Computational Regression Analysis on a Dataset of Journals to Evaluate H – Index

[1]B. Annapurna, [2]Dr. M. M. Naidu
[1]HoD, Computer Science Ch.S.D.St. Theresa's College for Women Autonomous *Eluru[1]*
[2]R.V.R & J.C. College of Engineering Guntur[2]
Andhra Pradesh, India.

Abstract - While submitting the research works to the journals for publication, Bibliometrics of journals is considered as the best tool for authors to submit their research papers. There are a variety of indices in use to estimate the quality of the research publications. Every indexing method has its own merits and demerits. The use of indices was initiated by Hirsh with G-index and H-index. This paper proposes a modified k-means algorithm to generate datasets of computer science journals. In this paper, a python based linear regression algorithm was implemented to evaluate H – Index and G – Index.

***Keywords* -** h-index, m-index, A-index, e-index, linear regression

I. INTRODUCTION

The importance of the quantification of the journal citation is increasing progressively. The increasing growth of knowledge and advanced scientific techniques and methods has led to a proliferation of journals in various fields of disciplines. To assess the quality of publications in journals of scientific discipline, several authors proposed and delineated the purpose of bibliometric indices for journals. Bibliometrics of journals has become an important and a promising tool for authors to submit their research papers. Hence, impact factor proposed to assess the quality of journals, however, several controversies existed [1]. The journal impact factor developed by Eugene Garfield and published by the Thomson Reuters is the first biblio-metric evaluator. However, the potentialities and several limitations of the impact factor have been well discussed [2] [3] [4] [5]. Alternative journal rankings [6] have been proposed, however, they deal with a small subset of the literature in any discipline.

The **Institute for Scientific Information** (**ISI**) offered bibliographic database services. A **bibliographic database** is a database of bibliographic records, an organized digital collection of references to published literature, including journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc. In contrast to library catalogue entries, a large proportion of the bibliographic records in bibliographic databases describe articles, conference papers, etc., rather than complete monographs, and they generally contain very rich subject descriptions in the form of keywords, subject classification terms, or abstracts [1].

The ***h*-index** is an author-level metric that attempts to measure both the productivity and citation impact of the publications of a scientist or scholar. The index is based on the set of the scientist's most cited papers and the number of citations that they have received in other publications. The index can also be applied to the productivity and impact of a scholarly journal [6].

II. MATERIALS AND METHODS

2.1 Dataset – A dataset of 150 computer science journals were extracted from SCIMago [http://www.scimagojr.com/journalrank.php] website developed from the information contained in the Scopus database [7]. The parameters calculated by SCImago are Sci Journal Ranking (SJR), h-index, Total Docs, Total References, Total cites, citable docs, cites/doc and references/doc etc. can be used to assess and analyze scientific domains.

2.2 Python - Using Python programming language, a program was written to perform k-means analysis. Python [8] has several advantages like open source, cross platform, object-oriented programming, dynamic typing features, simple and easy to learn and rich set of supporting libraries for mathematics, statistics, and visualization. Python modules like Numpy (Scientific Computing Tools for Python—Numpy), Scipy (Open Source Library of Scientific Tools), Python-Sklearn and matplotlib are used.

Regression analysis - The input data is a csv file with each index value obtained from SCImago [www.scimagojr.com] as descriptors, in a column and h-index of all journals as data in last column. Initially, the python script picks up all independent variables which have good correlation with dependent variable and rank them, and a Pearson correlation matrix is calculated. A predictive mathematical model is built using selected descriptor(s), all statistical terms associated with the model like Multiple Linear Regression r2, Adjusted r2, F statistics are calculated. Index values regarded as dependent variable as they are dependent on the number of papers published, number of citations received for each paper, total citations, total references etc. Citation parameters such as Total Docs. (2013), Total Docs. (3years), Total Refs, Total Cites (3years), Citable Docs. (3years), Cites/Doc. (2years) and Ref./Doc considered as Independent variables as they are not dependent on indices or any values/parameters.

III. RESULTS AND DISCUSSION

The dataset was selected based on the outcome of modified k-means algorithm on a dataset of 150 journals resulted in 3 clusters (k=3) (unpublished). Each cluster data grouped based on the titles appeared in each group and one dataset containing journal subjects such as Artificial intelligence, Programming and Computing was considered to perform linear regression analysis.

Several indicators which assess the scientific merits of researchers reported in literature quantify both the number of published papers and their citations in various other journals. To some extent, some indicators rely on the citation of articles published in journals. Few such important indices are e-index, h-index, A-index and m-index.

The dataset (Table 1) obtained from modified k-means algorithm used as input in regression procedure. The main aim of regression analysis is to assess how to improve index values for a particular journal. Because, a high index value would definitely attract more number of papers from scientific community. Hence, the objective is to identify which citation parameter is important and how such citation parameter behaves in a dataset. The outcome of the program is prediction of dependent variable. Plots constructed between actual and predicted dependent variables and the obtained equation discussed. However, to assess the quality of prediction, several validation parameters are analysed such as r2, F-statistic, Durbin-Watson, Jarque-Bera, Skew, Kurtosis etc. Outliers or outlying data (not shown) detected in the analysis and regression re-run after removing outliers from the dataset, based on relative error calculation.

h-index - The regression analysis on citation parameters of journals listed in dataset-1 carried out against h-index. The data given in Table-2 and the correlation between actual and predicted h-index values presented as graph in Figure-2.

Table 1: List of journals and citation parameters along with predicted h-index values by Eq(3).

S. No.	Journals	h-index	TD3	TC3	CD3	CD2	RD	Predicted e-index	Residual value	Relative error
1	Foundations and Trends in Machine Learning	11	10	281	10	8.14	155.67	41.69	-30.69	**-2.79**
2	Foundations and Trends in Computer Graphics and Vision	14	7	63	7	9.67	0	31.12	-17.12	-1.22
3	IEEE Transactions on Pattern Analysis and Machine Intelligence	221	584	6273	556	9.69	45.69	218.40	2.60	0.01
4	Computer Methods in Applied Mechanics and Engineering	120	764	2614	754	2.93	44.37	80.95	39.05	0.33
5	ACM Computing Surveys	90	84	876	84	9.91	119.15	55.99	34.01	0.38
6	IEEE Transactions on Evolutionary Computation	111	159	1363	153	9.21	52.1	69.07	41.93	0.38
7	SIAM Journal on Computing	68	227	450	221	1.92	32.49	39.58	28.42	0.42
8	Computers and Education	77	757	3724	744	4.52	50.64	119.94	-42.94	-0.56
9	Mathematical Programming	75	285	670	272	2.43	30	43.28	31.72	0.42
10	Proceedings - IEEE Symposium on Security and Privacy	43	117	628	110	4.78	42.95	50.27	-7.27	-0.17
11	ACM Transactions on Mathematical Software	54	83	282	82	4	32.53	40.65	13.35	0.25
12	IEEE Transactions on Mobile Computing	80	414	2034	401	4.75	33.12	81.39	-1.39	-0.02
13	Computers and Geotechnics	48	329	812	316	2.5	33.35	45.41	2.59	0.05
14	Journal of Machine Learning Research	94	756	2379	739	2.51	33.82	73.60	20.40	0.22
15	Foundations of Computational Mathematics	29	78	204	73	2.45	29.91	40.02	-11.02	-0.38
16	Artificial Intelligence	101	203	799	196	3.91	51.47	51.74	49.26	0.49
17	Computational Intelligence and Neuroscience	23	110	419	103	5.16	50.21	42.80	-19.80	-0.86
18	Journal of Computer Assisted Learning	48	138	485	129	2.52	46.23	46.34	1.66	0.03
19	Proceedings of the Annual IEEE Conference on Computational Complexity	21	99	128	93	1.43	23.81	37.00	-16.00	-0.76
20	IEEE Computational Intelligence Magazine	26	117	313	82	3.2	16.72	41.35	-15.35	-0.59
21	INFORMS Journal on Computing	48	146	240	139	1.29	28.03	38.13	9.87	0.21
22	Statistics and Computing	41	181	325	177	1.71	31.19	38.31	2.69	0.07
23	IEEE Transactions on Affective Computing	16	76	533	70	6.79	59.87	47.40	-31.40	**-1.96**
24	Journal of Scientific Computing	42	303	613	259	2.08	34.39	41.41	0.59	0.01
25	International Journal of Machine Learning and Cybernetics	15	67	330	65	4.9	35.3	42.49	-27.49	-1.83
26	IEEE/ASME Transactions on Mechatronics	74	356	1738	350	4.63	22.1	74.40	-0.40	-0.01
27	Automated Software Engineering	29	50	131	40	2.46	39.58	39.48	-10.48	-0.36
28	Computational Geometry: Theory and Applications	35	195	180	141	1.21	18.4	34.16	0.84	0.02
29	Journal of Graph Algorithms and Applications	24	62	76	56	1.35	24.34	37.70	-13.70	-0.57
30	Advanced Engineering Informatics	43	194	638	177	3.58	45.98	47.16	-4.16	-0.10
31	Artificial Intelligence and Law	22	44	65	40	1.41	44.39	38.53	-16.53	-0.75
32	Theory and Practice of Logic Programming	26	116	222	112	2.1	28.91	38.48	-12.48	-0.48
33	Fuzzy Optimization and Decision Making	29	66	150	63	2	21.21	39.26	-10.26	-0.35
34	Computers and Mathematics with Applications	69	2108	5600	2048	2.39	28.54	100.93	-31.93	-0.46
35	International Journal of Artificial Intelligence in Education	10	24	62	23	2.09	53.33	38.96	-28.96	**-2.90**
36	Journal of Artificial Intelligence Research	76	160	495	160	2.23	58.28	45.36	30.64	0.40

37	Mechanism and Machine Theory	60	433	992	421	2.08	25.39	45.33	14.67	0.24
38	Empirical Software Engineering	39	90	303	74	3.76	51.87	41.82	-2.82	-0.07
39	Algorithmica	50	393	433	330	1.21	26.51	30.38	19.62	0.39
40	International Journal of Intelligent Systems	46	176	463	169	2	27.4	43.38	2.62	0.06
41	IEEE Transactions on Parallel and Distributed Systems	78	558	1969	517	3.75	31	71.37	6.63	0.08
42	Control Engineering Practice	67	405	1372	393	3.12	28.64	59.81	7.19	0.11
43	Artificial Intelligence Review	40	161	449	109	2.95	51.38	44.39	-4.39	-0.11
44	Mathematics of Control, Signals, and Systems	26	45	63	39	1.53	31.31	38.22	-12.22	-0.47
45	ACM Transactions on Software Engineering and Methodology	53	51	207	49	4.33	49.19	39.86	13.14	0.25
46	Engineering Applications of Artificial Intelligence	54	430	1336	422	2.95	36.58	56.96	-2.96	-0.05
47	Mathematical and Computer Modelling	59	1171	2517	1114	2.27	24.81	52.21	6.79	0.12
48	Integrated Computer-Aided Engineering	25	89	327	81	4.3	45.57	41.85	-16.85	-0.67
49	Artificial Intelligence in Medicine	53	172	426	158	2.35	40.77	42.23	10.77	0.20
50	Advances in Engineering Software	39	342	760	332	2.23	46.5	42.96	-3.96	-0.10

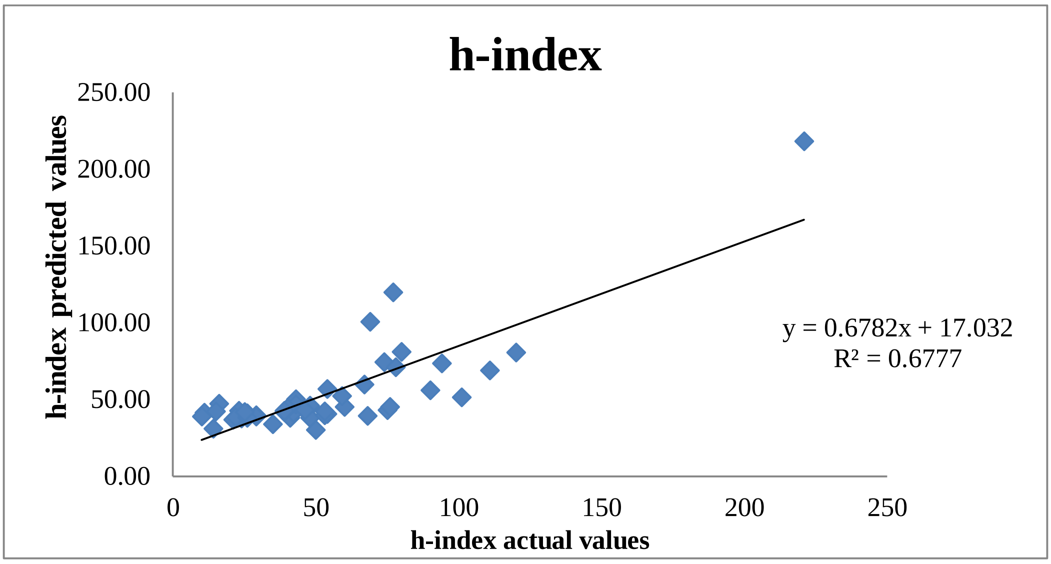

Figure 1: Correlation plot of actual vs predicted h-index values of journals

$$\text{h-index} = -0.0357*TD3 + 0.0362*TC3 - 0.0314*CD3 - 1.1339*CD2 + 0.0074*RD + 40.2711$$
$$r = 0.823,\ r^2 = 0.678,\ \text{adjusted } r^2 = 0.641,\ n = 50 \qquad (1)$$

Outlier Detection - Equation (1) describes the relationship between e-index and citation parameters applied in the study. It is evident from equation that a nominal increase in values of TC3 and RD contributes positively to enhance h-index factor of journals, whereas on the other hand, TD3, CD3 and CD2 has negative effect on h-index. It is observed from above equation that if the total cites to the papers and references per doc in journals are increased to a marginal extent along with decrease in the total docs, number of cites per doc, then an increase in e-index is possible. A negative value which represents reduced behavior of TD3 means that the number of total papers should be low, however dependent on the citations received by the number of papers.

In the next step, outliers removed from the dataset-1 and regression analysis performed. Corresponding data with graphs and tables is given below.

$$\text{h-index} = -0.0235*TD3 + 0.0335*TC3 - 0.0378*CD3 - 0.7615*CD2 + 0.3842*RD + 27.7776$$
$$r = 0.856,\ r^2 = 0.732,\ \text{adjusted } r^2 = 0.699,\ n = 47 \qquad (2)$$

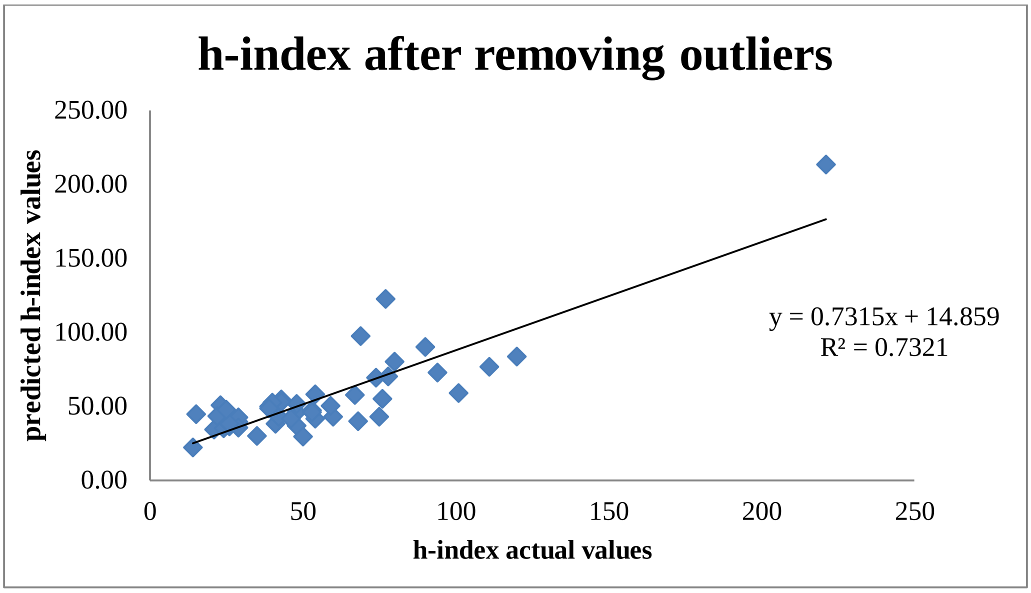

Figure 2: Correlation plot of actual vs predicted h-index values of journals after removing three outliers.

IV. CONCLUSION

For the selected dataset of journals, considering e-index, it was observed that a nominal increase in values of TC3, CD3, CD2 and RD contributes positively to enhance e-index factor of journals and TD3 has no effect on e-index. For h-index, after removing outliers, an increase in value of TC3 and RD contributes positively whereas TD3, CD3 and CD2 has negative effect on h-index. A negative value which represents reduced behavior of TD3 means that the number of total papers should be low, however, dependent on the citations received by the number of papers. Three outliers detected in m-index, and the outcome of regression analysis is similar to h-index where an increase in values of TC3 and RD contributes positively and the remaining citation parameters have negative effect on m-index. Therefore, considering the suggestions made in this work, publishers can benefit from increase in number of paper submissions to their journal, in particular.

V. REFERENCES

[1]. Feather, John; Sturges, Paul, eds. (2003). International Encyclopedia of Information and Library Science (Second ed.). London: Routledge. p. 127. ISBN 0-415-25901-0.

[2]. Garfield, E.1996. How can impact factors be improved British Medical Journal, 313, 411–413.

[3]. Glänzel, W., & Moed, H. F. 2002. Journal impact measures in bibliometric research. Scientometrics, 53(2), 171–194.

[4]. Saha, S., Saint, S. & Christakis, D.A. 2003. Impact factor: a valid measure of journal quality? Journal of the Medical Library Association 91:42-46.

[5]. Dong, P., Loh, M. & Mondry, A. 2005. The "impact factor" revisited. Biomedical Digital Libraries 2:7 doi:10.1186/1742-5581-2-7.

[6]. Suzuki, Helder (2012). "Google Scholar Metrics for Publications" googlescholar.blogspot.com.br.

[7]. http://www.elsevier.com/

[8]. http://www.python.org