Determinants of the Impact Factor of Publications: A Panel Model for Journals Indexed in Scopus 2017

Jenny Paola Lis-Gutiérrez[1]([⊠]), Carolina Henao[2], Álvaro Zerda[3],
Mercedes Gaitán[1], Juan Carlos Correa[1], and Amelec Viloria[4]

[1] Fundación Universitaria Konrad Lorenz, Bogotá, Colombia
{jenny.lis,mercedes.gaitana,
juanc.correan}@konradlorenz.edu.co
[2] Corporación Universitaria Minuto de Dios, Bogotá, Colombia
linda.henao@uniminuto.edu
[3] Universidad Nacional de Colombia, Bogotá, Colombia
azerdas@unal.edu.co
[4] Universidad de la Costa, Barranquilla, Colombia
aviloria7@cuc.edu.co

Abstract. This article has the purpose of establishing which are the variables that allow explaining the behavior of the SJR between 2014 and 2016, for the journals indexed in Scopus. To do this, journals that had a SJR value greater than eight in 2016 were selected, that is, 103 of the 22,231. For the analysis, a model of standard errors corrected for panel was used, for which a coefficient of determination of 81% was obtained, and a model of feasible generalized least squares. From these it was possible to establish that variables such as open access, the number of areas in which the publication is registered and the language of publication, are not significant to explain the impact of a publication. On the contrary, variables such as belonging to health sciences or social sciences.

Keywords: SJR · Open access · Publications · Journals · Indexing Scopus

1 Introduction

When talking about the impact of publications, it is possible to identify two approaches. The first is associated with the analysis of impact indicators of journals, and the second on the importance given by institutions and researchers to impact measurements. Within the first approach, the debate between open access (OA) [1, 2] and restricted [3] regarding scientific communication goes back to the 1990s, Although some open access journals exist since the mid-twentieth century (The Faculty Advisory Council, 2012) [4]. On the one hand, it is discussed if the OA could lead to the loss of quality of publications [5, 6], and on the other it is argued that the traditional system of scientific communication is in crisis [1] since it limits the dissemination and exchange of knowledge.

Also, empirical works such as [7] are presented, in which it was established that the publications of the first quarter had more citations than the papers published in the last quarter. For their part, [8], applied a quantile regression to predict a probability distribution that establishes the future number of citations of a publication.

Within the second approach are the works of [9], who from a survey applied to researchers from the United Kingdom and the development of two logit models, concluded that the institutional factors, the intensive research contexts and non-academic work experience, explain the preference of researchers for high-impact publications. [10] found that researchers do not interpret to the same extent the number of citations and the perceived impact, and that in contexts of expertise in a specific area, there is a bias to choose their own works.

In this context, [11] showed that self-citations of journals greatly affect the impact factors of journals at the meso and micro levels, since it is a way to artificially increase evaluation indicators impact of these, which severely undermines the objective authenticity of these indicators.

[12] argued that academic journals favor authors who share their institutional affiliation, came to this conclusion after examining citation counts, as a substitute for the quality of the article, for articles published in four leading journals in international relations during the years 2000–2015 and concluded that there is evidence of the existence of bias in some journals towards the publication of works by the teachers of their institution of origin, at the expense of the quality of the article.

[13] discern about five types of citations: application, affirmation, denial, revision and superficial mention, and their results encourage scientific stakeholders to go beyond the quotation count to evaluate the scientific contribution of an article or an academic, all of the above, from a panel data model with fixed effects.

This paper aims to establish, what are the variables that explain the behavior of the SJR between 2014 and 2016, for the journals indexed in Scopus? To do this, a panel data model will be used whose dependent variable is the SJR and the explanatory variables: (i) years of coverage of the journal, (ii) the areas in which the journal publishes, (iii) dichotomous variables where it is specified if it is open access, if it includes press articles, and the study area of the journal (life sciences, social sciences, physical sciences or health sciences), (iv) economic control variables to measure the relevance of the country in research took the per capita GDP, the number of patents and the percentage of GDP dedicated to research and development in the country of origin of the journal, the ratio of the number of documents cited and documents published from the country of origin of the journal, as a proxy of the degree of relevance in research of the country of origin of the journal.

2 Methodology

2.1 Data

The initial data were obtained from the page https://www.scopus.com/sources.uri?DGCID=Scopus_blog_post_check2015 and the list of Scopus sources "Scopus Source List" was downloaded as of October 2017. The treatment to perform first, the filter was

made between active and inactive publications and the active publications were chosen. Secondly, a filter was made to choose the types of sources, within the options were "Journal" "Trade Journal" and "Book Series", for this study "Journal" and "Trade Journal" were chosen. The data of the economic variables used were extracted from World Bank statistics.

2.2 Model

An econometric exercise was carried out to analyze the determinants of the journal impact indicator. To avoid this bias due to heterogeneity, the panel data econometric technique was used since it reduces the problems related to the identification of the models. The individuals analyzed (103 in total) were journals with a SJR value greater than eight in 2016. The period studied is comprised between 2014–2016.

The dependent variable used was the SJR and the explanatory variables were:

- Years of coverage of the journal is expected a positive relationship with the SJR, since to the extent that the journal has more years of coverage, there is a higher probability of citation.
- The number of areas published in the magazine. We seek to corroborate if there is empirical evidence that more areas of study, more citation, since it would be expected that as there is a broader scope of study increase the SJR, however, encompassing more areas of study may hinder the search for articles for researchers and in such a case we would expect a negative relationship with the dependent variable.
- A dichotomous variable that specifies if it is open access, in general it is expected that open access journals are more cited; However, [14] found no significant differences in mean values or growth rates between the Gold Open Access and Non Gold Open Access journals, taking bibliometric and bibliographic data from 27,141 journals (indexed between 2001 and 2015 in the SCImago Journal & Country Rank (SJR)). [15] conducted a study where they documented the growth in the number of journals and articles together with the increase in the normalized citation rates of the open access journals listed in the Scopus bibliographic database between 1999 and 2010, concluded that journals/open access articles have grown much faster than subscription journals, but still represented a very low percentage in journals in Scopus. In addition, the averages of two-year citations for journals where the processing of articles was financed by APC have reached the same level as the subscription journals. Average citations for open access journals funded by other media continue to lag behind open access journals funded by APC and subscription journals.

– A dichotomous variable where it is specified if it includes press articles. A positive relationship with the SJR is expected, since the inclusion of press articles attracts more readers
– Dichotomous variables where specified, and the area of study of the journal (life sciences, social sciences, physical sciences or health sciences). It is expected that the citation grade will be affected by the study science. [16] presented the characteristics of citations in the text in more than five million full-text articles from two databases (PMC, Open Access subset and Elsevier journals) and found the fields of biomedical and health sciences; life and earth sciences; and physics, science and engineering have similar reference distributions, although they vary in their specific details and the fields of mathematics and computer science; and social sciences and humanities, they have different reference distributions of the other three. In this context [14] conclude that the average values of reference density in some categories of Social Sciences and Arts and Humanities were equal or superior to those of the "hard sciences." Since citations to references occur at least as often in these two areas of knowledge as they do in some of the hard sciences, despite the potentially minor impact of the journal and, in general, did not correspond to the growth rates in the number of documents produced.
– Economic control variables to measure the degree of development and the resources of the country of origin destined for research, the per capita GDP was taken, and the percentage of GDP dedicated to research and development of the country of origin of the journal. It is expected that the citation has a positive relationship with the quality and impact of the investigation of its environment. However, [17], taking more than 600,000 publications from ISI Web of Science (WOS), processed data on the total number of citations received in ten years for all ISI-indexed journals, and found that the distribution of scientific citations for selected publications with different rules (author, subject, institution, country, magazine, etc.) collapses in a single curve if the citations are drawn relative to their average value; therefore, they affirm that the number of citations of a publication is basically a measure of social popularity, while it is considered to reflect the quality and impact of the research. If so, the proposed variables would not be significant.
– The reason for the number of documents cited and documents published from the country of origin of the journal and the number of patents, as proxies of incidence in the global research of the country of origin of the journal. It is expected to present a positive relationship with the SJR. Since [18] stated that the number of citations is used to measure the impact of academic research or the quality of an academic department and showed that there are other factors, other than the journal that are important including the extension of the article, the number of references and the status of the institution of the first author. Therefore, it is expected that the relevance in research of the country of origin of the journal is greater, either due to a higher quality of the educational institutions.

The model is specified as follows:

$$\begin{aligned}
\text{logsjr}_{jt} = {} & \beta_0 + \beta_1 \log\text{cobertura}_{jt} + \beta_2 \log\text{areas}_{jt} + \beta_3 \log\text{pibs}_{it} + \beta_4 \text{open}_{jt} \\
& + \beta_5 \text{articulosprensa}_{jt} + \beta_6 \log\text{imp1}_{it} + \beta_7 \log\text{gasto}_{it} \\
& + \beta_8 \log\text{paten}_{it} + \beta_9 \text{lifesciences}_{jt} + \beta_{10} \text{socialsciences}_{jt} \\
& + \beta_{11} \text{Physicalsciences}_{jt} + \beta_{12} \text{Health sciences}_{jt} + \varepsilon_{jt}
\end{aligned} \tag{1}$$

j corresponds to the journal, t is the year, i is the country of origin of the journal, logsjr is the logarithm of the sjr, logsjr is the logarithm of the journal's coverage years; logarithms is the logarithm of the number of areas in which the journal publishes, logpib is the logarithm of per capita gdp, open is a dichotomous variable where it is specified if it is open access, newspaper articles is a dichotomous variable where it is specified if it includes press articles, logimp1 is the logarithm of the ratio of the number of documents cited and documents published from the country of origin of the journal, loggasto is the logarithm of the percentage of GDP dedicated to research and development in the country of origin of the journal, logpaten is the logarithm of the number of patents of the country of origin of the journal, lifesciences is a dichotomous variable where it is specified if the study area of the journal is life sciences, socialsciences is a dichotomous variable where it is specified if the area of study is the journal is social sciences, Physicalsciences is a dichotomous variable where it is specified if the study area of the journal is physical sciences as, Health sciences is a dichotomous variable that specifies if the study area of the journal is Health Sciences, and ε_{jt} is a random disturbance that is supposed $\varepsilon_{jt} \sim N(0, \sigma^2)$.

3   Result of Model

In order to know if the random effects model or the pooled data model should be used, the Lagrange Multiplier for Random Effects test was performed, since the null hypothesis is rejected, it is preferable to use the random effects estimate instead of the pooled one (see Algorithm 1).

The Hausman test finds that the null hypothesis can not be rejected because the difference between the random and fixed effect coefficients is not systematic. Therefore, the random effects method was used (see Algorithm 2).

Algorithm 1. Breusch and Pagan Lagrangian multiplier test for random effects

logsjr[revista,t] = Xb + u[revista] + e[revista,t]
Estimated results:
 | Var sd = sqrt(Var)
---------+----------------------------
logsjr | .1581571 .3976897
 e | .0234742.153213
 u | .1211393 .3480507

Test: Var(u) = 0
 chibar2(01) = 176.53
Prob > chibar2 = 0.0000

Algorithm 2. The Hausman test

Ho: difference in coefficients not systematic
chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
=4.46
Prob>chi2 =0.4852

When performing the Wooldridge test for autocorrelation, the null hypothesis is rejected at a level of significance of 5% (see Algorithm 3).

Algorithm 3. Wooldridge test for autocorrelation in panel data

H_0: no first-order autocorrelation
 F(1, 82) = 3.995
 Prob > F = 0.0489

The likelihood ratio test of the null hypothesis homoscedasticity is rejected at 1% level of significance (see Algorithm 4).

Algorithm 4. Proof of the likelihood ratio.

lrtest hetero . , df(`df')
Likelihood-ratio test LR chi2(88) = 264.16
(Assumption: . nested in hetero) Prob > chi2 = 0.0000

When performing the Pesaran test of cross section dependence for unbalanced panel data set, the null hypothesis can not be rejected, therefore the errors of the cross section are weakly dependent [10].

The problems of contemporary correlation, heteroskedasticity and autocorrelation detected can be solved together with feasible generalized least squares estimators or with standard errors corrected for panel, therefore the two estimations are made (see Table 1).

Table 1. Feasible generalized least squares estimators or with standard errors corrected for panel

Prais-Winsten regression, heteroskedastic panels corrected standard errors		Cross-sectional time-series FGLS regression	
logsjr		logsjr	
logcobertura	0.016	logcobertura	0.019
	(1.48)		(3.37)**
logareas	−0.177	logareas	−0.160
	(2.07)*		(4.26)**
logpib	0.100	logpib	0.087
	(1.41)		(1.77)
open	−0.304	open	−0.122
	(1.26)		(0.28)
articilosprensa	0.063	articilosprensa	0.094
	(1.05)		(3.14)**
logimp1	0.988	logimp1	1.233
	(0.95)		(2.00)*
loggasto	−0.071	loggasto	−0.025
	(0.52)		(0.25)
logpaten	0.006	logpaten	−0.003
	(0.29)		(0.20)
lifesciences	0.020	lifesciences	0.091
	(0.19)		(1.80)
socialsciences	0.059	socialsciences	0.126
	(0.51)		(2.34)*
physicalsciences	0.023	physicalsciences	0.006
	(0.20)		(0.10)
healthsciences	0.294	healthsciences	0.303
	(2.88)**		(5.72)**
_cons	1.461	_cons	1.587
	(1.92)		(3.06)**
R2	0.81		
N	260	N	259
* p < 0.05; ** p < 0.01		* p < 0.05; ** p < 0.01	

4 Discussion and Conclusions

The model of standard errors corrected for panel has a coefficient of determination of 81%, and the significant variables in the model at 5% of the level of significance are the number of areas in which the journal is published and the dichotomous variable associated with the area of the study of health sciences, therefore it can be concluded that there is empirical evidence that the journals that publish in health sciences are more

cited and show a negative relationship between the SJR and the number of areas published by the journal, therefore, it can not be affirmed that the more the journal publishes in more areas, the more it will be cited.

Regarding the model of feasible generalized least squares, the significant variables in the model at 5% of the level of significance are the years of coverage of the journal, the number of areas in which the journal is published, the press articles included, the reason of the number of documents cited and documents published from the country of origin of the journal and the dichotomous variables associated with the area of study of social and health sciences, therefore, it can be concluded that there is empirical evidence that the journals that have more years of coverage, which includes press articles, which originate from countries with greater relevance in research, that publish in social and health sciences have a greater impact indicator. And it is corroborated that there is a negative relationship between the number of areas in which the journal publishes and the SJR.

The economic control variables to measure the degree of development and the resources of the country of origin of the journal destined for research are not significant in either of the two models, in line with what was found by [17]; since in the model of generalized least squares feasible to be significant the variable that is associated with the magazines that include press articles; It can be inferred that the citation is associated with social popularity, rather than with the quality and impact of the research.

The two specified models show that open access is not significant. Therefore, it can not be affirmed that an open access journal presents a greater impact indicator corroborating the conclusions of [14].

For future research it would be worthwhile to carry out this same analysis with the journals indexed in WoS; use other indicators as an independent variable (for example, CiteScore, SNIP, among others); perform comparative analyzes between different areas of knowledge, in order to identify if the findings are maintained.

References

1. Sánchez Tarragó, N.: El movimiento de acceso abierto a la información y las políticas nacionales e institucionales de autoarchivo. Acimed. **16**(3) (2007)
2. Peters, M.: Open science, philosophy and peer review. Educ. Philos. Theory (2013). https://doi.org/10.1080/00131857.2013.781296
3. Swan, A.: The Open Access citation advantage: studies and results to date. Technical report, School of Electronics and Computer Science, University of Southampton (2010)
4. The Faculty Advisory Council Memorandum on Journal Pricing (2012)
5. Gaitán-Angulo, M., Viloria, A.: Economic evaluation, for the production of white Cachama and Nilotíca Tilapia (Black) in cultivation systems Bioflocs. J. Control Theory Appl. ISSN, 0974–5572 (2016)
6. Suber, P.: Will open access undermine peer review? The SPARC Open Access Newsletter, issue 113 (2009). http://www.earlham.edu/~peters/fos/newsletter/09-02-07.htm
7. Donner, P.: Effect of publication month on citation impact. J. Inf. **12**(1), 330–343 (2018)
8. Stegehuis, C., Litvak, N., Waltman, L.: Predicting the long-term citation impact of recent publications. J. Inf. **9**(3), 642–657 (2015)

9. Salter, A., Salandra, R., Walker, J.: Exploring preferences for impact versus publications among UK business and management academics. Res. Policy **46**(10), 1769–1782 (2017)
10. Radicchi, F., Weissman, A., Bollen, J.: Quantifying perceived impact of scientific publications. J. Inf. 11(3), 704–712 (2017)
11. Yu, T., Yu, G., Wang, M.Y.: Classification method for detecting coercive self-citation in journals. J. Inf. **8**(1), 123–135 (2014)
12. Reingewertz, Y., Lutmar, C.: Academic in-group bias: An empirical examination of the link between author and journal affiliation. J. Inf. **12**, 74–86 (2018)
13. Stremersch, S., Camacho, N., Vanneste, S., Verniers, I.: Unraveling scientific impact: citation types in marketing journals. Int. J. Res. Mark. **32**(1), 64–77 (2015)
14. Sánchez, S., Gorraiz, J., Melero, D.: Reference density trends in the major disciplines. J. Inf. **12**, 42–58 (2018)
15. Solomon, D., Laakso, M., Björk, B.: A longitudinal comparison of citation rates and growth among open access and subscription journals. J. Inf. **7**(3), 642–650 (2013)
16. Boyack, K.W., Van Eck, N.J., Colavizza, G., Waltman, L.: Characterizing in-text citations in scientific articles: a large-scale analysis. J. Inf. **12**, 59–73 (2018)
17. Néda, Z., Varga, L., Biró, T.S.: Science and Facebook: the same popularity law! PLoS ONE **12**(7), e0179656 (2017). https://doi.org/10.1371/journal.pone.0179656
18. Mingers, J., Xu, F.: The drivers of citations in management science journals. Eur. J. Oper. Res. **205**(2), 422–430 (2010)