Lotka's Law and Bioinformatics Research in India

Dr. Manya R. Gopal

Librarian, Kendriya Vidyalaya, Pangode, Thiruvananthapuram, Kerala

Dr. K. G. Sudhier

Assistant Professor, Dept. of Library & Information Science, Central University of Tamil Nadu

Abstract

The study aims to examine the validity of Lotka's law and authorship distribution to the research outcome of Indian bioinformatics research output indexed in the web of science during 2008-2017.This study has applied Lotka's law to assess authors' productivity pattern of bioinformatics literature and further Chi- square test and Kolmogorov-Smirnov (K-S) goodness-of-fit test applied for testing of observed and expected author productivity data. A total of 8732 papers has been scattered in 1723 journals during the period were compiled from web of science database. Using 'straight count' of authors, a total of 3835 authors were identified and the Lotka's law was tested using K- S goodness- of - fit test and Chi- square test. The results indicate that the author productivity distribution of Lotka's law is applicable to bioinformatics literature, while applying K- S test and not applicable when the Chi- square test is employed. This study is significant, as there have been no such studies conducted in the area of Indian bioinformatics research output and it will definitely set a baseline for future studies on author productivity of bioinformatics research output.

Keywords: Bioinformatics, Chi-Square test, Kolmogorov-Smirnov (K-S) test, Lotka's Law, Web of Science.

1. Introduction

The computational methods for comparative analysis of genome data was referred as bioinformatics from the late 80s and this new branch has become an integral part for storage, retrieval and analysis of biological data since then (Hogeweg, 2011). Bioinformatics, in a broader sense, deals with the usage of computer science in solving biological problems. The recent years has seen enormous amount of data being generated at an unprecedented pace in the field of biology due to the fast growing technology. According to Lesk (2014) "the reduction in the cost of genome sequencing has made it possible to obtain data on almost every species and this huge data needs to be analyzed for answering various biological questions ranging from the field of disease biology, developmental biology to evolutionary biology". The immensity and complexity of the biological data has helped the bioinformatics to take shape as separate field of study and without basic knowledge of bioinformatics tools modern biological research has become impossible.

India has experienced an extensive growth of research in bioinformatics in the past decades since 1960s. The field of bioinformatics has undergone significant evolution in India. The number of publications has been increasing significantly during the years within the field of life science. The DBT, Government of India is primarily responsible for the extensive infrastructure and network that was initiated way back in 1980s for the spread of bioinformatics centres across India, and is being supported at 168 locations across the country. According to Krishnaswamy and Mohan (2016), "India was the first country to conceptualise and establish, during 1986-87, a national distributed bioinformatics network (BTISNET), which is now the largest in the world. Even as the term 'bioinformatics' was just coined, the DBT took a bold step in initiating the Biotechnology Information System Network".

Research evaluation of any organization is based on indicators of science and technology (S&T) activity. For research evaluation, bibliometrics and scientometrics tools were applied. It provides quantitative approach in science studies. The application of bibliometric laws such as Lotka's law of scientific productivity is the one of the areas of scientomeric research. Previously a number of studies have been made in to describe scientific productivity in various branches of knowledge. These studies suggest that law of scientific productivity follows an inverse law of growth and measures the frequency of occurrence of scientific output.

2. Previous studies

Pao (1986) and Nicholls (1989) studied the validity of the law, who found that the Lotka model fitted the majority of the data sets studied. Barik and Jena (2021) verified whether the authors' productivity pattern of Library and Information Science (LIS) open access journals adheres to Lotka's inverse square law of scientific productivity. The validity of Lotka's law and author productivity in the field of Economic literature was conducted by Tunga (2021).

Again Tunga (2020) studied author productivity and the application of Lotka's law in the field of Horticulture. Radhika, Thanuskodi and Sudhakar (2020) analysed marine pollution literature retrieved from the Scopus database during 1989-2018. They tested the applicability of Lotka's law by applying Chi-Square test and confirmed that the law did not fit the marine pollution research output. The Lotka's law on authorship productivity of artificial intelligence literature has been tested by Ahmad and Batcha (2019) to confirm the applicability of the law. K-S test was applied and was found that the inverse square law of Lotka follows the data set. Kumar and Kumar (2019) studied the scientific productivity pattern of authors and applicability of Lotka's law in the field of astronomy and astrophysics research in India (2013-2017) retrieved from WoS.

Batcha (2018) studied the applicability of law on the dengue global publication. The study uses Lotka's empirical law of scientific productivity, and proved that the law didn't follow the data. Kumar (2018) examined the authorship pattern of 556 papers published in the Journal of Documentation and verified the Lotka's law by applying K-S test. Found that the law was applicable in LIS literature. Da Silva (2018) described that "the law cannot be dismissed after considering a massive sample of the number of publications of Brazilian researchers in journals listed on the SCImago Journal Rank and the Journal Citation Reports". Sharma and Chakravarty (2018) studied the research of LIS by the faculties of north Indian central universities (1978-2014). It was found that the Lotka's law was applicable to the data. Dhoble and Kumar (2017) made a study on the authorship pattern and applicability of Lotka's law in the mustard research output in India. The study applied Chi-square test to validate the data and disclosed that collaboration of more number of authors per article dominates in publications activities in this research. Naqviand Fatima (2017) analysed international business literature to study the applicability of Lotka's law. Further, K-S goodness of fit test and Chi square test were applied to compare and confirm the data set and found that, Lotka's law confirmed the author productivity distribution. A bibliography record of oncology research output in India (2005-2015) in WoS database covering 10,298 research items was evaluated by Muthukrishnan and Kumar (2017), to test the applicability of Lotka's law. It was found that oncology research output conformed particularly well to Lotka's law. Sudhier (2013) in his paper analysed research output and examined the law and found that the author distribution pattern didn't found suitable. Again he conducted a study on the application of the law in the author distribution appended in the IISc physics doctoral theses (Sudhier,2010).

A number of studies on the validation of Lotka's law were available, but no studies on bioinformatics literature were found. Hence, this paper attempts to study the validity of Lotka's law in author distribution pattern of Indian bioinformatics literature.

3. Objectives

i. To study the authorship pattern of bioinformatics research output

ii. To examine the validity of Lotka's law

iii. To apply Chi- square test and Kolmogorov-Smirnov (K-S) goodness-of-fit test for testing of observed and expected author productivity data.

4. Methodology

The study was undertaken based on the data downloaded from web of science database for the period 2008-2017 using the following search strategy.

TS= (("Bioinformatics" OR "Computational biology" OR "Biology, Computational" OR "Computational Molecular biology" OR "Biology, computational molecular" OR "Biologies, Computational Molecular" OR "Computational Molecular Biologies"OR "Molecular biology, computational" OR "Molecular Biologies, Computational" OR "Bio-informatics" OR "Bioinformatic" OR "Bio Informatics" OR "Bio- Informatic" OR "Sequence analysis" OR "Genomics")) AND CU= India AND PY = 2009-2018. Refined by: [excluding] Countries/ territories: The key words used for searching was taken from MeSH (Medical Subject Headings). First author count is employed for the author counting. All the searched results are first saved in text files and then imported into Microsoft Excel latest version and Statistical Package for Social Science (SPSS) version 17.0 was used for analysis.

A total of 8732 journal articles scattered in 1723 journals have been contributed by 3835 authors during 2008-2017, were retrieved using"straight count of the authorship method". Goodness-of-fit test was done using K-S and Chi-square tests for the conformity of the Lotka's law by using SPSS and Excel software. The methods proposed by Pao (1985) was followed.

4.1 *Lotka's law of author productivity*

Lotka (1926) was the first to study the author productivity from empirical data and proposed an inverse square law relating to author's productivity distribution to understand the worth of authors in progress of a subject field. Lotka investigated the productivity patterns of authors in a sample data from physics and chemistry. The general formula is:

$$x^n y = k \qquad (1)$$

where, y is the frequency of authors making 'n' contributions each and 'k' is a constant

The Lotka's inverse square law can mathematically be written as:

$$g\,(x) = (6/p)(1/x^2),\ x = 1,2,3\ldots.. \qquad (2)$$

where, g (x) is the proportion of authors making x contributions.

A generalised form of Lotka's law was presented by Bookstein (1976) as:

$$g\,(x) = kx^{-n},\ x = 1,2,3,4\ldots x_{max},\ k>0 \qquad (3)$$

where g(x) represents the fraction of authors publishing x articles; k and n are parameters to be estimated from the data; x_{max} represents the maximum size or value of productivity variables x; and n is usually ≥1.

The procedure proposed by Pao (1985)is followed in studying the validity of the law:

a. Calculation of 'n'

In the first step, determine the value of 'n' which is to be calculated either by following using LLS regression method or by an equivalent formula.

$$n = \frac{N.\sum XY - \sum X.\sum Y}{N.\sum X^2 - \left(\sum X\right)^2} \qquad (4)$$

where N = number of pairs of data considered $x = 1,2,3 \ldots\ldots x_{max}$;

X = logarithm of articles (x) and Y= logarithm of authors (y).

b. Calculation of *k*

The value of k, which is the theoretical number of authors with a single article, is determined from the following formula:

$$k = \frac{1}{\sum_{x=1}^{p-1} 1/x^n + 1/(n-1)(p^{n-1}) + \frac{1}{2}pn + n/24\times(p-1)^{n+1}} \quad (5)$$

Here *p* is assumed to be 20 and n is the experimentally computed value of the exponent from the observed distribution. Once the value of n and *k* is determined, then using equation 3, determine the number of authors writing 1, 2,3,…x articles.

To verify the observed distribution of author productivity fits the estimated distribution, applying the Kolmolgorov-Smirnov (K-S) goodness-of-fit test. The maximum difference between the real and estimated accumulated frequencies was calculated, and this value was then compared with the critical value (cv) obtained from the following equation:

Critical Value (CV) = $1.63/\{\ \sum y_x + (\sum y_x/10)\ \}^{1/2}$

D = Dmax = Differences between the columns of the observed and expected cumulative frequencies = $\sum f(x) - \sum(y_x/\sum y_x)$.

5. Analysis and discussion

Calculation of parameter 'n'

Initially for the validation of the law is to determine the value of 'n' and is calculated by linear least square method by using the formula (4).

Table 1: Calculation of the parameter *n*

x	g(x)	ln(x)	ln(gx)	ln(x) * ln(gx)	ln(x) * ln(x)
1	3835	0	3.58	0	0
2	886	0.3	2.95	0.887	0.091
3	270	0.48	2.43	1.16	0.228
4	120	0.6	2.08	1.252	0.362
5	80	0.7	1.9	1.33	0.489
6	39	0.78	1.59	1.238	0.606
7	17	0.85	1.23	1.04	0.714
8	12	0.9	1.08	0.975	0.816
9	9	0.95	0.95	0.911	0.911
10	5	1	0.7	0.699	1
Total	5273	6.56	18.5	9.419	5.215

By substituting the values in the equation (4), the value of'n' is calculated as:

$$n=\frac{N.\sum XY-\sum X.\sum Y}{N.\sum X^2-\left(\sum X\right)^2}$$

n = -2.9

Calculation of value 'k'

The value of parameter 'n' is calculated as, n = -2.9

Substituting the given value of 'n', the value of 'k' is estimated from the table of exponents given by Rousseau (1993) as, k = 0.82.

By replacing the value of 'n' and 'k' in Lotka's model equationg (x) = kx^{-n} and the values calculated are shown in the Table 2.

5.1 *Application of K-S test*

Coile (1977) suggests the K-S test for validating the law. For applying the K- S test, convert the observed and expected number of authors into fractional values, and take the difference between cumulative fractional values of observed and expected numbers as shown in table 2.

Table 2: K-S test on observed and expected distribution of authors

x	g (x)	FOF	CFOF	FEF	CFEF	DOECF
1	3835	0.721	0.721	0.82	0.82	0.099
2	886	0.167	0.887	0.11	0.93	0.043
3	270	0.051	0.938	0.034	0.964	**0.026**
4	120	0.023	0.961	0.015	0.978	0.018
5	80	0.015	0.976	0.008	0.986	0.011
6	39	0.007	0.983	0.005	0.991	0.008
7	17	0.003	0.986	0.003	0.994	0.008

x	g (x)	FOF	CFOF	FEF	CFEF	DOECF
8	12	0.002	0.988	0.002	0.996	0.007
9	9	0.002	0.99	0.001	0.997	0.007
10	5	0.001	0.991	0.001	0.998	0.007
11	7	0.001	0.992	0.001	0.999	0.007
12	10	0.002	0.994	0.001	0.999	0.005
13	4	0.001	0.995	0	1	0.005
14	6	0.001	0.996	0	1	0.004
15	3	0.001	0.997	0	1.001	0.004
16	1	0	0.997	0	1.001	0.004
17	2	0	0.997	0	1.001	0.004
18	2	0	0.998	0	1.001	0.004
19	2	0	0.998	0	1.001	0.004
20	2	0	0.998	0	1.002	0.003
24	2	0	0.999	0	1.002	0.003
26	1	0	0.999	0	1.002	0.003
28	1	0	0.999	0	1.002	0.003
29	2	0	0.999	0	1.002	0.002
47	2	0	1	0	1.002	0.002
59	1	0	1	0	1.002	0.002

where,

g (x) : Number of authors contributing x number of papers

FOF: Fraction of observed frequency of authors

CFOF: Cumulative fraction of observed frequency of authors

FEF: Fraction of expected frequency of authors

CFEF: Cumulative fraction of theoretical frequency of authors

DOECF: Absolute difference of the observed and expected cumulative frequency of authors.

The maximum difference value, Dmax representing the maximum deviation is identified as 0.026.

The value of 'n' and 'k' were calculated to be -2.90 and 0.82 respectively. The K- S goodness of fit test was conducted at 0.05 % level significance. The D_{max} value is 0.026 and the resulting critical value is 0.264. Since critical value is greater than D_{max}, we must fail to reject the null hypothesis that the distribution is not different from the distribution predicted by Lotka's law. Hence the Lotka law is applicable to the bioinformatics publications. Therefore, the author's productivity data on bioinformatics fit modified Lotka's law with the value ?? = -2.90.

5.2 *Application of Chi-square test*

The details of the results and analysis are given in the table 3.

Table 3: Chi-Square test of observed and expected authors

x	fo	fe	fo-fe	(fo-fe)2	Chi
1	3835	3835	0	0	0
2	886	886	0	0	0
3	270	376	-106	11236	30
4	120	205	-85	7225	35
5	80	128	-48	2304	18
6	39	87	-48	2304	26
7	17	63	-46	2116	34
8	12	47	-35	1225	26
9	9	37	-28	784	21
10	5	30	-25	625	21
11	7	24	-17	289	12
12	10	20	-10	100	5
13	4	17	-13	169	10
14	6	14	-8	64	5
15	3	13	-10	100	8
16	1	11	-10	100	9
17	2	10	-8	64	6
18	2	9	-7	49	5
19	2	8	-6	36	5
20	2	7	-5	25	4
24	2	5	-3	9	2
26	1	4	-3	9	2
28	1	3	-2	4	1
29	2	3	-1	1	0
47	2	1	1	1	1
59	1	1	0	0	0
Chi-Square					286.5

Where,

x = number of papers

fo = observed number of authors

fe = expected number of authors

To find out the suitability of Lotka's law in the observed author productivity distribution, compare the calculated Chi-square value obtained 286.25 with the critical value of Chi- square at 0.05 significance is 37.65. On comparing, it is found that the calculated value of Chi-square is greater than the critical value and thus the Lotka's law does not fit in the observed given author productivity distribution of bioinformatics literature..

5.3 *Graphical representation*

The graphical representation of the author productivity data is shown in figure 1. The graph is plotted as the number of authors in X-axis and number of articles in the Y-axis.

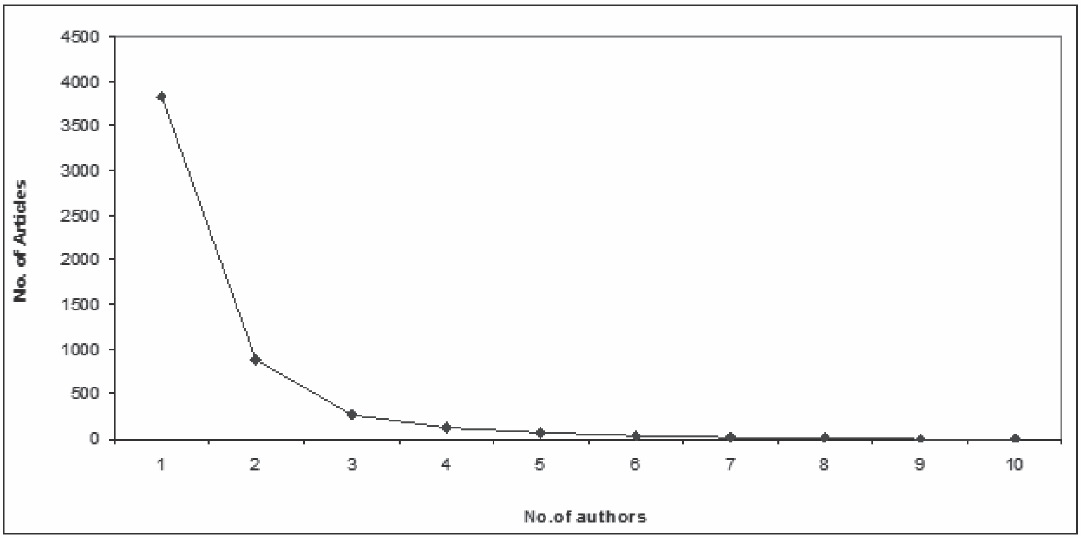

Figure 1 : Plot of number of authors and number of articles

The study shows that the Lotka's law in its generalised form follow the author productivity distribution pattern prepared for the first authors while applying K-S test. The application of the statistical tests, Chi-square shows that the author productivity distribution pattern does not fit Lotka's law.

6. Findings

Lotka's law of author productivity is regarded as one of the classical laws of bibliometrics. The present study has showed that Lotka's generalised law is not applicable to bioinformatics literature. The K-S test and chi-square test were applied to verify the applicability of Lotka's law of scientific productivity.

- The study shows that the Lotka's law in its generalised form follow the author productivity distribution pattern prepared for the first authors while applying K-S test.
- While applying the Chi-square test, it was found that the calculated value of Chi-square is greater than the critical value and thus the Lotka's law does not fit in the observed given author productivity distribution of bioinformatics literature.
- The productivity trend of authors of journal articles of bioinformatics found to be different while applying K-S test and Chi- square tests.

7. Conclusion

This study examines the validity of Lotka's law to author productivity distribution in the field of bioinformatics research in India, covered in the web of science database. A list of 8732 journals articles on various branches of bioinformatics research covered in 1723 journals during 2008-2017 was compiled for analysis. Using 'straight count' of authors, a total of 3835 personal authors were identified and the Lotka's law was tested using K- S goodness- of - fit test and Chi- square test. The results indicate that the author productivity distribution of Lotka's generalised inverse square law is applicable to bioinformatics literature, while applying K- S test and is not fit when the Chi- square test is employed.

Lotka's law of scientific productivity has been widely tested in the LIS field over the last several decades, but the results of the studies are inconclusive due to varying methods employed by the researchers. This study finds that literature in the field of bioinformatics research shows different results in confirmation with the validity of the Lotka's law. As a result, Lotka's law can be used in bioinformatics research as a standardised means of measuring author publication productivity which will lead to findings that are comparable on many levels (department, institution, national levels), as suggested by Askew (2008) in LIS studies. The results of the study would be useful for researchers, scientists and policy makers in the country in the area of bioinformatics. This may be the pioneer study in the area of Indian bioinformatics research and it may trigger more such studies for the purpose of testing Lotka's law in the various branches of bioinformatics and computational biology discipline. Future research could be directed to authorship and productivity studies in bioinformatics on various institutions in the country and on different databases. This study is significant from the viewpoint of the growing research on bioinformatics as well as various scientomtric studies in the field of LIS to identify the authorship pattern, collaboration trend and author productivity pattern of bioinformatics research output.

References

Askew, Consuella Antoinette (2008).An Examination of Lotka's law in the Field of Library and Information Studies. FIU Electronic Theses and Dissertations. Paper 182.http://digitalcommons.fiu.edu/etd/182.

Abrizah, A. & Wee, M.C. (2011). Malaysia's computer science research productivity based on publications in the web of science, 2000-2010. *Malaysian Journal of Library & Information Science,* 16 (1), 109-124.

Ahmad, Muneer & Batcha, M Sadik. (2019). Testing Lotka's Law and pattern of author productivity in the scholarly publications of artificial intelligence. *Library Philosophy and Practice (e-journal).* 2716.Retrieved fromhttps://digitalcommons.unl.edu/libphilprac/2716.

Barik, N. & Jena, P. (2021). Author productivity pattern and applicability of Lotka's inverse square law: a bibliometric appraisal of selected LIS open access journals. *Digital Library Perspectives,* 37 (3), 223-241. https://doi.org/10.1108/DLP-10-2020-0103

Batcha, S. M. (2018). Lotka's applicability on global dengue research publication: A scientometric study. *DESIDOC Journal of Library and Information Technology,* 38(4), 266-270. https://doi.org/10.14429/djlit.38.4.12361

Biotechnology Information System Network. Retrieved fromhttp://www.btisnet.gov.in/

Coile, R. C. (1977). Lotka's frequency distribution of scientific productivity. *Journal of the American Society for Information Science,* 28 (6), 366-370.

Da Silva, Sergio., Perlin, Marceelo., Matsushita, Raul., Santos, Andre A P., Imasato, Takeyoshi & Borenstein, Denis. (2019).Lotka's law for the Brazilian scientific output published in journals. *Journal of Information science,* 45(5), 705-709.https://doi.org/10.1177/0165551518801813

Dhoble, S. & Kumar, S. (2017).Applicability of Lotka's law in mustard research publications in India- a scientometric study. *SSARSC International Journal of Library, Information Networks and Knowledge,* 2(1), 1-10.

Hogeweg, P. (2011). The roots of Bioinformatics in theoretical Biology. *PLoS Computational Biology,* 7(3): e1002021. https://doi.org/10.`1371/journal.pcbi.1002021

Kawamura M, Thomas C.D., Tsurumoto, A., Sasahar, a H. & Kawaguchi, Y. (2000).

Lotka's law and productivity index of authors in a scientific journal. *Journal of Oral Science,* 42(2), 75?78. https://doi:10.2334/josnusd.42.75.

Krishnaswamy, S & Mohan, T Madhan. (2016).The largest distributed network of bioinformatics centres in the world: Biotechnology Information System Network (DBT-BTISNET). *Current Science,* 110 (4), 556-561

Kumar, P. K. Suresh. (2018). Author productivity and the application of Lotka's Law in LIS publications. *Annals of Library and Information Studies,* 64(4), 234-241.

Kumar, Satish & Kumar, R. Senthil. (2019). Applicability of Lotka's Law in Astronomy & Astrophysics Research of India. *Library Philosophy and Practice* (e-journal). 2129. Retrieved from http://digitalcommons.unl.edu/libphilprac/2129

Lesk, Arthur M. (2014). *Introduction to Bioinformatics.* Great Clarendon Street: Oxford University Press.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences,* 16 (2), 317-323.

Muthu Krishnan, M. & Kumar, R. Senthil. (2017). Author productivity of Oncology research output in India: testing Lotka's law. *International Journal of Information Dissemination and Technology,* 7(3), 187-189.

Naqvi, S. H. & Fatima, N. (2017). Authorship patterns in international business literature: Applicability of Lotka's law. *Annals of Library and Information Studies,* 64(4), 253- 259.

Nicholls, P T. (1986). Empirical validation of Lotka's law. *Information Processing & Management,* 22 (5), 417-419.

Pao, Miranda Lee. (1985). Lotka's law: A testing procedure. *Information Processing and Management,* 21 (4), 305-320.

Pao, Miranda Lee. (1986). An empirical examination of Lotka's law. *Journal of the American Society for Information Science,* 37(1), 26-33.

Radhika, N., Thanuskodi, S. & Sudhakar, K. (2020). Lotka's Law and the Pattern of Scientific Productivity in the Marine Pollution Research. *International Journal on Emerging Technologies,* 11(2), 332-341.

Rousseau, R. (1993). A table for estimating the exponent in Lotka's law. *Journal of Documentation,* 49(4), 409-12.

Sharma, Jyothi & Chakravarty, Deepak (2018).Application of Lotka's Law in Library Science literature of select Central Universities in North India. *Journal of Advances in Library and Information Science,* 7 (1), 36-39.

Sudhier, K G. (2013).Lotka's law and the pattern of author productivity in the area of physics research. *DESIDOC Journal of Library and Information Technology,* 33 (6), 457-464.

Sudhier, K.G. Pillai. (2010). Application of Lotka's law to author productivity distribution of physics literature. 6th international conference on webometrics, Informetrics and scientometrics (ICWIS) and 11th COLLNET Meeting, 2010, October 19-22, University of Mysore, Mysore.

Suresh, Kumar. (2003). Lotka's law and author productivity in the field of computer science in India. *Library Herald,* 41(2), 90-98.

Tunga, S. K. (2020). Author productivity and the application of Lotka's law in the field of Horticulture. *Library Philosophy and Practice (e-journal),* 47770. https://digitalcommons,unl.edu/libphilprac/4770

Tunga, S. K. (2021). Lotka's law and author productivity in the Economic literature: a citation study. *Indian Journal of Information Sources and Services,* 11(2), 1-8. DOI: https://doi.org/10.51983/ijiss-2021.11.2.2998.