Determinants of Research Citation Impact: A Combined Statistical Modelling

Fereshteh Didegah

Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY UK

E-mail: f.didegah@wlv.ac.uk

Abstract: This study investigates a range of metrics available when an article is published to see which metrics most associate with its eventual citation count. The purposes are to contribute to developing citation model and to inform policy makers about which predictor variables associate with citations in different fields of science. Despite the complex nature of reasons for citation, some attributes of a paper's authors, journals, references, abstract, field, country and institutional affiliations, and funding source are known to associate with its citation impact. The thesis investigates some common factors previously assessed and some new factors, including 8 main variables (internationality, interdisciplinarity, impact, size, collaboration, social network, readability, and funding) and 24 sub-variables (journal author internationality, journal citing author internationality, cited author internationality, cited journal author internationality, cited journal citing author internationality, reference interdisciplinarity, impact of author(s), publishing journal, references, institution of affiliation, and country of affiliation, paper, abstract, and title lengths, number of keywords and references, size of the field, number of authors, institutions, and countries, social institutional and co-authorship networks, readability of abstract and research funding). The internationality and social network variables are the new factors introduced in this study and the Gini coefficient is used to measure internationality. The h-index is used to gauge author impact and the Median Normalized Citation Score (MNCS) is used to gauge institutional and country impact. The Flesch Reading Ease Score is also used to measure readability of abstract. A sample of articles and proceedings papers in the 22 Essential Science Indicators subject fields from the Web of Science constitute the research data set. Using Negative Binomial Hurdle Models, an appropriate statistical model to simultaneously assess the citation factors, this study assesses the above factors using large scale data. Preliminary findings show that internationality and impact factors are the most effective determinants of citation counts in most subject fields.

Introduction and research background

This study investigates properties of an article as a text document when it is published to find the determinants that associate with the number of citations to the article. Knowledge of these factors could be useful to science evaluators to help them to make early estimates of the number of citations that a set of published articles is likely to receive.

Some factors result from authors' intellectual perceptions of an article and these reasons have been explored through questionnaires or interviews. Context or content analyses employing text analysis and semantic content analysis methods are two other approaches to explore citers' motives. Owing to the time-consuming nature of qualitative research and the complex and discipline-dependent nature of citers' motives, such qualitative studies usually involve only a small sample of scholars and documents.

A number of empirical studies have been carried out to seek associations between citation counts and various objective and easily measurable properties of research. These include the impact of the publishing journal (Boyack & Klavans, 2005), collaboration (Gazni & Didegah, 2010), the interdisciplinarity of the article references (Larivière & Gingras, 2010), the number and impact of references (Boyack & Klavans, 2005), and the size of the related field (Lovaglia, 1989). These factors may not directly determine future citation counts, but can provide indirect evidence of likely future citation impact. Although a number of studies have investigated citation factors in some subject areas, many areas and some factors have not yet been examined.

This study examines whether research collaboration, journal, reference, author, country of affiliation and institution of affiliation impacts, journal and reference internationality, reference interdisciplinarity, social networks, research funding, abstract readability, and article and field size attributes affect citation counts. Some factors such as research collaboration, journal and reference impact, research funding, abstract readability, and article size attributes are to some extent under the control of the authors and so it would be useful to know whether researchers should pay attention to them to ensure that their research has the greatest possible impact. Research collaboration has been frequently analysed (Sooryamoorthy, 2009) and some other factors have also been examined (Zhao, 2010) but they have not been examined simultaneously for multiple research fields using an optimal statistical model. In addition, this study assesses two new determinants of the citation impact of papers: journal and reference internationality and social networks.

van Raan (1998) criticises the claim that a theory is needed for citation analysis and suggests replacing the theory with a feasible model that provides a possible approximation of reality. This study also helps to address this goal with its new, more integrated statistical model.

Significance of the research

The development of a model for citation behaviour is the main motivation for conducting this study. Article citation impact factors have been widely scrutinized in the previous literature but have been considered separately (and mostly within a single field) whereas, in reality, citation impact results from interactions between different factors. This is an important omission because inappropriate models may generate misleading conclusions and non-simultaneous tests may identify apparently important factors that are not relevant when other factors are also considered. A simultaneous assessment of these factors will fill this gap in the literature and represent a model closer to reality. Given that this research is conducting a comparison across all fields of science, the results would be informative and significant for scholars in all fields of science, scientific policy makers and research administrators and will help developing a citation model.

Methodology

Data collection

A sample of publications from 22 different subject categories covered by Thomson Reuters' Web of Science (WoS) from 2000-2009 were extracted by systematic sampling based upon the year of publication and the sub-fields. Using the list of journals provided by *ScienceWatch.com* classifying each journal into one of the 22 fields, a single subject field was assigned to each document in the dataset. The subject categories are the 22 fields of Essential

Science Indicators (ESI) Subject Classification. Although the subject classification is journal-based, it is well-established and has frequently been used by scientometricians to classify individual papers. Only two types of documents, articles and conference proceedings, were included because original research is mainly published in these two types of documents (Milojević & Leydesdorff, 2012).

Variables and measurements

The dependent variable is the citation count of papers and the independent variables and their measures are presented in Table 2. Eight main variables and 24 sub-variables are examined in this study. The statistical model is fitted for the eight main variables separately from the 24 sub-variables in each subject area.

More specifically, the new internationality factor is gauged through five different approaches. Two approaches are related to the publishing journal and three approaches are related to the paper's references (See Figure 1). Previous studies have applied two methods to measure internationality: relative and absolute methods. Absolute methods implement concentration indices such as the Gini coefficient. Relative methods use normalization techniques. To measure the internationality variables in this study, the Gini coefficient was selected as the most straightforward approach automatically calculated for each journal in the dataset. The Gini coefficient is a value between 0 and 1 where the value of 0 represents absolute concentration (e.g., all authors from a single country) and the value of 1 represents absolute dispersion (Buela-Casal, Perakakis, Taylor & Checa, 2006). Hence, journals with Gini coefficients closer to 1 are more international. The Gini formula is:

$$\mathrm{Gini} = \left| 1 - \sum_{i=1}^{N} (X_{i-1} - X_i)(Y_{i-1} + Y_i) \right|$$

Where:
N = Number of countries contributing to the journal;
X_i = Cumulative proportion of countries for ith country ($X'_i = i/N$);

Y_i = Cumulative proportion of authors publishing in or citing the journal from countries 1 to i, where the countries are arranged in descending order of the number of authors contributing to the journal.

There are numerous formulae to measure the readability of a text but their validity is still a matter of debate. To prevent readability formula limitations affecting the results of our study, seven different readability formulae were used: Kincaid formula, Automated Readability Index (ARI), Coleman-Liau formula, Flesch Reading Ease formula, Fog Index, Lix formula, and SMOG Grading. The STYLE program was used to automatically calculate these scores (Cherry & Vesterman, 1981). There was a significant correlation between the seven readability scores in all fields. The Flesch Reading Ease Score was used since it seems to be the most popular and also has a high correlation with the other six scores (r ~0.8). The Flesch Score ranges between 0 and 100 where 0 indicates a text that is the most difficult to read and 100 represents the easiest text to read.

Social network is another factor that is measured through two approaches: co-authorship network and institutional network. Based upon the co-authorship approach, scholars' co-authors will presumably be the scholars' citers in future. Due to author names' variations and since it needs to control over the time to measure the effect of previous collaboration links on the future citation links, this factor will be examined on a small sample of authors in a single field and will be presented as a further study to the research.

The assumption carried by the institutional network is that scholars affiliated by an institution may receive citations from their colleagues who are working in the same institution and have the same research interests. This factor will be also examined on the same sample of authors used for measuring the co-authorship network.

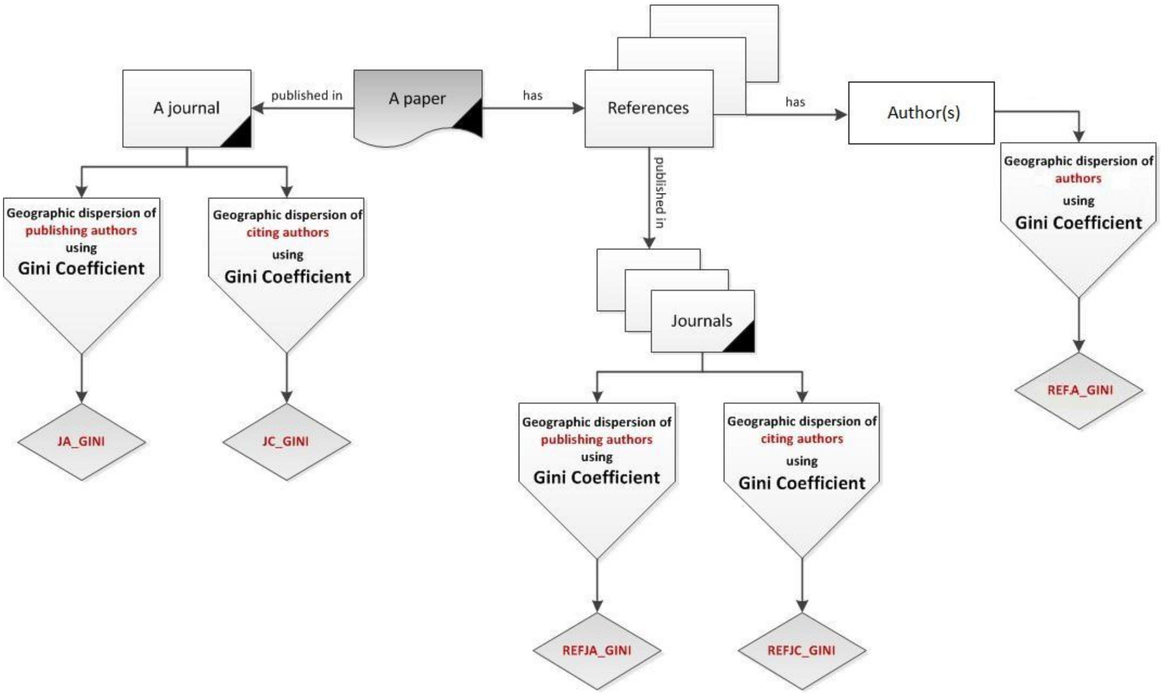
A journal
published in
A paper
has
References
has
Author(s)
published in
Journals
Geographic dispersion of publishing authors using Gini Coefficient
Geographic dispersion of citing authors using Gini Coefficient
Geographic dispersion of authors using Gini Coefficient
JA_GINI
JC_GINI
REFJA_GINI
REFJC_GINI
REFA_GINI

Figure 1. The calculation process for journal and reference internationality.

Table 1. The 22 ESI subject categories

Subject categories	**Sample size (no. of documents)**
Agricultural Sciences	15,488
Biology & Biochemistry	15,689
Chemistry	16,342
Clinical Medicine	16,387
Computer Science	15,698
Economics & Business	14,987
Engineering	16,059
Environment/Ecology	15,542
Geosciences	15,507
Immunology	15,104
Materials Science	15,907
Mathematics	15,570
Microbiology	15,254
Molecular Biology & Genetics	16,092
Multidisciplinary	14,248
Neuroscience & Behaviour	15,845
Pharmacology & Toxicology	14,411
Physics	16,203
Plant & Animal Science	15,867
Psychiatry/Psychology	14,636
Social Sciences, general	16,096
Space Science	14,614

Table 2. Independent variables and measures

Main factor	Sub-factors	Measure
Internationality of the paper	Journal author internationality (JAI)	Gini Coefficient
	Journal citing author internationality (JCI)	Gini Coefficient
	Cited author internationality (REFAI)	Gini Coefficient
	Cited journal author internationality (REFJAI)	Gini Coefficient
	Cited journal citing author internationality (REFJCI)	Gini Coefficient
Interdisciplinarity of the paper	Reference interdisciplinarity	Gini Coefficient
Impact of the paper	Author(s) impact	Maximum H-index of the publishing authors
	Journal Impact Factor	Journal Impact Factor
	Reference impact	Reference Median Citation Score
	Institution impact	Mean Normalized Citation Score (MNCS)
	Country impact	Mean Normalized Citation Score (MNCS)
Size of the paper	Length of paper	Number of pages
	Length of abstract	Number of words
	Length of title	Number of words
	Number of keywords	Number of keywords
	Number of references	Number of references
	Size of field	Number of publications in the related field
Research collaboration	Individual collaboration	Number of authors
	Institutional collaboration	Number of institutions
	International collaboration	Number of countries
Social networks	Co-authorship network	Matching author citers with the author co-authors and investigating the effect of co-authorship on future citations.
	Institutional network	The percentage of author citers with the same institution affiliation as the author
Readability of the paper	Readability of abstract	Flesch Reading Ease Score
Research funding		Funded (1); Unfunded (0)

Statistical procedures

Since the dependent variable is count data, count data regression models are most appropriate. The basic models for count data are the Poisson and Negative Binomial (NB) distributions. Because of data overdispersion, the Poisson model in which the mean and the variance are assumed to be equal does not fit and the NB model is more appropriate. The data had an unusual amount of zeros (i.e., uncited articles) for the NB distribution, however. One approach to deal with the common issue of too many zeros in count data is the hurdle model. The assumption behind the hurdle model is that zero counts and non-zero counts are generated by different underlying processes and should be modelled separately. With this model, after passing a hurdle in order to gain positive counts, the positive counts follow a Poisson or NB distribution.

This study models both zero citations and non-zero citations using the hurdle model. The number of citations to a paper has been previously shown to take a Poisson or negative binomial distribution after passing the zero barrier or the hurdle (Burrell, 2003). A hurdle model comprises of two parts: a count model and a binary model and has different types: NB-logit model, NB-cloglog (complementary log-log) model, Poisson-logit model and Poisson-cloglog model. For the count model, the NB model was the best fit to the data due to the data overdispersion. For the binary model, a logit model was the best fit and odds ratio in this model is in the form of Log [P(citations>1)/P(citations=0)] (Hilbe, 2011).

To examine the hurdle model for the entire ten years, publication year has been included in the model as a logarithmically transformed year of publication.

An example of obtained results

The results of the negative binomial-logit hurdle model provide coefficients for both the negative binomial (non-zero citation counts) and the logit (proportion of uncited papers) components of the model for a number of factors in Clinical Medicine (Table 3).

Journal impact and internationality

With respect to the negative binomial model, the journal internationality (JAI) and the journal impact factor (JIF) significantly associate with increased citations and a unit increase in the JAI and JIF increases the mean citation counts by a substantial 43.6% and 22.6%, respectively. With respect to the logit model, a unit change in the JAI and the JIF significantly contributes to 79.2% and 57.9% decreases in the mean zero citations, respectively.

Research collaboration

The coefficients of the negative binomial model show that among the patterns of research collaboration, international and individual collaborations significantly associate with increased citations whereas institutional collaboration is not a significant determinant of citation counts. One additional country increases the mean citation count by 17.1% and one additional author increases the mean by 1.8%. With respect to the logit model, a unit change in the number of countries and the number of authors decreases the mean zero citations by 43.6% and 5.7%, respectively.

Article properties impact

The impacts of author, institution and country of affiliation were examined. With respect to the negative binomial model, the institution impact is the only significant determinant of citation counts and a unit increase in this factor associates with 1.7% increase in the mean citation count. With respect to the logit model, this factor associates with 18.1% decrease in the mean zero citations.

References characteristics

The three article reference features, internationality, impact and number, associate with increased citation counts. A unit increase in the internationality and impact of an article’s references associates with 89.1% and 0.6% increases in the mean citations, respectively. Each additional reference also associates with 0.9% increase in the mean citations. With respect to the logit model, the three reference factors significantly associate with decreased zero citations.

Abstract readability and length

Abstract readability is a significant determinant of decreased citation counts whereas abstract length is a significant factor of increased citations. A unit increase in the readability score decreases the mean citation count to 99.8% which has no practical significance. With respect to the logit model, abstract readability is not a significant determinant of zero citations. A unit increase in the abstract length increases the mean citation count to 0.2% and with respect to the logit model, this factor significantly associates with 0.4% decrease in zero citations.

Table 3. The results of hurdle model in Clinical Medicine

Logit Model	**Coef.**	**Exp.(Coef.)**	**Std. Err.**	**z**	**P>z**	**[95% Conf. Interval]**	
Journal IF	0.457	1.579	0.019	24.21	0.000	0.42	0.494
No. of authors	0.056	1.057	0.011	4.88	0.000	0.033	0.078
No. of institutions	-0.073	0.929	0.023	-3.14	0.002	-0.119	-0.027
No. of countries	0.362	1.436	0.071	5.1	0.000	0.223	0.501
Author impact	0.000	1	0.000	4	0.000	0.000	0.000
Institution impact	0.166	1.181	0.018	9.33	0.000	0.131	0.201
Country impact	-19.224	0.000	8.797	-2.19	0.029	-36.466	-1.981
Abstract readability	-0.003	0.997	0.002	-1.41	0.157	-0.007	0.001
Abstract length	0.004	1.004	0.000	10.12	0.000	0.003	0.004
No. of references	0.018	1.018	0.002	7.92	0.000	0.014	0.023
JAI	0.583	1.792	0.237	2.47	0.014	0.12	1.047
REFJCI	7.815	2476.786	0.684	11.43	0.000	6.475	9.155
Reference impact	0.002	1.002	0.001	3.04	0.002	0.001	0.003
Constant	-6.843	0.001	0.535	-12.79	0.000	-7.892	-5.794
NB Model	**Coef.**	**Exp.(Coef.)**	**Std. Err.**	**z**	**P>z**	**[95% Conf. Interval]**	
Journal IF	0.204	1.226	0.005	39.63	0.000	0.193	0.214
No. of authors	0.018	1.018	0.004	4.51	0.000	0.01	0.026
No. of institutions	0.01	1.01	0.008	1.15	0.249	-0.007	0.026
No. of countries	0.158	1.171	0.02	7.8	0.000	0.118	0.198
Author impact	0.000	1	0.000	1.64	0.101	0.000	0.000
Institution impact	0.017	1.017	0.003	5.6	0.000	0.011	0.023
Country impact	-1.74	0.176	4.184	-0.42	0.678	-9.94	6.461
Abstract readability	-0.002	0.998	0.001	-2.22	0.027	-0.004	0.000
Abstract length	0.002	1.002	0.000	15	0.000	0.002	0.003
No. of references	0.009	1.009	0.001	10.7	0.000	0.008	0.011
JAI	2.006	7.436	0.113	17.79	0.000	1.785	2.227
REFJCI	3.254	25.891	0.351	9.26	0.000	2.565	3.942
Reference impact	0.006	1.006	0.000	24.18	0.000	0.006	0.007
Constant	-3.414	0.033	0.271	-12.61	0.000	-3.945	-2.883
Alpha	**0.472**	**1.603**	**0.023**	**20.45**	**0.000**	**0.427**	**0.517**

References

Boyack, K.W. & Klavans, R. (2005). Predicting the Importance of Current Papers. *Proceedings of ISSI 2005,* 335–342. Edited by P. Ingwersen and B. Larsen. July 24-28, Stockholm, Sweden.

Buela-Casal, G., Perakakis, P., Taylor, M., & Checa, P. (2006). Measuring internationality: reflections and perspectives on academic journals. *Scientometrics*, 67 (1), 45-65.

Burrell, Q.L. (2003). Predicting Future Citation Behaviour. *Journal of the American Society for Information Science and Technology,* 54(5), 372-378.

Cherry, L. L. & Vesterman, W. (1981). *Writing tools: the STYLE and DICTION programs*. Bill Cronin Laboratories, Murray Hill, New Jersey.

Gazni, A. & Didegah, F. (2010). Investigating Different Types of Research Collaboration and Citation Impact: A Case Study of Harvard University's Publications. *Scientometrics*, 87(2), 251-265.

Hilbe, J.M. (2011). *Negative Binomial Regression, second edition* (5th print, 2012). Cambridge, UK: Cambridge University Press.

Larivière, V. & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61, 126-131.

Lovaglia, M.J. (1989). Status characteristics of journal articles for editor's decisions and citations. *The Society for Social Studies of Science Annual Meeting*, November 15- 18, University of California at Irvine, Irvine, CA.

Milojević, S. & Leydesdorff, L. (2012). Information Metrics (iMetrics): A Research Specialty with a Socio-Cognitive Identity? *Scientometrics*.

van Raan, A. F. J. (1998). In matters of quantitative studies of science: The fault of theorists is offering too little and asking too much. *Scientometrics*, 43(1), 129-139.

Zhao, D. Z. (2010). Characteristics and impact of grant-funded research: a case study of the library and information science field. *Scientometrics,* 84(2), 293-306.

Schedule of completion

Table 4. Research schedule of completion

	Year	**2013**						**2014**	
	Month	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**
Tasks	Running statistical tests and models	■	■						
	Reporting results and producing tables and graphs			■					
	Doing the last research paper out of the thesis				■				
	Writing up the thesis chapters					■	■	■	■
	Thesis submission								■

Budget

I have a plan to take part in the ASIS&T annual meeting (November 1-6, Montreal, Canada) and shall submit a paper to the workshop on Informetric and Scientometric research as part of the ASIS&T annual meeting. Hence, this scholarship will financially support me to travel to Canada to attend the meeting and the workshop.

Thesis advisor: Prof. Mike Thelwall