Statistical Analysis of the Impact of International and Industry Collaboration on Citations per Paper for Universities

Ruben Aurelio Puebla Gutierrez, Laura Jaideny Perez Gomez

May 29, 2021

Abstract

Scientometrics has been brought to the spotlight, in part, thanks to university rankings heavily relying on research indicators. One of the core research indicators used by such rankings is citations per paper as its commonly accepted as a good measure of research quality. Universities compete for better positions in such rankings to attract both better students and researchers so special attention has been brought into what factors contribute to the increase of citations per paper that an university achieves. Several studies have tackled the problem of identifying such factors; however, none has considered international and industry collaboration without identifying the distribution for the main variable – citations per paper – first. Papers that have considered such factors have either employed other methodologies beyond statistics or have relied on the central limit theorem under a frequentist approach. Using a dataset of the top 200 universities according to the 2020 QS rankings, this work identifies that Weibull is the best distribution for citations per paper and proceeds to test the impact of international and industry collaboration using three approaches – Fisherian, Bayesian, and regression analysis –, all of which make use of the identified distribution. The results of testing hypotheses with the three methodologies support previous findings that, in fact, international and industry collaboration are major factors that contribute to citations per paper, providing more evidence for supporting this relationship under the scope of a previously not considered approach.

Keywords— publication quality, international collaboration, industry collaboration, Weibull

1 Introduction

Scientometrics has emerged as a quantitative study of science, particularly, for the context of this paper, institutional research's science indicators. Indicator databases of publications have allowed extracting quantifiable criteria related to each publication. Analysis around scientometrics has recently gained importance due to its capacity to unveil relevant indicators that may help evaluate and compare research among institutions. Some of these institutions are educational such as universities whose research productivity is considered on several university rankings [1].

World university rankings have been used as a criterion of excellence by higher education institutes. Being positioned in a higher rank allows universities to be more attractive to new students and potential researchers. Universities are commonly ranked based on academic and research indicators such as publication per author, citations per paper, and papers indexed in top journals. Every year, the QS World University Ranking ranks universities using Scopus database indicators, university's portfolio, and surveys, distributed in six indicators. These indicators include academic reputation, employer survey, citations per faculty, faculty-student ratio, international students percentage, and international faculty percentage [2]. As it can be seen, research indicators are relevant in the university evaluation performed by the QS ranking as well as for others such as the ARWU and the Times Higher Education (THE). Therefore, it is essential to conduct studies that provide universities with reliable information regarding areas of opportunity that allow them to improve the quality of their scientific publications and aspire to better global positioning.

Considering different research indicators, a contrast-pattern-based model was used to identify patterns that describe the top universities in different university rankings. The research conducted by Loyola-Gonzalez, et al. [3] implements this approach, where a set of contrast patterns that describes the top 100 world university ranking, according to the Quacquarelli Symonds (QS) World University Ranking, as well as another one describing universities ranked from 101 to 200 are provided. Its main aim was to provide patterns with research indicators describing the top 100 universities to help other institutions identify improvement areas and conduct better decision-making.

This paper follows the same context described in [3] with the primary motivation of detecting what drives scientific publication quality on the top 200 universities based on the QS ranking. Therefore, this article's main aims are to find which scientometric factors impact the most a university's scientific publication quality and hence provide universities worldwide with reliable insights that can help them generate new research strategies. To achieve this, the database compiled by Loyola-Gonzalez, et al. [3] for their research was used. This database contained 34 features extracted from SciVal that describe relevant research indicators from 2016 to 2018 from the top 200 ranked world universities based on the 2020 QS ranking.

Concerning the database features, two main research questions arise as part of this work:

1. Does a higher percentage of university publications having international collaboration increase the university's scientific publication quality, measured in a higher ratio of citations per publication?
2. Does a higher percentage of university publications having industry collaboration will increase the university's scientific publication quality, measured in a higher ratio of citations per publication?

The present research aims to answer them to fulfill the objective of this study by conducting statistical hypothesis tests regarding some of the original database universities' research indicators. Therefore, the impact of international and industry collaboration on the number of citations per paper received by the universities' scientific publications will be tested. The main contributions of this paper can be described as follows:

- This is the first article that identifies the distribution of the main variable – citations per paper –, in this case as Weibull, rather than relaying on the central limit theorem for testing the impact of international and industrial collaboration factors.
- Consequently, it is too the first article to employ Fisherian, Bayesian, and regression analysis under this distribution consideration to prove that international and industrial collaboration are factors that contribute significantly to citations per paper.

The rest of the article is structured as follows. The next section briefly reviews the related literature; section 3 deeply describes the problem definition and formulation; section 4 presents the analytical approach conducted to test the proposed statistical hypotheses; finally, conclusions and future research are presented in the last section.

2 Literature Review

This study aims to explore the influence of two factors on the quality of universities scientific publications: international and industry collaboration. This section explores works that conduct a similar approach based on identifying research indicators that could be directly related to the quality of scientific publications.

Measuring the quality of a scientific article using research indicators has gained relevance in recent years, mainly since the quality of research carried out in a university institution contributes to its classification within different university rankings. The number of citations is often considered as an indicator of a scientific publication's impact [4]. For this reason, identifying the factors that influence the number of citations has become a common research topic for both researchers in general and universities.

One approach was developed by the authors in [5] who conducted a study on the number of citations received over thirteen years, for articles published in five major marketing journals, based on the characteristics of the article and the author. To do this, they performed statistical hypothesis tests with which they tested those factors that significantly impacted the citations received from an article. They found that the number of citations received depends on the domain of the article, as well as the visibility and personal promotion of the author, while the expository clarity of the article and the use of attention grabbers have less influence.

Other works have applied regression analysis approaches. Authors in [6] considered the number of citations received by articles published in six management science journals over 1990 to evaluate the impact that factors related to the author, the journal, and the article itself could have on it. To do this, the authors developed a regression equation with a generalized linear model (GLM) between the citations and these factors, finding that the rank of the author's institution and the journal of publication, as well as the number of references and the length of the article, have a significant influence on the number of citations received. The GLM model they used was based on the negative binomial family as they found it is adequate for single article citation numbers. Similarly, authors in [7] presented a linear regression analysis to prove that the length of the article and the impact factor of the journal could predict the number of citations in major medical journals articles. Besides, a linear regression analysis was used in [8] to confirm that the number of co-authors significantly influences the citation of Italian publications, concluding that there is a consistent and significant linear growth in the number of citations received by a publication in most of the thematic categories, being more notable in the areas of Social Sciences and Art and Humanities.

Authors in [9] and [10] used linear mixed models to test the effects of the open-access state and the characteristics of article reference lists, respectively, on the number of citations that an article receives. In [9], it was found that the open access status of a publication increases the citation rate of that article by approximately one citation per year, concluding that better decisions can be made considering the advantages of open electronic accessibility of scientific articles. Additionally, researchers in [10] found a positive and significant relationship between the number of citations and the total of citations referenced. Also, logistic regression models were used in [11] to analyze the probability that an article is cited based on the year of publication, number of authors, type of publication, and language of publication. Data were taken from articles published in the journal Gaceta Sanitaria, a public health journal in Spanish, and it was found that only the year of publication influenced the probability of an article being cited. Finally, in [12], a meta-regression analysis was performed to identify the relationship between the length of an article and citations, finding that the longer an article is, the greater the number of citations it receives.

As it can be seen, several factors related to the authors and articles have been explored using a range of different techniques; however, given the aim of this research, publications considering different types of collaboration factors were found. In [13], the author starts from the premise that international collaboration in scientific publications influences their impact, measured by citation-based indicators, due to the phenomenon of self-citation by its various authors. He exposes a simple mathematical relationship between self-citations and publication impact to show that the high rates of self-citations in international collaboration do not entirely determine the number of citations per publication. Similarly, authors in [14] presented an analysis carried out with Spanish Web of Science publications related to the field of computer science to discover the relationship between collaboration in research, the number of documents, and the number of citations. It was found that international collaborations has the highest average number of citations per document and year compared to documents with institutional and national collaboration. Additionally, using the Kruskal-Wallis non-parametric statistical test, it was shown that the number of authors does not affect the number of citations per document and year at the international level.

Following a similar purpose, other investigations have relied on regression analysis to discover the impact of collaborations on the citations received by an article. An example is found in [15] where authors developed a regression analysis considering the number of citations as a dependent variable to examine the citation patterns regarding the collaborative dimensions of the publications of South African scientists. The results obtained indicated that the number of authors could predict the number of citations received by an article, the number of authors from foreign countries, and the type of collaboration, national or international, that the authors have between them. In such a way that if the authors of a publication have international collaboration or internal-institutional collaboration between them, the publication is destined to have a greater impact on citations. Likewise, researchers in [16] constructed a multiple linear regression model taking as the dependent variable the total number of citations received representing the quality of the article. The authors used SPSS software to perform step-by-step regression between different samples, and through this analysis, they concluded that the number of countries that cooperate with the institution has a significant positive impact on citations per article.

It was explored in [17] how different types of collaboration influence the variation in the average number of citations per article or impact of the article, as well as the impact of publications involving at least one industrial partner. The types of collaboration analyzed included collaboration with authors from the same institution, national collaboration, and international collaboration. Using regression analysis, the authors found that there is a linear relationship between the impact of the article and the number of international contributors, increasing by approximately 1.6 citations for each additional international contributor. Therefore, concluding that publications with international collaboration have a greater impact than articles with collaboration from the same institution or with any other national institution, and founding also that when industry collaborates, the impact of the paper is greater than when it does not collaborate. More specifically, authors in [18] studied the collaboration between academia and industry, arguing that it contributes to a higher level of scientific production and verifying through regression models that university-industry collaboration produces significantly more follow-on-citation-weighted publications. In addition, the authors in [19] showed that publications that involve collaborations between the university and industry in Canada receive a greater number of citations, significantly increasing their scientific impact.

The literary review allows one to confirm the general use of the number of citations as a performance measure. In addition, it is noticeable that the impact different qualitative and quantitative factors could have on the number of citations that an article receives has been a research topic for several years. In terms of collaboration factors, articles have shown special attention to international and industry collaboration. However, no article studying those two factors has identified and used the best distribution for citations per paper – used to evaluate university-wide publication quality rather than a single publication quality. Particularly, it also means that the three approaches – Fisherian, Bayesian, and a regression analysis – with the distribution consideration are novel in terms of its application.

3 Problem Definition

This work's intent is to provide statistical support for the factors that drive a university's research quality. This complements previous findings based on other methodologies that arrived to similar conclusions to the ones shown here, as seen in the previous section. Particularly, the goal is to analyze from a Fisherian, Bayesian, and regression standpoint how international collaboration and collaboration with the industry improve the number of citations per paper. Ergo, international and industry collaboration are the factors under consideration, while citations per paper is the proxy for research quality.

The dataset used for this study was retrieved from the research performed by Loyola-Gonzalez, et al. [3], which has features that describe relevant research indicators from 2016 to 2018 for the top 200 universities as considered in the QS World University Ranking. From this dataset, only the variables present in the statistical hypothesis are used. A detailed definition of the variables is found in [3]. Two scientific hypotheses are presented, and for each of them, two statistical hypotheses are considered:

1. International collaboration increases a university's scientific publication quality.

   Statistical Hypothesis 1: There is a positive linear regression slope between the percentage of papers with international collaboration for universities (intColYYYY) and citations per paper for universities (citPPYYYY) for a given year YYYY.

   $$\begin{gathered}\beta_1 = \text{linear regression slope between intColYYYY and citPPYYYY} \\ H_0 : \beta_1 \leq 0 \\ H_a : \beta_1 > 0\end{gathered}$$

   Statistical Hypothesis 2: Citations per paper for universities (citPPYYY) are statistically significantly greater for universities with international collaboration (intColYYYY) above the median vs universities with international collaboration (intColYYYY) below the median for a given year YYYY.

   $$\begin{gathered}\mu_1 = \text{mean of citPPYYYY with intColYYYY above median} \\ \mu_2 = \text{mean of citPPYYYY with intColYYYY below median} \\ H_0 : \mu_1 \leq \mu_2 \\ H_a : \mu_1 > \mu_2\end{gathered}$$

2. Industry collaboration increases a university's scientific publication quality.

   Statistical Hypothesis 3: There is a positive linear regression slope between the percentage of papers with industry collaboration for universities (acaColYYYY) and citations per paper for universities (citPPYYYY) for a given year YYYY.

   $$\begin{gathered}\beta_1 = \text{linear regression slope between acaColYYYY and citPPYYYY} \\ H_0 : \beta_1 \leq 0 \\ H_a : \beta_1 > 0\end{gathered}$$

   Statistical Hypothesis 4: Citations per paper for universities (citPPYYY) are statistically significantly greater for universities with industry collaboration (acaColYYYY) above the median vs universities with industry collaboration (acaColYYYY) below the median for a given year YYYY.

   $$\begin{gathered}\mu_1 = \text{mean of citPPYYYY with acaColYYYY above median} \\ \mu_2 = \text{mean of citPPYYYY with acaColYYYY below median} \\ H_0 : \mu_1 \leq \mu_2 \\ H_a : \mu_1 > \mu_2\end{gathered}$$

In the case of mean difference hypotheses (i.e. statistical hypotheses 2 and 4), citations per paper for a university is the main variable. Fisherian and Bayesian analyses of these hypotheses are presented in section 4.1 and 4.2 respectively.

The rationale for splitting the data points using the median of the two considered factors is to be able to compare two equally sized groups: one containing the upper ranked universities given a factor and another with the lower ranked universities given the same factor.

In the case of regression hypotheses (i.e. statistical hypotheses 1 and 3), citations per paper for a university is the response variable. Regression analyses are presented for these hypotheses in section 4.3.

Regression and mean difference hypotheses in this case complement each other, supporting in more than one way the objective of demonstrating the significance of the two factors.

4 Analytic Solution

As it was mentioned at the end of section 3, citations per paper for a university in a given year (citPPYYY), depending the type of hypothesis, is either the main variable or the response variable for all hypotheses. It is important, therefore, to analyze it and determine the best statistical distribution for it.

From Fig. 1, it can be noted that citPPYYYY tends to decrease for later years. This is the case because newer papers have had less time than older papers to be cited. Also, considering that the dataset contains 200 entries, it is observed that the number of outliers is below 3%. In fact, it is at most 1.5%. Finally, a slight skewness is observed. Although mostly well-behaved, in an attempt to improve the behaviour of the distribution for this variable, different transformations were considered.

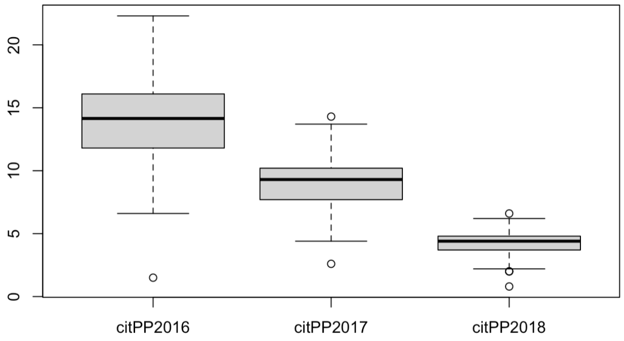
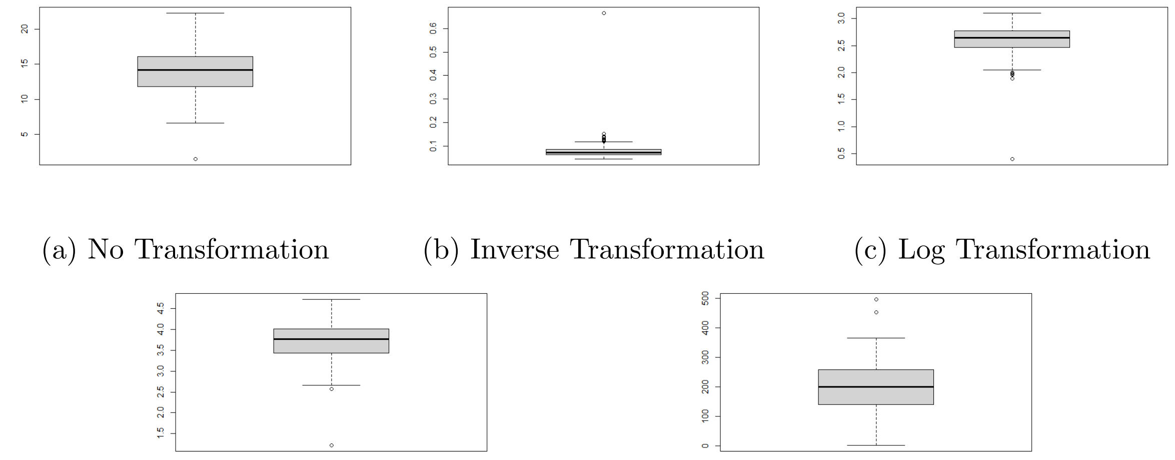


Figure 1: citPPYYYY box plots.

For the considered transformations, only the results for the year 2016 are shown in this document as other years matched the behaviour herein. In Fig. 2, the box plots for the resulting data points for different common transformations are shown. It can be observed that all transformations increased the number of outliers vs no transforming the data. Also, in no case was the slight skewness reduced.



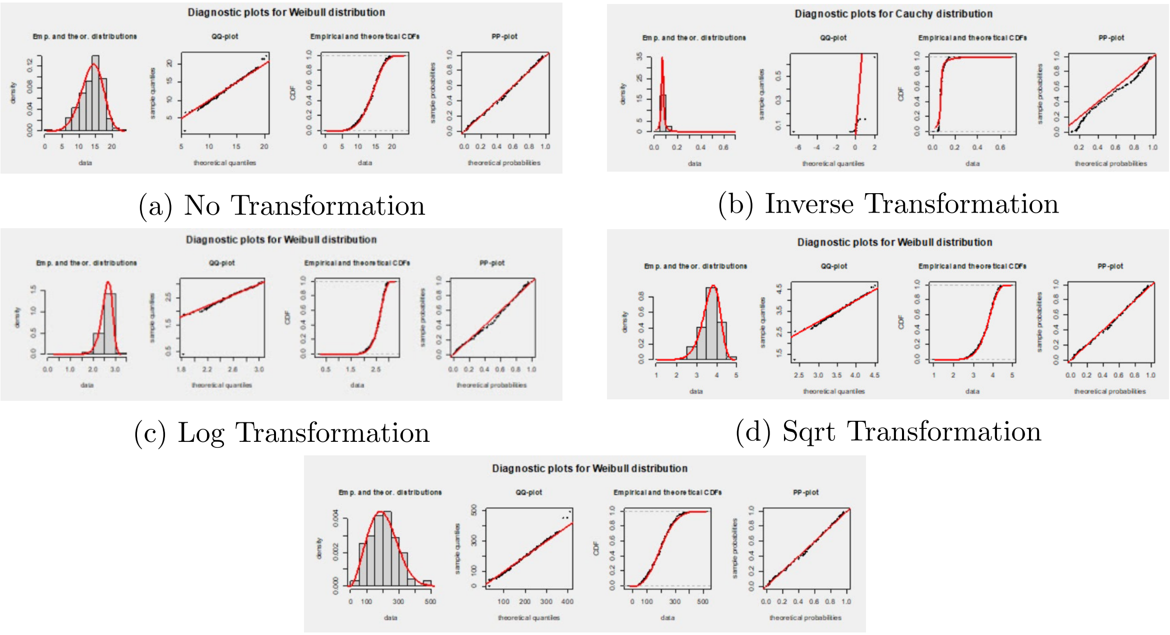
(a) No Transformation (b) Inverse Transformation (c) Log Transformation

(d) Sqrt Transformation (e) Square Transformation

Figure 2: Transformation box plots for citPP2016.

To support the decision of whether or not to transform citPPYYYY, and, if so, what transformation to follow, the best distribution fit for each transformed data was found. Diagnostics plots are shown in Fig. 3. Note that the PP and QQ plots for different transformations are either as well behaved or worse than those of the data without transformation. Taking in consideration this in addition to the results discussed for the box plot results, it was decided that proceeding without transformation was the best way to move forward.

Now, in order to find the best distribution for citPPYYYY robustly, bootstrap was used. For every bootstrap iteration, the maximum log likelihood for different distributions was calculated on the sampled data with replacement, and the fitted distribution with the greatest maximum log likelihood of the iteration received one point. After 1000 iterations, the distribution with the highest score was selected. The results outlined in Table 1 show that Weibull was the winner by a large margin for all years. This also aligns with the well behaved QQ and PP plots observed for the data with no transformation for the Weibull distribution (see Fig. 3).



(a) No Transformation

(b) Inverse Transformation

(c) Log Transformation

(d) Sqrt Transformation

(e) Square Transformation

Figure 3: Best distribution fit diagnostic plots for different citPP2016 transformations.

Table 1: Bootstrap best-fit by maximum log likelihood scores.

Variable	Logistic	Gompertz	**Weibull**	Normal	Gamma
citPP2016	207	55	630	107	1
citPP2017	173	7	666	153	1
citPP2018	275	170	519	36	0

4.1 Fisherian Analysis

Having determined that Weibull is the best distribution for citPPYYYY, the main variable of statistical hypotheses 2 and 4, Maximum Likelihood Estimation (MLE) is used to determine the best parameter values for the distribution - both its scale λ and shape k parameters.

The pdf for Weibull distribution is defined as follows:

$$f_{k,\lambda}(x_i) = \begin{cases} \frac{k}{\lambda}\left(\frac{x_i}{\lambda}\right)^{k-1} e^{-(x_i/\lambda)^k} & x_i \geq 0 \\ 0 & x_i < 0 \end{cases} \tag{1}$$

where x_i is a single data point.

Because the density function is zero when x is less than zero, Weibull can only be a good fit for data that is non-negative; the likelihood for a sample would be zero regardless of the distribution parameters if there is a negative data point. citPPYYYY is always non-negative, so the log likelihood for a sample x of size N can be defined as

$$\ell(k,\lambda) = N\log\frac{k}{\lambda} - \sum_{i=1}^{N}\left(\frac{x_i}{\lambda}\right)^k + (k-1)\sum_{i=1}^{N}\log\frac{x_i}{\lambda} \tag{2}$$

and its score function is

$$\dot{\ell}(k,\lambda) = \nabla\ell(k,\lambda) = \begin{bmatrix} \frac{\partial\ell(k,\lambda)}{\partial k} \\ \frac{\partial\ell(k,\lambda)}{\partial\lambda} \end{bmatrix} = \begin{bmatrix} \frac{N}{k} - \sum_{i=1}^{N}\left[\left(\frac{x_i}{\lambda}\right)^k - 1\right]\log\frac{x_i}{\lambda} \\ k\lambda^{-k-1}\sum_{i=1}^{N} x_i^k - \frac{kN}{\lambda} \end{bmatrix} \tag{3}$$

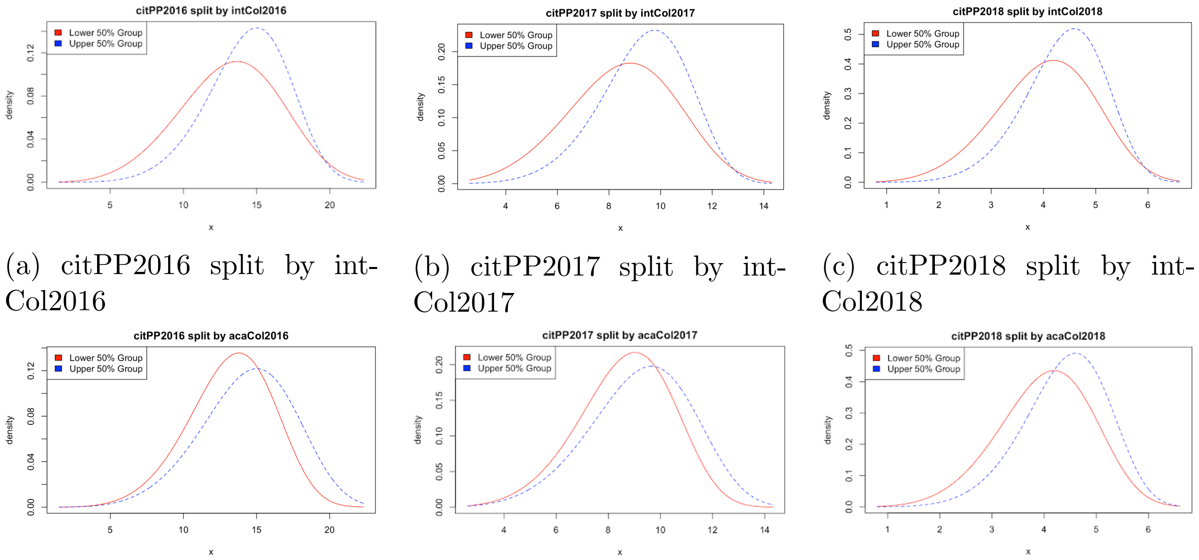
Solving for zero to find the MLE k and λ parameters is difficult to do analytically, so numerical methods were used to find for the MLE parameters. This was done for each data split of every year for the two considered factors, yielding 12 estimations in total as seen in Table 2. In addition, the mean (μ) value of the distribution was calculated with $\lambda\Gamma(1+1/k)$ as that's the mean for the Weibull distribution given its parameters.

Table 2: Weibull MLE parameters for each data split group.

Main Variable	Split Variable	Group	shape (k)	scale (λ)	mean (μ)
citPP2016	intCol2016	lower 50%	4.293487	14.52335	13.21759
citPP2016	intCol2016	upper 50%	5.940575	15.48840	14.36096
citPP2017	intCol2017	lower 50%	4.510490	9.329910	8.515377
citPP2017	intCol2017	upper 50%	6.264777	10.04647	9.342437
citPP2018	intCol2018	lower 50%	4.819969	4.403007	4.034126
citPP2018	intCol2018	upper 50%	6.558878	4.703421	4.384671
citPP2016	acaCol2016	lower 50%	5.191100	14.36925	13.22196
citPP2016	acaCol2016	upper 50%	5.082883	15.67596	14.40685
citPP2017	acaCol2017	lower 50%	5.424488	9.366747	8.640597
citPP2017	acaCol2017	upper 50%	5.286839	10.03708	9.245373
citPP2018	acaCol2018	lower 50%	5.071977	4.375917	4.021148
citPP2018	acaCol2018	upper 50%	6.227510	4.730834	4.397879

A visualization of the lower 50% vs the upper 50% groups by main variable and year is found in Fig. 4 which uses the parameters found in Table 2. Here, it is possible to see that the mean of the upper group tends towards being higher than that of the lower group, already providing some indication that the means might be higher for the upper 50% groups as expected by the alternative hypotheses. What remains to evaluate statistical hypotheses 2 and 4 is estimating the variance of the mean of the distributions.

The expected value of the derivative of the score function is negative N times its Fisher information $E[\ddot{\ell}(k,\lambda)] = -NI(k,\lambda)$ and Fisher's fundamental theorem for the MLE states that the variance of a parameter is approximated by $\frac{1}{NI(k,\lambda)}$, following a normal distribution. This means



(a) citPP2016 split by intCol2016

(b) citPP2017 split by intCol2017

(c) citPP2018 split by intCol2018

(d) citPP2016 split by acaCol2016

(e) citPP2017 split by acaCol2017

(f) citPP2018 split by acaCol2018

Figure 4: Upper (blue) vs lower (red) 50% group data distributions by factor and year.

that the variance of a parameter is approximated by $-\frac{1}{\ddot{\ell}(k,\lambda)}$.

The derivative of the score function for the scale parameter

$$\frac{\partial^2 \ell(k,\lambda)}{\partial \lambda^2} = -k(k+1)\lambda^{-k-2}\sum_{i=1}^{N} x_i^k + \frac{kN}{\lambda^2} \tag{4}$$

was used to calculate the variance for the scale parameter as seen in Table 3, and considering that the greatest parameter variance came from the scale parameter, the variance of the mean (μ) was obtained using the delta method. The mean of a Weibull distribution given its parameters is $\lambda\Gamma(1 + 1/k)$, so the variance of μ was approximated by

$$var[\mu] = \Gamma^2(1 + 1/k) \cdot var[\lambda] \tag{5}$$

with results shown in Table 3 as well.

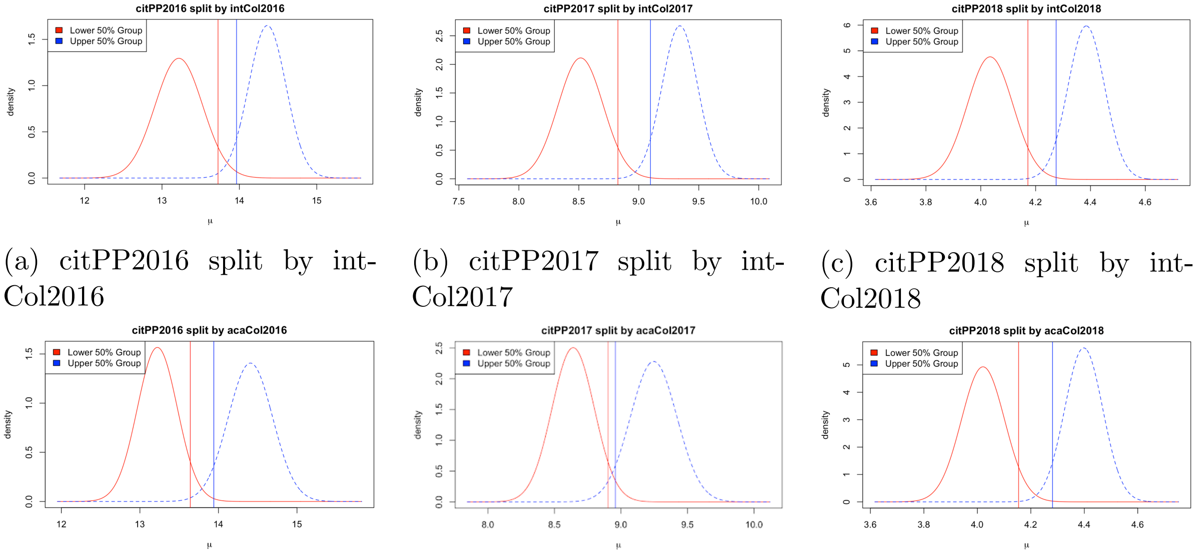
As a note, Cramer Rau Lower Bound (CRLB) plots for the estimated means all rendered expected results, yielding square mean differences for the estimated mean below the CRLB for randomized sample estimations in approximately 70% of the cases.

Using the information from Tables 2 and 3 the distribution of the mean for each data group can be plotted as seen in Fig. 5. Here, the 95th percentile for each lower 50% group is shown as a vertical line as well as the 5th percentile for each upper 50% group. Noticing that such lines do

Table 3: Scale parameter and mean variance for each data split group.

Main Variable	Split Variable	Group	var[scale (λ)]	var[mean (μ)]
citPP2016	intCol2016	lower 50%	0.1144628	0.09480597
citPP2016	intCol2016	upper 50%	0.0679631	0.05842884
citPP2017	intCol2017	lower 50%	0.0427912	0.03564573
citPP2017	intCol2017	upper 50%	0.0257290	0.02224937
citPP2018	intCol2018	lower 50%	0.0083458	0.00700601
citPP2018	intCol2018	upper 50%	0.0051414	0.00446816
citPP2016	acaCol2016	lower 50%	0.0765808	0.06484012
citPP2016	acaCol2016	upper 50%	0.0951296	0.08034990
citPP2017	acaCol2017	lower 50%	0.0298111	0.02536817
citPP2017	acaCol2017	upper 50%	0.0360401	0.03057881
citPP2018	acaCol2018	lower 50%	0.0077525	0.00654645
citPP2018	acaCol2018	upper 50%	0.0058292	0.00503757

not cross and the small overlap of the mean distributions, the null hypothesis can be refuted in all cases. The alternative hypothesis for the six cases is supported, evidencing the importance of both international and industry collaboration in increasing citations per paper for universities.



(a) citPP2016 split by intCol2016

(b) citPP2017 split by intCol2017

(c) citPP2018 split by intCol2018

(d) citPP2016 split by acaCol2016

(e) citPP2017 split by acaCol2017

(f) citPP2018 split by acaCol2018

Figure 5: Upper (blue) vs lower (red) 50% group mean distributions by factor and year.

4.2 Bayesian Analysis

Continuing with the consideration that citPPYYYY is Weibull distributed, it is important to mention that it can be expressed in exponential family form when its shape parameter k is known.

This means that the Weibull PDF (1) can be expressed as

$$f_\lambda(x_i) = x_i^{k-1} e^{\log k - k \log \lambda - \left(\frac{x_i}{\lambda}\right)^k} \tag{6}$$

to match the exponential family form.

From the previous MLE estimations shown in Table 2 and Fig. 4, the value for the shape parameter k is used. This allows making use of the existing conjugate prior distributions available to exponential families. In the case of the Weibull distribution, its conjugate prior is the inverse gamma distribution for the parameter $\theta = \lambda^k$. It is important to note, therefore, that both the prior and posterior distributions in this approach correspond to the scale parameter to the power of the fixed shape parameter.

The procedure followed in this approach to get the posterior distribution for each of the data groups in hypotheses 2 and 4 can be outlined as follows:

1. One thousand randomized samples of the original data with replacement were drawn.
2. Weibull MLE was applied on them, fixing the shape parameter according to Table 2 to get a thousand estimations of the scale parameter.
3. An inverse gamma distribution MLE parameter α and β estimation was run on the estimated scale parameters to the power of the fixed shape parameter to get the prior distribution.
4. The posterior distribution parameters were calculated using

$$\alpha_{posterior} = \alpha_{prior} + N \tag{7}$$

$$\beta_{posterior} = \beta_{prior} + \sum_{i=1}^{N} x_i^k \tag{8}$$

where N is the number of observed data points x_i of the group.

The results of both the prior and posterior parameters are exhibited in Table 4. Considering that k is fixed and that $\mu = t(\theta) = \theta^{\frac{1}{k}}\Gamma(1 + 1/k)$, it possible to use a simple PDF transformation $g(\mu) = f(t^{-1}(\mu))\left|\frac{dt}{d\mu}\right|$ from the posterior distribution f to its equivalent g in terms of μ. This enables comparing the results better with the ones from section 4.1. Note that such transformation does not alter the overlap when comparing two different data groups, it simply translates the x axis.

Comparing the transformed posterior distributions between the upper and lower 50% groups as in Fig 6, it is possible to see that the upper 50% group has a statistically significantly greater

Table 4: Prior and posterior inverse gamma parameters for the distribution of $\theta = \lambda^k$ in each data group.

Main Variable	Split Variable	Group	α_{prior}	β_{prior}	$\alpha_{posterior}$	$\beta_{posterior}$
citPP2016	intCol2016	lower 50%	73.56835	7104877	173.5684	16859164
citPP2016	intCol2016	upper 50%	7.957203	95954836	107.9572	1269214782
citPP2017	intCol2017	lower 50%	86.22589	2030182	186.2259	4399285
citPP2017	intCol2017	upper 50%	28.93195	52453764	128.932	241779602
citPP2018	intCol2018	lower 50%	97.62418	122278.7	197.6242	248985.9
citPP2018	intCol2018	upper 50%	68.88648	1742568	168.8865	4315060
citPP2016	acaCol2016	lower 50%	33.03357	33084891	133.0336	135072591
citPP2016	acaCol2016	upper 50%	23.9021	27907778	123.9021	146806942
citPP2017	acaCol2017	lower 50%	71.2076	13140100	171.2076	31779844
citPP2017	acaCol2017	upper 50%	58.4718	11295699	158.4718	31036579
citPP2018	acaCol2018	lower 50%	88.3992	155901.8	184.3992	327224.6
citPP2018	acaCol2018	upper 50%	90.22241	1426520	189.2224	3007101

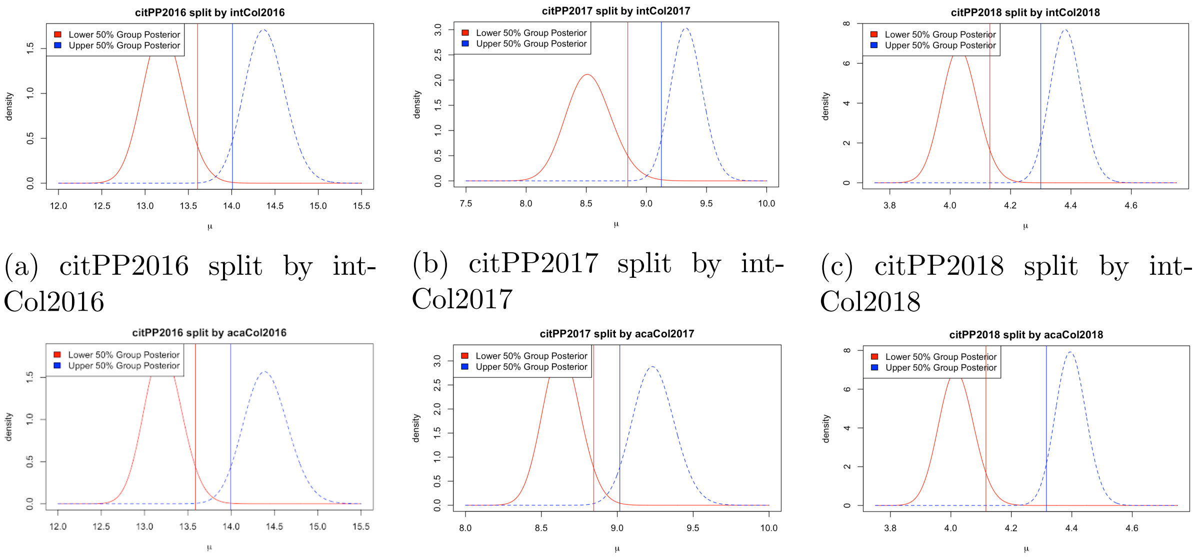
number of citPPYYYY than the lower 50% group for each of the two considered split factors for all years. Herein, the 95th percentile for the lower 50% group is shown as the red vertical line, and the 5th percentile for the upper 50% group is shown as the blue vertical line. Noticing that such lines do not cross and the small overlap of the mean distributions, the null hypothesis can be refuted in all cases, supporting the importance of both international and industry collaboration in increasing citations per paper for universities and agreeing with the results from the Fisherian approach in section 4.1.

Comparing with the results with the ones in section 4.1, the mean distributions gotten from the two approaches are centered around the same values, but the mean distributions for the Bayesian approach have a lower variability. The latter can be noticed as well by the greater gap between the vertical lines, achieving a more robust result for the Bayesian case.

4.3 Regression Analysis

As it was previously mentioned, the distribution of the response variable of statistical hypotheses 1 and 3 (citPPYYYY) is Weibull. In order to find if there exists a positive linear regression slope between such a response variable and the two considered factors (i.e. international and industry collaboration respectively), a Generalized Linear Model (GLM) for the Weibull distribution was fit on the data. Recalling that the Weibull distribution is part of the exponential family when its shape parameter is fixed, it is possible to proceed with GLM.

The linear predictor is related to the scale parameter λ by the logarithm. Essentially, $\log \lambda = \beta_0 + \beta_1 x$. Recalling that $\mu = \lambda \Gamma(1 + 1/k)$ and that k is fixed, it translates in the inverse link



(a) citPP2016 split by intCol2016

(b) citPP2017 split by intCol2017

(c) citPP2018 split by intCol2018

(d) citPP2016 split by acaCol2016

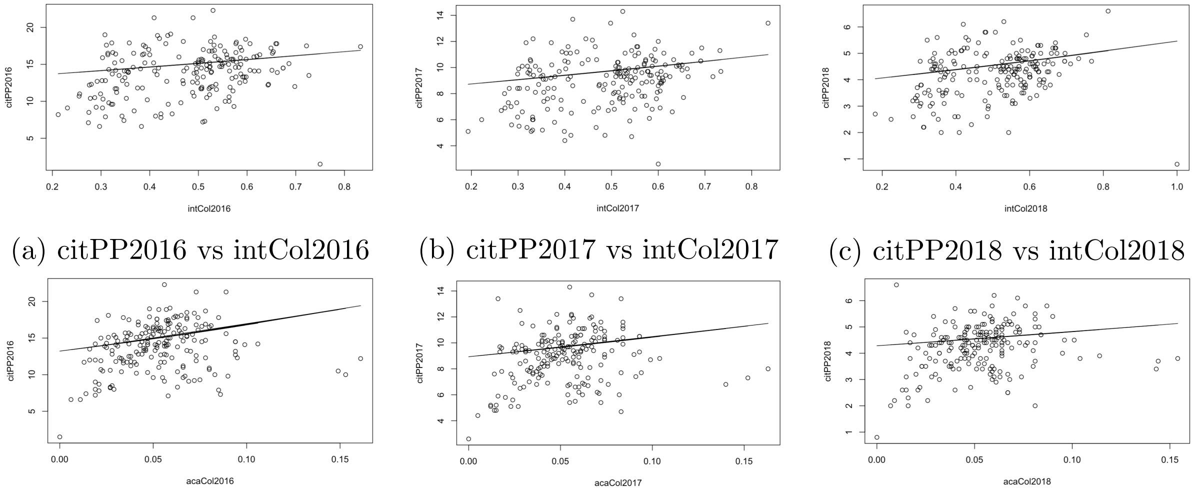
(e) citPP2017 split by acaCol2017

(f) citPP2018 split by acaCol2018

Figure 6: Upper (blue) vs lower (red) 50% group mean posterior distributions by factor and year.

function for the mean as $\mu = e^{\beta_0+\beta_1 x}\,\Gamma(1+1/k)$.

The results of fitting the linear predictor via maximum likelihood give the curves presented in Fig. 7. Here in, the positive slope of the curve is appreciated. The estimation, standard error, and significance of the slope of the linear predictor β_1 under hypotheses 1 and 3 is shown in Table 5.



(a) citPP2016 vs intCol2016

(b) citPP2017 vs intCol2017

(c) citPP2018 vs intCol2018

(d) citPP2016 vs acaCol2016

(e) citPP2017 vs acaCol2017

(f) citPP2018 vs acaCol2018

Figure 7: Citations per paper vs considered factors for all years scatter plots and fitted regression curve.

The regression fit was done using a numerical method know as Rigby and Stasinopoulos. For

Table 5: Regression β_1 estimations, standard errors, t values, and p values.

Response Variable	Independent Variable	β_1 estimate	std error	t value	p value
citPP2016	intCol2016	0.3335	0.1290	2.585	0.00525
citPP2017	intCol2017	0.3614	0.11384	3.172	0.00088
citPP2018	intCol2018	0.3692	0.10210	3.616	0.00019
citPP2016	acaCol2016	2.3902	0.79885	2.992	0.00156
citPP2017	acaCol2017	1.5490	0.71527	2.166	0.01575
citPP2018	acaCol2018	1.1732	0.64595	1.816	0.03454

all regressions, the number of iterations necessary to achieve stability was below 4. Also, residuals were well behaved in all cases, staying with a mean close to zero indicating a good distribution choice for the response variable. Finally, global deviance was low in all cases.

It is evidenced in Table 5 that the GLM regression slope β_1 is significant in all considered cases under an alpha-level of 0.05. Therefore the null is rejected in all cases just as it was too in both section 4.1 and 4.2. This further supports the evidence that both international and industry collaboration are significant factors for increasing the number of citations per paper universities receive. All three methodologies agree on this result.

5 Conclusion and Future Research

Different scientometric indicators related to universities' scientific publications influence the position they occupy within different university rankings. Being positioned as one of the best universities in the world increases the prestige of a university, making it attractive to new talent represented by students and researchers. Therefore, knowing which research factors could significantly impact the quality of their scientific publications has become a topic of interest. Citations per publication are often considered a measure of research quality; therefore, in this article, we explored the influence of two factors – international and industrial collaboration – in the number of citations per publication received by the 200 top universities according to the QS World University Ranking. One of the main findings of this study was that Weibull is the distribution that better fits the variable of citations per publication. With this, hypothesis tests were conducted following three approaches: Fisherian, Bayesian, and regression analysis. Through these tests, it was possible to verify that international and industrial collaboration significantly affect the number of citations in a scientific publication. Moreover, the result obtained was the same with the different approaches, further reinforcing the influence of the two factors studied on the quality of a university's scientific publication quality.

Future work is intended to evaluate the impact that other scientometric factors could have on the quality of university publications, following the same approach applied for the two factors considered in this article. In addition, it would be appropriate to contrast the results obtained herein with ones from distinct university rankings following the same methodology presented here.

References

[1] P. G. Altbach, “The globalization of college and university rankings”, *Change: The Magazine of Higher Learning*, vol. 44, no. 1, pp. 26–31, 2012.

[2] Y.-W. Hou and W. J. Jacob, “What contributes more to the ranking of higher education institutions? a comparison of three world university rankings”, *International Education Journal: Comparative Perspectives*, vol. 16, no. 4, pp. 29–46, 2017.

[3] O. Loyola-González, M. A. Medina-Pérez, R. A. C. Valdez, and K.-K. R. Choo, “A contrast pattern-based scientometric study of the qs world university ranking”, *IEEE Access*, vol. 8, pp. 206 088–206 104, 2020.

[4] I. Podlubny, “Comparison of scientific impact expressed by the number of citations in different fields of science”, *Scientometrics*, vol. 64, no. 1, pp. 95–99, 2005.

[5] S. Stremersch, I. Verniers, and P. C. Verhoef, “The quest for citations: Drivers of article impact”, *Journal of Marketing*, vol. 71, no. 3, pp. 171–193, 2007.

[6] J. Mingers and F. Xu, “The drivers of citations in management science journals”, *European Journal of Operational Research*, vol. 205, no. 2, pp. 422–430, 2010.

[7] M. E. Falagas, A. Zarkali, D. E. Karageorgopoulos, V. Bardakas, and M. N. Mavros, “The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals”, *PLoS One*, vol. 8, no. 2, e49476, 2013.

[8] G. Abramo and C. A. D’Angelo, “The relationship between the number of authors of a publication, its citations and the impact factor of the publishing journal: Evidence from italy”, *Journal of Informetrics*, vol. 9, no. 4, pp. 746–761, 2015.

[9] M. Tang, J. D. Bever, and F.-H. Yu, “Open access increases citations of papers in ecology”, *Ecosphere*, vol. 8, no. 7, e01887, 2017.

[10] S. Mammola, D. Fontaneto, A. Martınez, and F. Chichorro, “Impact of the reference list features on the number of citations”, *Scientometrics*, vol. 126, no. 1, pp. 785–799, 2021.

[11] A. Ruano-Ravina and C. Álvarez-Dardet, “Evidence-based editing: Factors influencing the number of citations in a national journal”, *Annals of epidemiology*, vol. 22, no. 9, pp. 649–653, 2012.

[12] J. Xie, K. Gong, Y. Cheng, and Q. Ke, “The correlation between paper length and citations: A meta-analysis”, *Scientometrics*, vol. 118, no. 3, pp. 763–786, 2019.

[13] A. Van Raan, “The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations”, *Scientometrics*, vol. 42, no. 3, pp. 423–428, 1998.

[14] A. Ibanez, C. Bielza, and P. Larranaga, “Relationship among research collaboration, number of documents and number of citations: A case study in spanish computer science production in 2000–2009”, *Scientometrics*, vol. 95, no. 2, pp. 689–716, 2013.

[15] R. Sooryamoorthy, “Do types of collaboration change citation? collaboration and citation patterns of south african science publications”, *Scientometrics*, vol. 81, no. 1, pp. 177–193, 2009.

[16] W. Fang, S. Dai, and L. Tang, “The impact of international research collaboration network evolution on chinese business school research quality”, *Complexity*, vol. 2020, 2020.

[17] J. Katz and D. Hicks, “How much is a collaboration worth? a calibrated bibliometric model”, *Scientometrics*, vol. 40, no. 3, pp. 541–554, 1997.

[18] M. Bikard, K. Vakili, and F. Teodoridis, “When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity”, *Organization Science*, vol. 30, no. 2, pp. 426–445, 2019.

[19] L.-M. Lebeau, M.-C. Laframboise, V. Larivière, and Y. Gingras, “The effect of university–industry collaboration on the scientific impact of publications: The canadian case, 1980–2005”, *Research Evaluation*, vol. 17, no. 3, pp. 227–232, 2008.