

An analysis of bibliometric indicators to JCR according to Benford's law

Alexandre Donizeti Alves¹ · Horacio Hideki Yanasse² · Nei Yoshihiro Soma¹

Received: 14 January 2016/Published online: 15 March 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Journal Citation Reports (JCR) is the main source of bibliometric indicators known by the scientific community. This paper presents the results of a study of the distributions of the first and second significant digits according to Benford's law (BL) of the number of articles, citations, impact factors, half-life and immediacy index bibliometric indicators in journals indexed in the JCR Sciences and Social Sciences Editions from 2007 to 2014. We also performed the data analysis to country's origin and by journal's category, and we verified that the second digit has a better adherence to BL. The use of the second digit is important since it provides a more sound, complete and consistent analysis of the bibliometric indicators.

Keywords Journal Citation Reports · JCR · Bibliometric indicators · Benford's law

Introduction

Bibliometric indicators have increasingly becoming of interest for they can be helpful in providing some measurement of the visibility of scientific publications. Bibliometric indicators can provide support in the evaluation of research success, the impact in the scientific community, and to research policy optimization. They can also help researchers in selecting the journals to which to submit their manuscripts (Durieux and Gevenois 2010). Funding agencies have also been looking at bibliometric indicators as they can offer

Federal University of São Paulo – UNIFESP, Avenida Cesare Mansueto Giulio Lattes, 1201 – Eugênio de Mello, São José dos Campos, SP 12231-280, Brazil



Aeronautics Institute of Technology – ITA, Praça Marechal Eduardo Gomes, 50 – Vila das Acácias, São José dos Campos, SP 12228-900, Brazil

quantitative measures to the results of the investment made in science (Cabezas-Clavijo et al. 2013).

As far as we know, the main source of bibliometric indicators known by the scientific community is the Journal Citation Reports (JCR). According to Thomson Reuters (2016), the owner of JCR, "Journal Citation Reports offers a systematic, objective means to critically evaluate the world's leading journals, with quantifiable, statistical information based on citation data. By compiling articles' cited references, JCR helps to measure research influence and impact at the journal and category levels, and shows the relationship between citing and cited journals". JCR contains several bibliometric indicators, which can reveal information about the performance of each journal and annually it has the Science and Social Sciences editions.

The bibliometric indicators listed in the JCR have already been used in many studies in the last years (Vanclay 2012; Sangwal 2013; Campanario 2014, 2015). Given the interest in bibliometric indicators and the information that can be extracted from the JCR database, in this paper we investigate whether the main bibliometric indicators listed in the JCR comply with Benford's law (BL).

BL is the empirical observation that in many data sets the significant digits are not uniformly distributed, as one might expect, but instead they tend to follow a very particular logarithmic distribution (Berger and Hill 2015). One of the applications of BL has been to aid researchers and professionals in identifying eventual anomalies in data sets, such as, financial data of religious community (Clippe and Ausloos 2012), the quality of occupational hygiene (De Vocht and Kromhout 2013), aggregated income taxes of Italian municipalities (Mir et al. 2014), birth time series (Ausloos et al. 2015), natural climatic process (Joannes-Boyau et al. 2015) and to distinguish different chaotic processes from stochastic processes (Li et al. 2015).

To the best of our knowledge, Campanario and Coslado (2011) was the first application of BL to bibliometric indicators listed in the JCR. They noted that the number of articles published, citations received to journals and impact factors of journals indexed in the JCR Science Edition from 1998 to 2007 not always comply with BL. They identified the first significant digit of each one of these indicators for each year separately, and compared them to the numbers predicted by BL.

In Alves et al. (2014), we extended the work of Campanario and Coslado (2011) analyzing the distribution of the first significant digit of the number of articles published of journals indexed in the JCR Science and Social Sciences Editions from 2007 to 2011. We also investigated their compliance with BL analyzing the number of articles published according to the country of origin and to the journal's category.

Other studies that analyses bibliometric indicators using BL (first digit) are of Egghe and Guns (2012) and Hürlimann (2015). The former, introduced a generalization of BL based on the same data used by Campanario and Coslado (2011), while the latter suggests a truncated Erlang distribution and he partially used data given in Campanario and Coslado (2011) and Alves et al. (2014) to illustrate his new approach to some bibliometric indicators.

In this paper, we study the distributions of the first and second significant digits according to BL of several bibliometric indicators in journals indexed in the JCR Sciences and Social Sciences Editions from 2007 to 2014. The indicators considered are: number of articles, citations, impact factors, half-life and immediacy index. We also investigate their compliance with BL analyzing them indicators according to the country of origin and to the journal's category.



Benford's law

BL, also known as the first digit law, is a logarithmic distribution function used to predict the first significant digit in numerical data. It asserts that the leading significant digit is not equally likely to be any one of the nine possible digits, but it is 1 more than 30 % of the time, and it is 9 less than 5 % of the time, with the probability of occurrence decreasing logarithmically in value as the digit increases from 1 to 9.

This was first observed by Newcomb (1881), who noted that the first pages of his book of logarithmic tables were more worn than the latter pages, which indicated that tables of logarithms were not used in a uniform way. However, only in 1938 the law was referred to as BL when Benford published a paper (Benford 1938) analyzing diverse data sets.

The general BL (Berger and Hill 2015) that specifies the probabilities of occurrence of the joint distribution of all the m significant digits is:

$$P(D_1 = d_1, D_2 = d_2, ..., D_m = d_m) = \log_{10} \left(1 + \left(\sum_{j=1}^m d_j 10^{m-j} \right)^{-1} \right), \tag{1}$$

for every positive integer m, where d_1 is in $\{1, 2, ..., 9\}$ and d_j is in $\{0, 1, ..., 9\}$ for all $j \ge 2$.

Table 1 gives the expected proportions of BL for the first and second digits (generally referred to as the 1BL and 2BL), based on Eq. (1).

Materials and methods

We get the data for this study from the JCR (Science and Social Sciences Editions) database available on the Web, covering the period from 2007 to 2014. Initially, we collected the following data for each journal: title, ISSN, year, edition, country of origin and the journal's category. It is worth noting that a journal can belong to one, two, or more JCR categories. The following bibliometric indicators for all journals indexed in the JCR in both editions were also collected: number of articles, total of citations and self-citations,

Table 1 Expected proportions of BL for the first and second digits

Digit (d)	P (d) First digit	P (d) Second digi		
0	_	0.11968		
1	0.30103	0.11389		
2	0.17609	0.10882		
3	0.12494	0.10433		
4	0.09691	0.10031		
5	0.07918	0.09668		
6	0.06695	0.09337		
7	0.05799	0.09035		
8	0.05115	0.08757		
9	0.04576	0.08500		
Total	1	1		



2-year impact factor (IF) with self-citations (2Y-IF) and without self-citations (2Y-IF_{WSC}), 5-year IF with self-citations (5Y-IF), cited and citing half-life, and immediacy index.

Figure 1 shows the total number of journals indexed in the JCR Science and Social Sciences Editions from 2007 to 2014. The actual number of journals considered in this study is smaller than the total available in the JCR, since we considered only those whose bibliometric indicators values are greater than zero. In addition, in the case of cited and citing half-life, all values equal to ">10.0" cannot be included in the study because it is not possible to obtain their significant digits directly. For these reasons, the number of journals considered varies for each bibliometric indicator. All impact factors where multiplied by 1000 and the half-life by 10 to avoid values smaller than 1.

We then identified the first and second significant digits of the bibliometric indicators of each journal indexed in the JCR, for each edition, to calculate their frequencies and, we compared them with the numbers predicted by BL. We also determined the Chi square test for 1BL and 2BL:

$$\chi^{2}(n-1) = \sum_{i=k}^{n} \frac{(N_{o}(d_{i}) - N_{e}(d_{i}))^{2}}{N_{e}(d_{i})}, \quad k = 1 \text{ for 1BL and } k = 0 \text{ for 2BL};$$
 (2)

where $N_{\rm o}(d_i)$ and $N_{\rm e}(d_i)$ are the observed values from the data and expected values according to BL, respectively. For n=9 we have n-1=8 degrees of freedom, and $\chi^2(8)=15.507$ for a 95 % confidence level. This is the critical value for the acceptance or rejection of the *Null Hypothesis*. For n=10 we have n-1=9 degrees of freedom, and $\chi^2(9)=16.919$ for a 95 % confidence level.

We also tested each of the proportions separately using the Z-statistic, which verifies whether the observed proportion for a digit differs significantly from the expected value based on BL (Nigrini 2012). The Z-statistic has the following equation:

$$Z = \frac{|P_{\rm o} - P_{\rm e}| - \left(\frac{1}{2N}\right)}{\sqrt{\frac{P_{\rm e}(1 - P_{\rm e})}{N}}},\tag{3}$$

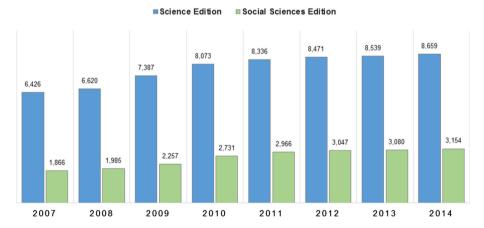


Fig. 1 Total number of journals indexed in the JCR from 2007 to 2014



where $P_{\rm o}$ denotes the observed proportion value, $P_{\rm e}$ the expected proportion value, and N is the total numbers of observations. The term (1/2N) is a continuity correction factor and it is considered only when it is smaller than the other term in the numerator. For a significant level of 5 %, the cutoff level is 1.96. The Z-statistic is important to recognize and indicate which numbers need further investigation.

Results and discussion

In Alves et al. (2014), we analyzed only the distribution of the first significant digit of the number of articles published of journals indexed in the JCR from 2007 to 2011. In this study, we extend our analysis to the second significant digit, and we investigate several other bibliometric indicators listed in the JCR from 2007 to 2014. In Fig. 2 we present a summary of our analysis. In the case of the number of articles, there is a small difference in relation to the values presented in Alves et al. (2014) because we considered now a larger number of decimal places in the calculation of the Chi square values. However, this difference did not affect the compliance (or not compliance) with BL in relation to our previous work.

According to all 288 tests performed, it is possible to observe that 191 (66.32 %) do not comply with BL. The result for each edition of the JCR is almost equal, 95 (65.97 %) for Science Edition and 96 (66.67 %) for Social Sciences Edition. By considering 1BL, 112 (77.78 %) do not comply with BL, while for 2BL 79 (54.86 %) do not comply with BL, which is a better result.

	Science Edition						Social Sciences Edition					
	1BL	2BL	1BL	2BL	1BL	2BL	1BL	2BL	1BL	2BL	1BL	2BL
	Article		Citation		Self-Citation		Article		Citation		Self-Citation	
2007		19.07	8.33	8.12	4.27	2.74			4.26	12.63	4.02	4.97
2008		11.57	15.52	10.41	12.40	7.64			8.93	14.17	14.27	6.99
2009	34.31	5.07	29.45	14.36	6.14	7.61		18.39	7.25	2.85	2.95	10.13
2010	44.97	7.51	20.29	7.90	5.09	8.02		12.19	7.13	7.19	15.59	10.81
2011	43.25	27.30	10.26	13.85	8.99	13.50			9.57	13.46	14.78	7.64
2012		12.98	3.06	18.25	7.80	6.10		7.76	7.34	12.15	13.92	6.76
2013		14.54	5.64	16.50	7.16	10.58		16.87	8.57	5.96	14.03	11.66
2014	44.70	11.30	15.53	6.69	10.51	9.18	335.21	24.19	2.99	4.40	17.59	7.74
	2Y-IF		2Y-IF _{wsc}		5Y-IF		2Y-IF		2Y-IF _{wsc}		5Y-IF	
2007	41.28	7.24	38.62	10.30	92.45	11.44		15.57	14.81	20.10	37.54	11.00
2008		13.32	57.07	33.46	145.26	7.41			30.78	16.22	61.82	21.42
2009		17.91	78.24		179.62	2.65			44.29		80.59	12.76
2010			48.41	19.28	159.90	10.84			48.09		78.24	2.68
2011		26.71	70.15	17.38	167.70	5.92			28.65		78.01	5.63
2012	90.26	21.23	66.89	16.60	146.72	7.45			36.78		88.16	9.72
2013		16.60	96.30	13.29	140.04	13.82			54.17		75.10	9.48
2014	104.04	17.45	89.68	27.21	127.15	4.28	54.86	24.60	27.48	27.53	67.41	8.03
	Cited Half-Life		Citing Half-Life		Immediacy Index		Cited Half-Life		Citing Half-Life		Immediacy Index	
2007	4,813.99	99.20	10,965.70		29.21	22.97			3,449.90		6.49	
2008		114.38	11,687.41		33.13	15.15			3,902.81		5.99	
2003			13,398.59		39.98	8.07			4,572.59		22.11	
2010			14,917.60	155.73	64.35	25.68			5,503.78		8.89	
2011	5,205.18		16,063.36	166.30	42.47				5,869.77		15.09	
2012			16,559.10		48.91	27.27			6,117.84		22.82	
2013	5,637.26	154.76	16,972.45	152.28	89.78	18.76			6,165.98		14.00	
2014	5,765.54	160.82	17,566.24	168.49	95.26	36.78	3,163.14	38.21	6,570.56	99.08	17.85	46.21

Fig. 2 Chi square values for the 1BL and 2BL of the bibliometric indicators in journals indexed in the JCR from 2007 to 2014 (the *shaded numbers* indicate noncompliance with BL)



The average percentage of journals considered varied from an indicator to another. For instance, the total number of articles to the Science Edition is 97.82 % of the total number of journals for 1BL and 96.40 % for 2BL. For the Social Sciences Edition, the average percentage is very similar, 97.86 % for 1BL and 95.50 % for 2BL. The percentages are similar for the citations, self-citations, 2Y-IF, 2Y-IF $_{\rm WSC}$, 5Y-IF and immediacy index. However, for the cited and citing half-life the percentages are smaller than 77 % in the Science Edition and smaller than 63 % in the Social Sciences Edition.

Number of articles

Campanario and Coslado (2011) noted that the number of articles published in journals indexed in the JCR Science Edition do not comply with 1BL from 1998 to 2007. Alves et al. (2014) also noted the same in journals indexed in the JCR Science and Social Sciences Editions from 2007 to 2011. The same occurred from 2007 to 2014, but for the 2BL, some Chi square values in some years are smaller than the critical value. For instance, only 2 years (2007 and 2011) do not comply in the JCR Science Edition.

The number of articles published in a journal varies a little from 1 year to another, since the number of issues in a journal seldom changes yearly. This may explain the noncompliance of the first digit and the slight improvement on the second digit for the number of articles published.

We also investigated the journals according to their country of origin and to their JCR category. In Fig. 3 we present the percentage of countries and journal's categories that complies with 1BL and 2BL considering the Chi square values for the number of articles published in journals indexed in the JCR Science and Social Sciences Editions from 2007 to 2014. The results for the 1BL are generally very good, for 2BL the results are always better than 1BL. An outlier occurred for the 1BL in the journal's category in the JCR Social Sciences Edition. In this case, the average percentage is 52.52 % for 1BL and 93.04 % for 2BL.



Fig. 3 Percentage of countries and journal's categories that complies with 1BL and 2BL for the number of articles published in journals indexed in the JCR



Citation

The number of citations indicate the total number of times that each journal was cited by all journals included in the database within the current JCR year. For this indicator we observed that almost all data complies with 1BL, being the result still better for 2BL, that has a sole exception of non-compliance in the year 2012. Considering the self-citations, all years comply with 2BL. We also considered the number of citations without self-citations and to 2BL they all comply with BL. For the 1BL, non-compliance was observed in the years from 2009 to 2011 in the Science Edition. In 2009 and 2010, the Chi square values are smaller compared to those of the indicator with self-citations; nevertheless, they do not obey BL.

Considering the country of origin and the journal's category for the citations and self-citations, both 1BL and for 2BL have a very high average acceptance percentage, around of 95 %, for the two JCR editions. For all cases, except by a small difference the journal's category in Social Sciences Edition for self-citations, the percentage is higher for 2BL compared to 1BL.

Impact factor

The IF identifies the frequency with which an average article from a journal receives citations by other articles in a particular period. To 2Y-IF, the last 2 years are considered and for the 5Y-IF the last 5 years. To the IF values we noticed that almost all of them do not follow 1BL, expect the year 2007 in the Social Sciences Edition for 2Y-IF_{WSC}. Additionally, the 5Y-IF do not follow 1BL, but it follows the 2BL, except in 2008 in the Social Sciences Edition. This is interesting since neither 2Y-IF nor 2Y-IF_{WSC} follow BL in almost all years. It provides support for the preferential use of the 5Y-IF over the 2Y-IF or 2Y-IF_{WSC}.

In Fig. 4 we present the percentage of countries and journal's categories that complies with 1BL and 2BL considering the Chi square values for the impact factors in journals indexed in the JCR from 2007 to 2014. It is possible to observe that there is a significant difference from 2BL to 1BL in all cases, with an overall average percentage for compliance with 2BL around 95 %.

Half-life and immediacy

The cited half-life gives the number of years back from the current year that accounts for 50 % of the total number of citations to a journal. The citing half-life identifies the number of years from the current year that accounts for 50 % of the cited references from articles published by a journal. The immediacy index measures how frequently, in average, an article from a journal is cited within the same year of publication and it is a useful metrics for evaluating journals that publish cutting-edge research.

We noted that not all years of the cited and citing half-life indicators comply with 1BL and 2BL for the two JCR editions. We expect that the great majority of half-life values ">10.0" is probably bounded from 10 to 20 years. For the sake of calculation we, therefore, assumed "1" to be the first digit of all the half-life values ">10.0". The new Chi square values obtained are slightly smaller but the conclusions for the first digit remain the same.

For the immediacy indicator the same pattern occurred with 2BL, but to 1BL, the Chi square values are in accordance to BL for some years.

It is worth noting the differences in percentage of compliance of 1BL and 2BL of cited and citing half-life by country and journal's categories shown in Fig. 5. We observed 1BL





Fig. 4 Percentage of countries and journal's categories that complies with 1BL and 2BL for the 2Y-IF, 2Y-IF_{WSC} and 5Y-IF in journals indexed in the JCR

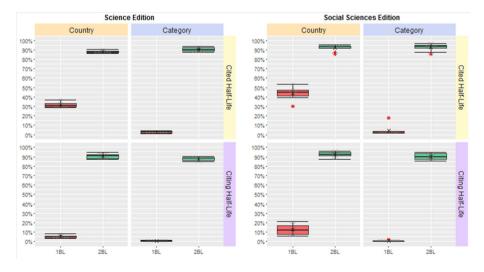


Fig. 5 Percentage of countries and journal's categories that complies with 1BL and 2BL for the cited half-life and citing half-life in journals indexed in the JCR



smaller than 1 % in some cases, while for the 2BL the percentage stays around 90 %. To the immediacy index, the average percentage is around of 90 % for the 1BL and 95 % for the 2BL, except in journal's categories in the JCR Social Sciences Edition that the average percentages are the opposite, that is, around 95 % for 1BL and 90 % for the 2BL.

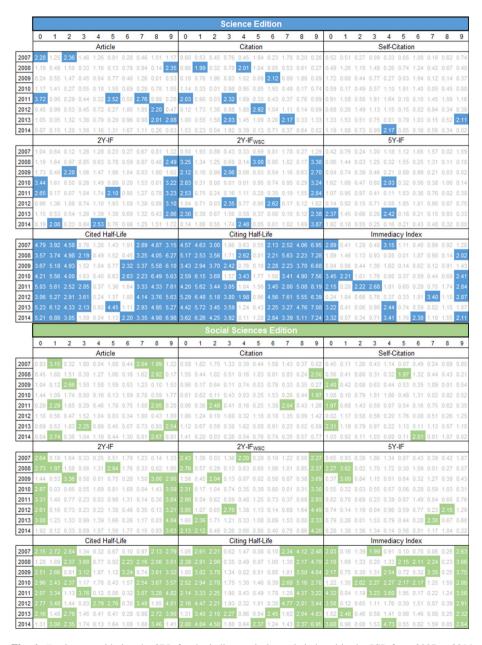


Fig. 6 Z values considering the 2BL for the indicators in journals indexed in the JCR from 2007 to 2014 (the *shaded numbers* indicate values greater than the cutoff levels)



Z-statistic

We used the Z-statistic to test each digit proportion separately, since it allows the identification of which ones may need further investigation. To 1BL, the result is that the Z values of almost all digits are greater than the cutoff level (1.96) for all bibliometric indicators but to citation and self-citations. Figure 6 presents the result to 2BL to all bibliometric indicators considered in this study. Observe that for digits 0 and 9, in almost all years, the values are greater than the cutoff value.

Conclusions

In this paper, we analyzed the number of articles, citations, impact factors, half-life and immediacy index bibliometric indicators in journals indexed in the JCR Science and Social Sciences Editions from 2007 to 2014 according to 1BL and 2BL. We also performed the data analysis to country's origin and by journal's category.

In Alves et al. (2014) we verified that for countries and for journal's categories the majority complies with 1BL. In this study, we verified that the second digit has a better adherence to BL, since the average percentage of compliance is around of 95 % to almost all bibliometric indicators. The non-compliance with 1BL of the number of articles for the period 1998–2007 was observed in Campanario and Coslado (2011), for the period 2007–2011 in Alves et al. (2014), and again in the current study for the period 2007–2014. Here we observed a slight improvement on the compliance with 2BL. This may be explained by the observation that the number of articles published in a journal varies a little from 1 year to another.

From the data analyzed an interesting result came out related to 5Y-IF. 5Y-IF has a better compliance to BL than 2Y-IF and 2Y-IF_{WSC}. This result gives support for the preferential use of the 5Y-IF over 2Y-IF or 2Y-IF_{WSC}. For both the cited and citing half-life they do not comply with 1BL and 2BL in all the years considered. The non-compliance with BL may be explained by the fact that we expect that these indicators to be quite stable over the years. The result for the immediacy index for some of the years was better for the 1BL compared with 2BL. This occurred also considering the journal's category for the Social Sciences Edition. Finally, the result for citation and self-citations is very good, for both 1BL and for 2BL.

This study indicates that to consider the country of origin and journal's category is relevant either to 1BL or to 2BL. The use of the second digit is important since it provides a more sound, complete and consistent analysis of the bibliometric indicators.

Further studies can be carried out using the generalized law of Benford (Egghe and Guns 2012).

Acknowledgments The authors acknowledge the financial support of CAPES and CNPq. We thank the suggestions of the anonymous reviewers that improved the presentation of the paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



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.
- Ausloos, M., Herteliu, C., & Ileanu, B. (2015). Breakdown of Benford's law for birth data. *Physica A: Statistical Mechanics and its Applications*, 419(1), 736–745.
- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572.
- Berger, A., & Hill, T. P. (2015). An introduction to Benford's law. Princeton: Princeton University Press. Cabezas-Clavijo, Á., Robinson-García, N., Escabias, M., & Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? PLoS One, 8(6), e68258.
- Campanario, J. M. (2014). The effect of citations on the significance of decimal places in the computation of journal impact factors. Scientometrics, 99(2), 289–298.
- Campanario, J. M. (2015). Providing impact: The distribution of JCR journals according to references they contribute to the 2-year and 5-year journal impact factors. *Journal of Informetrics*, 9(2), 398–407.
- Campanario, J. M., & Coslado, M. A. (2011). Benford's law and citations, articles and impact factors of scientific journals. Scientometrics, 88(2), 421–432.
- Clippe, P., & Ausloos, M. (2012). Benford's law and Theil transform of financial data. *Physica A: Statistical Mechanics and Its Applications*, 391(24), 6556–6567.
- De Vocht, F., & Kromhout, H. (2013). The use of Benford's law for evaluation of quality of occupational hygiene data. *Annals of Occupational Hygiene*, 57(3), 296–304.
- Durieux, V., & Gevenois, P. A. (2010). Bibliometric indicators: Quality measurements of scientific publication. *Radiology*, 255(2), 342–351.
- 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.
- Hürlimann, W. (2015). On the uniform random upper bound family of first significant digit distributions. *Journal of Informetrics*, 9(2), 349–358.
- Joannes-Boyau, R., Bodin, T., Scheffers, A., Sambridge, M., & May, S. M. (2015). Using Benford's law to investigate Natural Hazard dataset homogeneity. Scientific Reports, 5, 12046.
- Li, Q., Fu, Z., & Yuan, N. (2015). Beyond Benford's law: Distinguishing noise from chaos. PLoS One, 10(6), e0129161.
- Mir, T. A., Ausloos, M., & Cerqueti, R. (2014). Benford's law predicted digit distribution of aggregated income taxes: The surprising conformity of Italian cities and regions. The European Physical Journal B. 87(11), 1–8.
- Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1), 39–40.
- Nigrini, M. (2012). Benford's law: Applications for forensic accounting, auditing, and fraud detection. London: Wiley.
- Sangwal, K. (2013). Some citation-related characteristics of scientific journals published in individual countries. *Scientometrics*, 97(3), 719–741.
- Thomson Reuters. (2016). Journal Citation Reports. http://wokinfo.com/products_tools/analytical/jcr/. Accessed 6 January 2016.
- Vanclay, J. K. (2012). Impact factor: Outdated artefact or stepping-stone to journal certification? Scientometrics, 92(2), 211–238.

