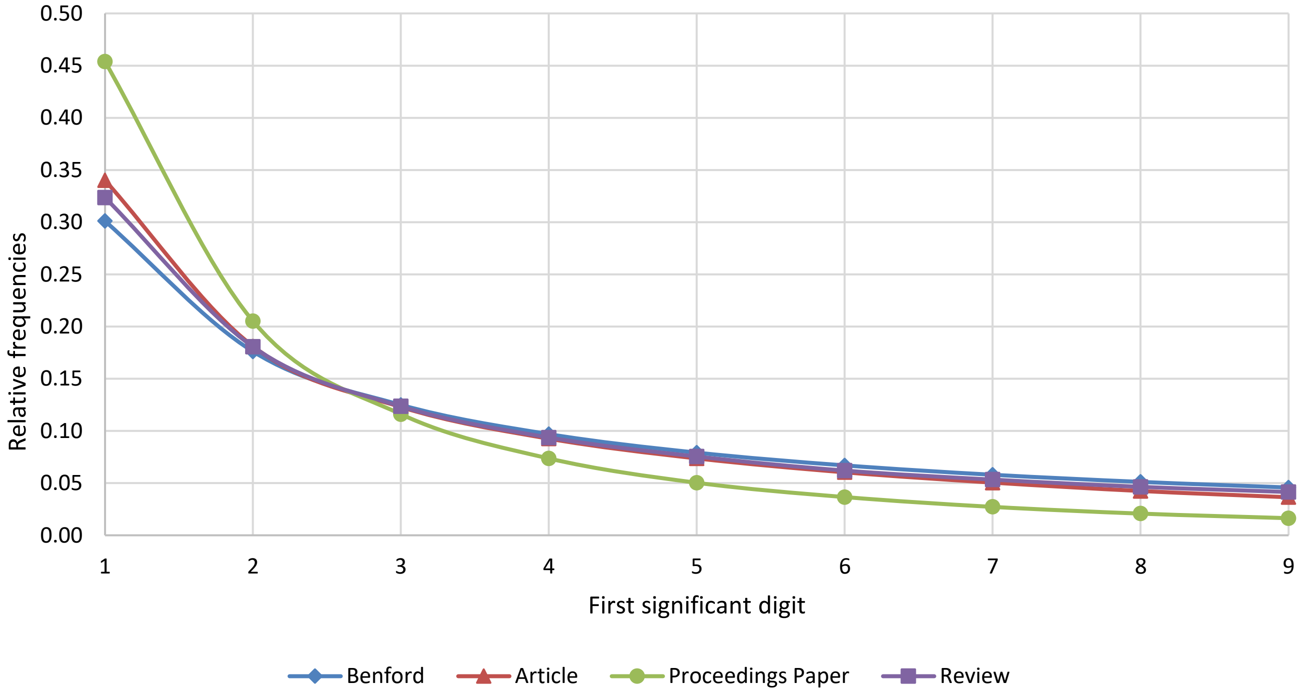

Figure 2. *Conformity with Benford's law for relative frequencies of first significant digits and for three types of documents.*

Table 6. *Relative frequency of occurrence for first significant digit for individual years and MAD and SSD results.*

Analysis – Factor of document type												
1st significant digit	1	2	3	4	5	6	7	8	9	**AVG TC**	**MAD**	**SSD**
Benford	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046			
Article	0.340	0.181	0.123	0.092	0.074	0.060	0.050	0.042	0.036	10.89	0.0099	18.8
Proceedings paper	0.455	0.205	0.116	0.073	0.050	0.037	0.027	0.021	0.016	4.50	0.0406	296.2
Review	0.324	0.181	0.124	0.094	0.075	0.062	0.053	0.046	0.041	22.79	0.0060	6.4

The results of this analysis show that the best conformity with Benford's law is achieved by the citation counts for documents of the review type (perfectly Benford). The article document type has a significantly better level of conformity with Benford's law (acceptably close) than the proceedings paper document type (non-Benford), but worse than the review document type, see also Figure 2.

4.4 Analysis addressing factor of Web of Science category – RQ4

This analysis contains the complete dataset divided by Web of Science categories. This approach allows determining conformity with of Benford's law between categories. Categories satisfying $F_{diff} > 2$ are considered in this analysis. The tables with the results are attached in (Šlosar, 2025). The table (sheet Selected) contains 236 Web of Science categories. Note that 18 Web of Science categories were excluded from this table due to non-compliance with the condition $F_{diff} > 2$. In this analysis, only data form sheet Selected are used.

As shown in Table 7, the most commonly represented level of conformity in the results is acceptably close. An examination of (Šlosar, 2025) shows that the best agreement is predominantly achieved by STEM (science, technology, engineering and mathematics) sciences. In the top 30 disciplines as ranked by the MAD value (ascending order), there is only one discipline from SSH (social sciences & humanities), namely behavioural sciences. On the other hand, the last positions according to MAD value predominantly come from SSH fields and, surprisingly, also from fields related to information technology.

Table 7. *Numbers of (sheet Selected) categories in SSD and MAD levels of conformity.*

Level of conformity	**MAD**	**SSD**
Perfectly Benford	33 (13.98%)	0 (0%)
Acceptably close	105 (44.49%)	138 (58.47%)
Marginally Benford	18 (7.63%)	62 (23.27%)
Non-Benford	80 (33.9%)	36 (15.25%)

5 DISCUSSION

The results of the analysis of the complete dataset show that the citation counts are in conformity with Benford's law (RQ1). Although the conformity of citation counts with Benford's law is classified as marginally Benford, it is still a reasonably good agreement. The rule that the relative frequency of a digit decreases as the value of the digit increases is followed. The graphs show a logarithmic curve for the relative frequencies of the first significant digits of the citation counts. This curve is very similar to the curve of probabilities according to Benford's law, but it is "rotated" in favour of the digits one and two.

It is problematic to rigorously compare the results of the present study and previous research. The present study examined conformity of the citation counts with Benford's law, whereas previous research has investigated conformity with journal indicators (Campanario & Coslado, 2011; Alves et al., 2016) or contribution of authors (Tošić & Vičič, 2021). Also, chi-square was used in some cases to assess the measure of conformity, instead of the recommended MAD (Nigrini, 2012) or SSD (Kossovsky, 2021). As a result of previous studies, it is concluded that scientometric data follow Benford's law reasonably well. Thus, the results of the present study are consistent with and extend previous research. The answer to this question revealed that Benford's law also applies to citation counts, confirming the theoretical derivation made by Egghe (2011).

The results of this study are most comparable to Gupta (2024), as it is the only study to contain an analysis of conformity of numbers of citations with Benford's law. They examined conformity of citation counts with Benford's law for some records published in 2021 from the Dimensions database. The evaluation of conformity was done only graphically, yet it provides relevant information. In both the present study and Gupta (2024), the citation counts follow Benford's law quite well. In the present study, the first digits one and two are represented more frequently than Benford's law predicts, while higher digits are represented less frequently. In Gupta (2024), the digits one, two and three were represented more frequently than predicted by Benford's law. In both studies, a sort of "rotation" of the distributions in favour of the frequencies of the lower-valued digits is evident. This is probably due to the fact that large parts of the datasets are made up of records that are cited only once (in the present study, 16.78% of the records).

The results suggest that conformity worsens slightly the more recently the records are released (RQ2). According to Bertin & Lafouge (2025), the conformity should not change significantly over time. On the contrary, the present study revealed an unexpected result, that the most recent examined year, 2018, shows dramatically worse conformity of citation counts with Benford's law compared to 2014-2017. This is due to too short a time gap in obtaining the data. This behaviour clearly shows that at least a two-year gap is needed when determining the conformity of citation counts of scientific outputs with Benford's law. "When measuring citation frequencies, the temporal dimension or time window is important. Usually, articles that have been published recently have hardly been cited yet and the number of citations increases over time as older papers have had more time to accrue citations." (Aksnes et al., 2019). The data show that the average citation rate of a record decreases the closer the year of publication is to the year of data collection. We consider this result exceptionally valuable, as the necessary time to accrue citations in the context of conformity with Benford's law has not yet been established.

The time window in the present study provides data from the pre-COVID period, downloaded at a time when COVID-19 did not yet have a potential impact on publishing practices (e.g., conference cancellations). This approach provided a basis for further potential research that would investigate, for example, differences in the pre-COVID and post-COVID data.

It has been shown that the type of document affects the level of conformity of the citation counts with Benford's law (RQ3). The review document type, which is the most intensively cited document type, achieves high conformity with Benford's law. As our analysis results showed, the average citation rate and conformity with Benford's law are correlated. The review document type is heavily cited (Miranda & Garcia-Carpintero, 2018). This citation "density" could be a precursor to good conformity with Benford's law. The low citation density and long indexing lag of the proceedings paper document type are some of the reasons for the low conformity of this document type with Benford's law. Thus, the representation of document types in the journals examined in this way should be taken into account when evaluating the conformity of journal citation counts with Benford's law.

It was also found that there are significant differences in the level of conformity with Benford's law among the Web of Science categories (RQ4). A surprising result is that when the disciplines are sorted according to MAD values, the disciplines with the lowest values are predominantly natural and medical sciences and those with the highest values are arts and humanities. We suggest that the level of conformity with Benford's law may be a useful tool in determining which disciplines are suitable for the use of scientometric methods to provide a basis for evaluation of science. The results suggest that the better the conformity of the Web of Science category citation counts, the more suitable the category is for analysis by scientometric methods. In this context, the present study provides a tool that allows a more detailed classification of disciplines than the commonly used STEM (science, technology, engineering and medicine) versus HASS (humanities, arts and social sciences) or FORD 6 disciplines (Ochsner, 2021).

Considering the results of the analysis of the factor of document type and the factor of Web of Science categories, it is likely that citation intensity has the strongest positive effect on the conformity of the citation counts of datasets with Benford's law.

Finally, the results of our empirical data analysis support the validity of Egghe's theoretical derivations of Benford's law from Zipf-Mandelbrot law (Zipf, 1949; Mandelbrot, 1965) and Lotka's law (Lotka, 1926).

The results of this study could be applied in further research into predatory journals. We believe that the level of conformity with Benford's law for the number of citations of relevant outputs published in a journal in a certain field should be similar to the conformity of the number of citations of all scientific output in that field. At the same time, it is necessary to consider what types of documents a given journal consists of. Journals that publish predominantly the review document type are more likely to achieve higher citation count conformity with Benford’s law for the records of that journal. In the case of predatory journals, we expect that the conformity of citation counts with Benford's law for documents published in a given journal will deviate from the conformity that would be expected in the field and for a given composition of document types. However, this has yet to be tested.

In this paper, the conformity levels with Benford’s law for each Web of Science category were mapped in detail. Using the results of conformity of these Web of Science categories, it is possible to estimate the level of conformity with Benford’s law for the citation rates of the content of scientific journals. Our hypothesis is that predatory journals will deviate from Benford's law in their citation counts.

6 LIMITATIONS

Data were downloaded from a single citation database. The level of conformity with Benford’s law for Scopus, OpenAlex or Dimension databases might be different from the Web of Science used here. In addition, data were downloaded over a five-year period, the length of which seems insufficient to investigate the whole effect of the time lag between the publication of a document and its citation data download. A wider time window would be more appropriate.

It is important to recognize that citation databases are dynamic in nature. If records from the publication years 2014-2018 were to be downloaded at present, the Web of Science database would likely contain a greater number of such records, and these would have accrued a higher citation count.

7 CONCLUSION

This study explored the conformity of citation counts with Benford's law. The conformity was investigated on a dataset of citation counts from the Web of Science database, consisting of 8.47 million records published between 2014 and 2018, with the article, proceedings paper and review document types. The MAD and SSD metrics recommended by the literature were used to measure the distance of empirical data from Benford’s law. The results show that the citation counts conform to Benford’s law quite well.

The effects of three factors on the level of conformity of the citation counts with Benford’s law were also investigated. A table, see (Šlosar, 2025), was provided showing the conformity of citation counts with Benford’s law for each Web of Science category. Citation intensity was shown to have a visible positive effect on conformity with Benford’s law. The review document type or Web of Science categories with high citation dynamics achieved high conformity with Benford’s law.

Based on the results, it is recommended to have at least a two-year gap between the date of downloading and the year of publication of records whose citation counts are to be examined for conformity with Benford's law.

Possible practical applications for further research include a tool for detecting predatory journals or determining whether a field is suitable for scientometric methods in evaluation of science. The results of this study confirm the theoretical assumptions and provide a basis for further research into Benford’s law in scientometrics.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding: The work was supported by the grant SVV 2025-260788 realized at the Charles University, Faculty of Arts.

Conflict of Interests: The author declares no conflict of interest.

Author Contributions: The author confirms being the sole contributor of this work.

Statement on the Use of Artificial Intelligence Tools: The author declares that artificial intelligence tools were used in the process of creating this article. Specifically, two tools were used: DeepL Translate and Microsoft Copilot M365.

These tools were used to consult on the phrasing of some sentences to increase the readability of the text. The author estimates the overall rate of sentences consulted in this way to be less than 10%.

Data Availability: Data cannot be made public due to the terms of the contract with Clarivate Plc. However, the procedure can be replicated after accessing the Web of Science database. The data that support the findings of this study are available from https://doi.org/10.5281/zenodo.16935993.

REFERENCES

Alves, A.D., Yanasse, H. H., & Soma, N. Y. (2014). Benford's Law and articles of scientific journals: Comparison of JCR® and Scopus data. *Scientometrics*, 98(1), 173–184. https://doi.org/10.1007/s11192-013-1030-8

Alves, A.D., Yanasse, H. H., & Soma, N. Y. (2016). An analysis of bibliometric indicators to JCR according to Benford's law. *Scientometrics*, 107(3), 1489–1499. https://doi.org/10.1007/s11192-016-1908-3

Aksnes, D. W., Langfeldt, L. & Wouters, P. (2019) Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. Online. *Sage Open*, 9(1), 1–17. https://doi.org/10.1177/2158244019829575

Bantange, Ch., Burgett, D., Haws, L., & Nelson, S. P. (2023). The "Benfordness" of Bach Music. *Journal of Humanistic Mathematics*, 13(2), 389–397. https://doi.org/10.5642/jhummath.SGFV8169

Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American philosophical society*, 78(4), 551–572.

Berger, A., Hill, T. P., & Rogers, E. (2009). Benford Online Bibliography. http://www.benfordonline.net

Berger, A., & Hill, T. P. (2015). *An introduction to Benford's law*. Princeton University Press.

Bertin, M., & Lafouge, T. (2025) Categorization of scientometric data in a Benfordian context. Online. *Quantitative Science Studies*, 6, 524–545. https://doi.org/10.1162/qss_a_00361

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45–80. https://doi.org/10.1108/00220410810844150

Campanario, J. M., & Coslado M.A. (2011). Benford's law and citations, articles and impact factors of scientific journals. *Scientometrics*, 88(2), 421–432. https://doi.org/10.1007/s11192-011-0387-9

Cerqueti, R., & Maggi, M. (2021). Data validity and statistical conformity with Benford's Law. *Chaos, Solitons & Fractals*. 144, 110740. https://doi.org/10.1016/j.chaos.2021.110740

Cleary, R., & Thibodeau, J. C. (2005). Applying Digital Analysis Using Benford's Law to Detect Fraud: The Dangers of Type I Errors. *Auditing: A Journal of Practice & Theory*, 24(1), 77–81. https://doi.org/10.2308/aud.2005.24.1.77

Crespo, J.A., Li, Y., Ruiz-Catillo, J., Bornmann, L. (2013). The Measurement of the Effect on Citation Inequality of Differences in Citation Practices across Scientific Fields. *PLoS ONE*, 8(3), e58727. https://doi.org/10.1371/journal.pone.0058727

Druica, E., Oancea, B., & Valsan, C. (2018). Benford's law and the limits of digit analysis. *International Journal of Accounting Information Systems*, 31, 75–82. https://doi.org/10.1016/j.accinf.2018.09.004

Egghe, L. (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Elsevier Academic Press.

Egghe, L. (2011). Benford's law is a simple consequence of Zipf's Law. *ISSI Newsletter*, 7(3), 55–56.

Egghe, L., & Guns, R. (2012). Applications of the generalized law of Benford to informetric data. *Journal of the American Society for Information Science and Technology*, 63(8), 1662–1665. https://doi.org/10.1002/asi.22690

Gupta, S., Singh, V.K., & Banshal, S.K. (2024). Altmetric data quality analysis using Benford's law. *Scientometrics*, 129, 4597–4621. https://doi.org/10.1007/s11192-024-05061-9

Kennedy, A.P., & Yam, S.C.P. (2020). On the authenticity of COVID-19 case figures. *PLoS ONE*, 15(12), e0243123. https://doi.org/10.1371/journal.pone.0243123

Kossovsky, A.E. (2015). *Benford's Law: theory, the general Law of relative quantities, and forensic fraud detection applications*. World Scientific.

Kossovsky, A.E. (2021). On the Mistaken Use of the Chi-Square Test in Benford's Law. *Stats*, 4(2), 419–453. https://doi.org/10.3390/stats4020027

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Sage.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12), 317–324.

Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In B.B. Wolman & E. Nagel (Eds.), *Scientific psychology*, (pp. 550–562). Basic Books.

Merton, R.K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Miranda, R., & Garcia-Carpintero, E. (2018). Overcitation and overrepresentation of review papers in the most cited papers. *Journal of Informetrics*, 12(4), 1015–1030. https://doi.org/10.1016/j.joi.2018.08.006

Newcomb, S. (1881). Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, 4, 39–40.

Nigrini, M. J. (2012). *Benford's Law: applications for forensic accounting, auditing, and fraud detection*. Wiley.

Ochsner, M. (2021) Bibliometrics in the Humanities, Arts and Social Sciences. In *Handbook Bibliometrics*, (pp. 117–124). Walter de Gruyter.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today*, 58, 49–54.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, 8(4), 917–934. https://doi.org/10.1016/j.joi.2014.09.006

Solla Price, D. (1965). Networks of scientific papers. *Science*, 149, 510–515.

Sorour, M. A., Marey, Y. A., Halim I. T. A., & Kasem M. M. (2024) Statistical Investigation of Scientific Journals Impact Factors in Relation to Benford's Law. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES),* (pp. 521–524). IEEE. https://doi.org/10.1109/NILES63360.2024.10753145

Šlosar, D. J. (2025). Dataset of first significant digits citation counts of documents from WoS Categories 2014–2018. *Zenodo.org*. https://doi.org/10.5281/zenodo.16935993

Tošić, A., & Vičič, J. (2021). Use of Benford's law on academic publishing networks. *Journal of Informetrics*, 15(3), 101163. https://doi.org/10.1016/j.joi.2021.101163

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic │ eISSN: 1805-4951