MPRA
Munich Personal RePEc Archive

The Brazilian scientific output published in journals: A study based on a large CV database

Perlin, Marcelo and Santos, André and Imasato, Takeyoshi and Borenstein, Denis and Da Silva, Sergio

2017

Online at https://mpra.ub.uni-muenchen.de/79662/
MPRA Paper No. 79662, posted 12 Jun 2017 04:44 UTC

The Brazilian scientific output published in journals: A study based on a large CV database

Marcelo Perlin[a]*, André A. P. Santos[b], Takeyoshi Imasato[a], Denis Borenstein[a], Sergio Da Silva[b]

[a] *School of Business Administration, Federal University of Rio Grande do Sul, Porto Alegre, R.S., 90010-460, Brazil*

[b] *Graduate Program in Economics, Federal University of Santa Catarina, Florianopolis, S.C., 88049-970, Brazil*

* Corresponding author. Email address: marcelo.perlin@ufrgs.br (M. Perlin).

Abstract

We assemble a massive sample of 180,000 CVs of Brazilian academic researchers of all disciplines from the *Lattes* platform. From the CVs we gather information on key variables related to the researchers and their publications. We find males are more productive in terms of quantity of publications, but the effect of gender in terms of research impact is mixed for individual groups of subject areas. Holding a PhD from abroad increases the chance for a researcher to publish in journals of higher impact, whereas domestic PhDs publish more articles, but in journals of less impact. Thus, there is a trade-off between quantity and research impact. We also find that the more years a researcher takes to finish his or her doctorate, the more likely he or she will publish less thereafter, although in outlets of higher impact. The data also support the existence of an inverted U-shaped function relating research age and productivity.

Keywords: *Lattes* platform, scholarly publishing, scientific productivity, Brazilian researchers

Acknowledgements. We thank the developers of the R (R Core Team 2015) packages that helped us to build the script used in this research: RSQLite, doParallel, DataTable, ggplot2, texreg, dplyr and xtable (Wickham et al., 2014; Analytics and Weston, 2014; Dowle et al., 2015; Wickham, 2009; Leifeld, 2013; Wickham and Francois, 2015; and Dahl, 2014). This research was partially supported by CNPq (Brazilian National Research Council), and Senescyt (Secretariat for Higher Education, Science, Technology and Innovation), Ecuador.

Published: Perlin M, Santos A, Takeyoshi I, Borenstein D, Da Silva S (2017) The Brazilian scientific output published in journals: A study based on a large CV database, *Journal of Informetrics* 11 (1), 18-31.

1. Introduction

To keeping score of the scientific output of individual researchers and institutions, the Brazilian National Research Council (*CNPq*) implemented an individual CV database countrywide, integrated with a web-based query interface. This is called the *Lattes* platform (http://lattes.cnpq.br/). All researchers and institutions are incentivized to keep their records up to date through *Lattes*, recording all the main academic activities and publications. For this reason, *Lattes* can be used to gather good-quality information on individual researchers, as well as to help conduct the performance evaluation of academic and research institutions. Here, we take advantage of the massive database provided by *Lattes* and assemble a sample of 180,000 CVs. Our main interest is to collect the particular information conveyed in the CVs regarding the journal publication destination of Brazilian researchers.

Research activity usually involves a large amount of financial resources. World Bank data (http://data.worldbank.org/) show that Brazilian expenditures on research accounted for 1.2 percent of GDP from 2005 to 2012, which translates to approximately $27 billion at current values.

Scholarly publishing is highly concentrated among a few authors (Lotka, 1926), and scientific research productivity is asymmetrically distributed (Chung and Cox, 1990; Ruiz-Castillo and Costas, 2014). Although what became known as Lotka's law has been disputed recently (Coile, 1977), there is a general agreement in literature that a relatively small proportion of scientists contribute most. Thus, "success breeds success" (Ruiz-Castillo and Costas, 2014; Rorstad and Aksnes, 2015).

Having this is mind, proper identification of the determinants of research productivity in different areas of knowledge can be not only useful for a better allocation of financial resources but also to improve the evaluation of researchers and academic institutions, providing new insights about the factors that directly explain the asymmetric productivity in one particular area and across areas.

Researcher productivity has been examined widely. Previous work points to key factors explaining it. Gender has been regarded as a factor that can explain research productivity asymmetry (Isfandyari-Moghaddam et al., 2012; Morganson et al., 2010; Leta and Lewison, 2003; Prpi, 2002; Puuska, 2009). In one study, female research productivity accounted for only 70.6 percent of male productivity (Prpi, 2002). This result is reinforced as one looks at a sample of associate and full professors, a group heavily skewed toward males (Rorstad and Aksnes, 2015; Abramo et al., 2015; Larivire et al., 2011; Barrios et al., 2013). Moreover, the scant number of female Nobel laureates has less chance to marry and have children as compared to both their male counterparts and the general population (Charyton et al., 2011). This posits a trade-off between career success and family life for female researchers, though this trade-off seems to be lessening for the current generation (van Arensbergen et al., 2012).

The effect of holding a PhD from a foreign university has also been considered in the literature regarding the Brazilian experience. Roos et al. (2014) show strong evidence that Brazilian researchers in the biological sciences with doctorate degrees from abroad are also those who publish less, though in journals of greater impact. Here, the student-supervisor relationship seems to play a role (Tuesta et al., 2015). Indeed, there is significant correlation between future productivity of a PhD candidate and the time spent in his or her interaction with the supervisor.

The previous literature also shows evidence of a non-linear impact of a researcher age on his or her productivity. This suggests the existence of an inverted U-shaped function. Soon after

PhD completion, productivity rises, reaches a peak at a certain age, and then declines (Bonaccorsi and Daraio, 2003; Rorstad and Aksnes, 2015).

The main objective of this study is to provide an overview of the Brazilian scientific output of different subject areas. By using statistical models, we identify the factors that explain researcher productivity in terms of quantity and impact across the different areas. We replicate some findings from literature, contradict others, and further present new findings and insights based on a much larger sample. The massive database allows us to adopt not only a time series approach, but also a cross-section perspective. These make it possible to consider the evolution of the Brazilian research output from almost all subject areas.

Assessing the factors affecting research output using statistical models is not new. This perspective is already present in the studies of Roos et al. (2014), Rorstad and Aksnes (2015), and Diniz-Filho et al. (2016). However, our study is more comprehensive in that it considers a much larger database of researchers from different subject areas in higher education institutions. Roos et al. (2014) consider the same database as ours (*Lattes*), but their study covers only the biological sciences. Rorstad and Aksnes (2015) consider the Norwegian academic staff, which is ten times smaller than the Brazilian one, and Diniz-Filho et al. (2016) analyze data from *Lattes* for only one Brazilian university. By exploring this large dataset of heterogeneous areas, we contribute to the mapping of the profile of the Brazilian academic output, and also to identify the factors explaining volume and impact of research within each subject area.

The next section provides a description of the dataset and methods employed in this study. Section 3 presents and discusses the results. Section 4 concludes.

2. Materials and methods

2.1 *Collecting the data*

The *Lattes* platform was launched in August 1999. It replaced the prior paper-based and non-integrated electronic system, which conveyed information from individual academic researchers. In the subsequent two years, the number of CVs posted under *Lattes* grew more than 300 percent, from 35,000 CVs to more than 100,000. In 2010, *Lattes* provided high quality data on about 1.6 million researchers and about 4,000 institutions (Lane, 2010).

An open, sound and consistent platform for measuring all the activities that make up academic productivity is a worldwide-acknowledged need. In this regard, a standardized platform such as *Lattes* is an asset because “metrics are data driven, so developing a reliable, joined-up infrastructure is a necessary first step” (Lane, 2010). *Lattes* is now internationally perceived as a powerful example of good academic practice.

Our study departs from others that use information from *Lattes* (such as Almeida and Guimarães, 2013; Leite et al., 2011) in at least one important aspect. We develop a software for mining big-scale data that renders us a sample significantly larger than any other previously considered in *Lattes* studies. The software, labeled *Nilrep-Lattes*, is written in Python and registered at the Brazilian National Institute of Industrial Property (code number BR512014000516-0). *Nilrep-Lattes* takes as an input a list of CVs from *Lattes*, such as http://lattes.cnpq.br/x, where x is a numeric value of 16 digits that is unique to each researcher. The software then collects information for every researcher, such as name, PhD origin, PhD completion time, years since PhD completion, current job address, and list of published papers, including publication year and journal ISSN.

Lattes is mostly used by academics, from undergrad students to full professors. While it provides a rich dataset, it also has shortcomings. The user supplies all information and updates are not automatic. Thus, some users may feed wrong information. However, because *Lattes* is widely used as a performance assessment tool, this creates a strong incentive for researchers to keep their *Lattes* profiles updated with accurate information. When signing up for *Lattes*, researchers are held legally responsible for posting false information. While we cannot track the occurrence of such cases in the data we analyze, we expect them to represent a very small proportion of the thousands of CVs in the dataset, thus having negligible effects on our results. We checked the accuracy of the information in our dataset by considering the frequency of the last date of update for the whole database and found 81.27 percent of the CVs had been updated after 2014. Thus, the bulk of the records reflect up-to-date information.

2.2 *SJR and "Qualis"*

The use of citation analysis (CA) to evaluate academic research is a controversial issue. Although there is an increasing consensus that there is no single metrics capable of representing research quality *per se* (Bollen et al., 2009; Moed, 2005), the use of CA is on the rise due to its great accessibility made it possible by the introduction of new information technologies. This fact can be confirmed by the proliferation of CA indicators over the last two decades. In particular, the ISI impact factor of a journal has been used as a proxy for quality and the expected impact of each of the papers published by the journal (Bordons et al., 2002). (ISI refers to The Institute for Scientific Information, which was subsequently acquired by Thomson Reuters). Here, we consider the CA indicators to gauge research impact rather than quality. After all, one should be cautious when assessing research quality in large scale datasets such as ours (Moed, 2005).

For measuring an academic journal impact, we consider the SCImago Journal Rank (SJR). SJR accounts for both the number of citations received by a journal and the prestige of the journals where such citations come from. It has emerged as an alternative for the ISI impact factor. Its advantages involve 1) open-access nature; 2) larger source databases (SJR has 29,713 indexed journals and analysis of several non-English-language journals, while ISI has 11,719); and 3) a gauge of the quality of citations (Hall, 2011). Falagas et al. (2008) compared both measures and concluded that the SJR poses as serious alternative to the ISI impact factor. Adopting a broader metrics seems more appropriate to our study, as we have in mind that Brazilian research is still maturing and has not yet reached its peak (see Gonzalez-Pereira et al., 2010 for a discussion). We also verified a sensitivity analysis of our results by replacing the use of the SJR with the Journal Citation Reports (JCR) of Thomson Reuters (available upon request). Even dropping a significant part of data, results did not change a great deal. Thus, the conclusions of our study are unlikely to be changed by the choice of impact indicator.

For each researcher, we collect a journal's ISSN from the CVs and then feed the SJR table (available at http://www.scimagojr.com/) with an ISSN to get the journal rank. The bulk of the Brazilian research output is published in domestic outlets. In our dataset, nearly 66 percent of published papers are not tracked by the SJR. To remedy this situation, we decided to consider further the alternative classification provided by the Brazilian "*Qualis*," which takes into account the majority of domestic journals.

Qualis provides journal rankings since 2000 (http://qualis.capes.gov.br/webqualis/). It is an initiative of the Coordination for the Improvement of Higher Education Personnel (*CAPES*), a federal agency responsible for establishing criteria for evaluating the performance of graduate education institutions. According to *CAPES*, *Qualis* has been conceived to meet the graduate programs' "particular needs of the evaluation process." All journals are evaluated by committees of each assessment area, following criteria previously defined for the area. These are then endorsed by *CAPES*'s scientific committee, which balances what is considered the relative importance of different journals for a given area.

The particularities of the subject areas perhaps justify the use of *Qualis*. Linguistics and arts, for example, usually have a comparatively lower number of publications indexed by international indicators. Ignoring such publications not tracked by the SJR generates a bias toward areas of higher international insertion, such as medicine. Thus, a more balanced approach is to consider *Qualis* along with the SJR.

Qualis is made up of eight strata: A1 and A2 (for the most highly rated journals), B1, B2, B3, B4, B5 and C (lower rating). *CAPES* suggests the committees responsible for generating the rankings for each subject area to openly display the criteria used, which should involve "impact factor, *h*-index, and other ways of measuring quality." In particular, *CAPES* recommends "the A1 and A2 strata be rigorously justified", considering only "journals indexed by the JCR, with an ISSN, which have recognized editorial board and a well-functioning refereeing process." The criteria for the rankings considered by each assessment area are made public in *CAPES*'s website (at http://www.avaliacaotrienal2013.capes.gov.br/documento-de-area-e-comissao).

Australia, Colombia, and Norway also have domestic systems of journal evaluation and stratification similar to *Qualis*. The particularity of the Brazilian evaluation system is that every publication is categorized according to one of the eight strata at an area committee's discretion, and then a number is assigned to it, in order to match order of a stratum and order of a number. The points ascribed to each stratum also vary across the disciplines. For instance, for business administration A1 = 100, A2 = 80, B1 = 60, B2 = 50, B3 = 30, B4 = 20, B5 = 10 and C = 0, while for economics A1 = 100, A2 = 80, B1 = 60, B2 = 40, B3 = 25, B4 = 15, B5 = 5 and C = 0. Three papers rated as B2 in economics, for example, amass 120 points, thus outweighing one rated as A1 (100 points). A same journal can also be rated differently across the assessment areas. In the last triennium, the journal *Disability and Rehabilitation* was rated as A1 by the *Qualis* of physical education, and rated as B2 by the *Qualis* of medicine.

The *Qualis* of economics was compared to the simple impact indicators provided by the economics repository *RePEc* (http://repec.org/) and major biases were identified, including a domestic journal bias and hostility toward journals from other subject areas (Da Silva, 2009; Guimarães, 2012). We speculate that similar biases may occur in other subject areas.

In view of the differences of *Qualis* classification across areas we were forced to aggregate them in order to provide a single local measure of impact for each journal. We then took the averages of different journal ratings from each subject area. For example, for a journal receiving 100 points in one area and 60 in another, we considered the average of 80 points. Adopting this procedure is likely to avoid biasing toward a particular discipline. Overall, despite the shortcomings of *Qualis* and its conflicting role with the SJR, we consider this cost to be outweighed by the benefit of not letting 66 percent of articles published in domestic journals out of our analysis.

Across the different disciplines, high impact journals display more articles. As a result, a journal's expected impact and quantity of papers become positively correlated (Huang, 2016). However, easily quantifiable measures of publication quality—such as number of citations and quantity of published papers in journals of high impact—have been increasingly questioned (Seglen, 1997; Frey and Rosa, 2010; Beets et al., 2016; Genova et al., 2016). It is argued that quality should be assessed by the value of new insights accrued from a published paper and, thus, additional qualitative measures are urged to be used along the conventional measures above.

Because this means the mediation of some sort of peer review evaluation, we note this is exactly what a classification method such as *Qualis* can possibly deliver. Unfortunately, it is very difficult to track quality as a multidimensional attribute for a massive database such as ours. For this reason, we limit our analysis to impact factors.

2.3 *Preliminary descriptive statistics*

We downloaded a list of addresses from *Lattes* for every academic with doctorates in all the fields of study, as shown in Table 1. The list ended up with 180,594 researchers and their 3,212,490 publications. Thus, Brazilian researchers published 18 papers on average. We removed from the list those articles without an ISSN, and those published either before 1990 or later than 2014. As observed, the ISSN information is critical for finding a journal's SJR and average *Qualis* ranking. We also removed from the list those "researchers" with no publication as well as those who completed their PhDs after 2014.

There is also no explicit information related to a researcher's gender in *Lattes*, whether male or female. Thus, to identify a gender we computed the given names of the researchers as they appear in *Lattes* and then contrasted them with a list of 35,855 Brazilian names from a website that explicitly relates given names to gender (http://www.listadenomes.com.br). After this procedure, we could associate 93.49 percent of all researchers in *Lattes* with their gender.

From *Lattes*, we then identified those researchers holding a doctorate along with the higher education institutions from which they received their PhDs, whether domestic or foreign. This piece of information may matter for researcher productivity (Roos et al., 2014). Here, our software searched for cues to identify words related to foreign countries and graduate education institutions, because such data were not directly available in any particular locations of *Lattes*.

After we have employed the filters above to detect gender as well as PhD origin, ambiguous data points were dropped from the raw sample and we ended up with 111,736 researchers (and 2,574,082 publications). As can be seen, the filtered sample is still huge and much larger than any other used so far, such as the study of Rorstad and Aksnes (2015) for Norwegian researchers, who consider only 12,403 researchers (and 35,798 publications). Table 2 shows the descriptive statistics of our filtered sample.

To measure publication productivity, we consider only scholarly articles published in journals with an ISSN. Books, book chapters, conference papers, and working papers were left out. We are aware this procedure is far from perfect as one considers the particularities of subject areas, such as differences in methodology, epistemology, and outlet preference (Shin and Cummings, 2010). Researchers from the natural and medical areas tend to target prestigious journal outlets, while researchers from other areas tend to prefer to get published through books rather than scholarly journal outlets (Bourke and Butler, 1996; Shin and Cummings, 2010; Piro et al., 2013). While it is quite feasible to assess the potential impact of scholarly articles based on the SJR for large datasets, it is very difficult to carry out the same sort of evaluation for books and book chapters. In our data, most books and book chapters are published domestically, a contingency that makes it almost impossible to evaluate their impact. This difficulty is not found in the previous literature where the quality of books and book chapters are usually assessed for short datasets. In such a situation, it is easier to perform a multidimensional evaluation of quality even if bibliometric data is lacking (Bourke and Butler, 1996; Dundar and Lewis, 1998; Seglen and Aksnes, 2000; Shin and Cummings, 2010, Piro et al., 2010).

Table 2 shows the average number of publications is not homogenous across the subject-area groups, a result that replicates the findings from previous work (Kyvik, 2003; Piro et al., 2010). For example, while the figure amounts to 30 in the health sciences, it is only 17 for engineering. This fact partially reflects the standards of scholarly publishing adopted by the distinct subject-area groups. There are particularities in each discipline that affect the quantity of publications delivered by a researcher. Papers with many co-authors are common in some disciplines but not in others. Furthermore, papers in some areas are usually too short if compared with the usual paper length in other areas.

Table 1. Labels for the subject areas according to the *Lattes* platform.

Lattes label for subject-area groups	*Selected subject areas*
Agricultural sciences	agricultural engineering, agronomy, fishing engineering, food technology, forest engineering, veterinary, zootechny
Applied social sciences	architecture and urbanism, business administration, communication, domestic economics, economics, industrial design, information science, law, museology, social services, tourism, urban and regional studies
Biological sciences	biochemistry, biology, biophysics, biotechnology, botany, ecology, genetics, immunology, morphology, parasitology, pharmacology, physiology, zoology
Engineering	aerospace engineering, biomedical engineering, chemical engineering, electrical engineering, energy engineering, mechanical engineering, metallurgical engineering, mining engineering, naval engineering, nuclear engineering, production engineering, sanitary engineering, traffic engineering
Exact and earth sciences	astronomy, chemistry, computer science, geoscience, mathematics, oceanography, physics, statistics
Health sciences	dentistry, medicine, nursing, nutrition, occupational therapy, pharmaceutical science, phono audiology, physical education, physiotherapy
Humanities	anthropology, archeology, education, geography, history, philosophy, political science, psychology, sociology, theology
Linguistics and arts	arts, linguistics
Others	environmental science

As for gender, Table 2 shows there are more male researchers (55 percent) than female ones after considering the average of all subject-area groups. Stronger male participation follows a general pattern already observed in literature (Rorstad and Aksnes, 2015). However, while engineering is predominantly a male subject-area group (74 percent), the health sciences are skewed toward females (46 percent of males). A more detailed analysis of the tails of the gender distribution reveals that male participation in electrical, mechanical and aerospace engineering is, respectively, 89, 88 and 83 percent. In contrast, females prevail in phono audiology, nursing and nutrition, with male participation at 8, 10 and 12 percent respectively.

Table 2 also shows only a minority of Brazilian researchers hold a PhD from a foreign institution. This is observed for all subject-area groups, though in some groups, such as the health sciences, the percentage is even lower. This may be partially explained by the availability of funding from foreign bodies as well as distinct patterns of academic competition across the areas. We present in the Appendix an expanded version of Tables 1 and 2 for the subject areas with at least 100 active researchers.

While both Table 2 and the table in the Appendix present a current snapshot of gender participation in research and PhD origin, Fig. 1 shows how these variables evolved since 1980. The left-hand side in Fig. 1 shows females are catching up over the years. Apart from engineering, exact, earth and health sciences, the proportion of male and female researchers reaches nearly 50 percent in 2014. The proportion of PhD graduates from abroad has been declining since 1980 (right-hand side in Fig. 1). This can be explained by the expansion of domestic doctoral programs. Since the nineties, the Brazilian funding bodies have been favoring the creation of domestic doctoral programs, drastically reducing grants for foreign PhD studies.

Table 2. Brazilian researchers and their publications: Descriptive statistics from the *Lattes* platform.

Lattes label for a subject-area group	*Number of researchers*	*Number of publications*	*Average number of publications*	*Male researchers, %*	*PhD from a foreign institution, %*
Agricultural sciences	15,703	405,527	25.82	59.54	12.16
Applied social sciences	5,286	102,230	19.34	61.43	23.31
Biological sciences	21,565	479,475	22.23	43.63	13.04
Engineering	11,419	191,094	16.73	74.17	22.13
Exact and earth sciences	19,883	403,130	20.28	67.71	21.44
Health sciences	25,619	775,930	30.29	46.11	7.01
Humanities	10,091	179,393	17.78	42.71	18.71
Linguistics and arts	1,720	30,176	17.54	40.23	25.23
Others	450	7,127	15.84	56.67	25.11
All subject-area groups	111,736	2,574,082	23.04	54.60	15.19

Whenever a journal's ISSN is not recorded in the SJR, we assign it a value of zero. This simple procedure allows us to keep a publication not listed on the SJR and, at the same time, to penalize it. Doing so, we are able to keep the same data points in both the quantity and impact analysis, avoiding a possible bias if the publications not listed on the SJR were removed from the sample. As a robustness check, we removed the publications not listed on the SJR from the sample, and the results did not change a great deal regarding the sign and statistical significance of the coefficients (available upon request).

Fig. 2 shows boxplots of the average SJR per researcher. Researcher productivity is very heterogeneous across the subject-area groups. On the left-hand side in Fig. 2, one can see that the group of the exact and earth sciences along with engineering and the biological sciences have larger median SJRs than the group of linguistics and arts, humanities and applied social sciences. For example, a researcher with an average SJR of 0.25 stands at the top of the distribution for humanities, but falls below average for engineering. However, the difference between the two distinct groups above may be caused by the existence of fewer journals with high impact for the second group. Therefore, this does not necessarily mean higher productivity for the first group. For this reason, the analysis of researcher productivity only makes sense after considering the disciplines in isolation, as pointed out by Rorstad and Aksnes (2015).

Fig. 2 (center) also shows the average number of points a researcher received from *Qualis* as well as boxplots of the average number of journal publications per year (right). We considered only the years after PhD completion for the latter. One can observe that subject-area groups with higher median number of publications per year are also the ones with lower median values for both average SJR and average *Qualis*. This strongly suggests a publication trade-off between expected impact and quantity.

The findings in Fig. 2 strongly suggest *Qualis* overrates the social sciences and humanities subject areas. The dispersion of *Qualis* across the areas is also much smaller if compared to the SJRs. The ratings are more related in the natural and medical areas, and present more inconsistencies in the others. *Qualis* seems to track better the differences within a subject area, however. In theory, *Qualis* can thus consider those qualitative aspects neglected by a pure citation analysis indicator, such as the SJR. One example of such a qualitative aspect that *Qualis* aims to track is reputation of editorial boards.

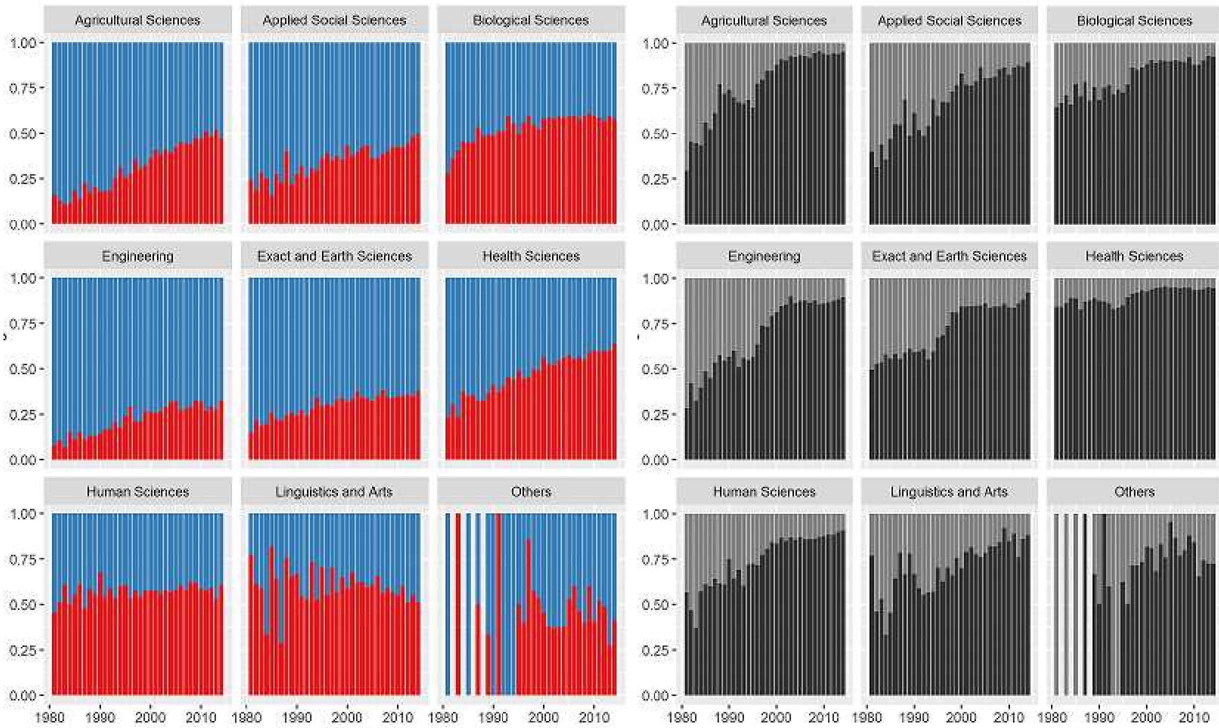

Fig. 1 The percentage of Brazilian researchers distributed by gender (left) and PhD origin (right), 1980-2014. Overall, female participation (red columns) is catching up toward the level of 50 percent, and domestic doctorates (black columns) are on the rise.

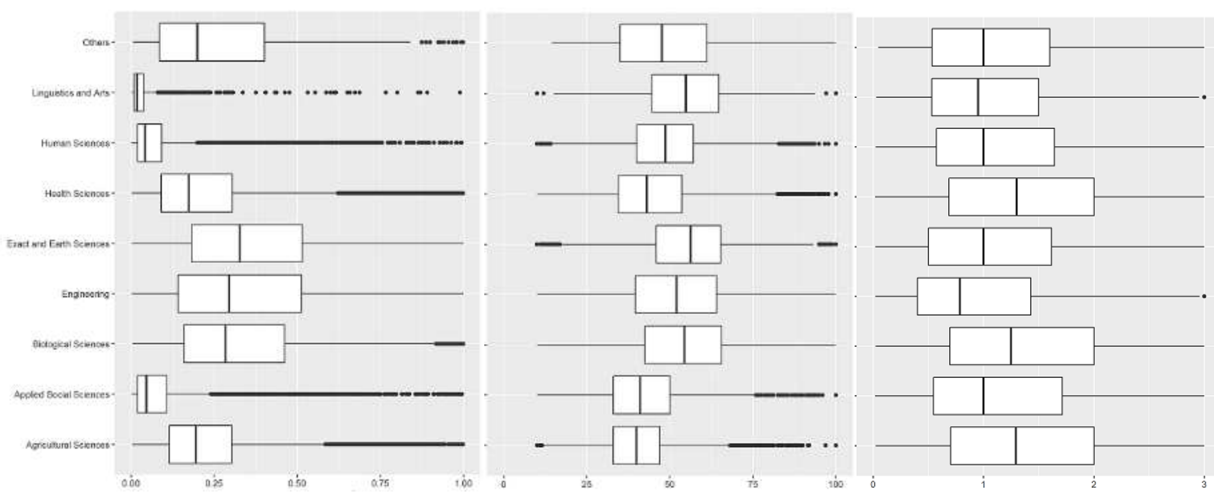

Fig. 2 Boxplots of the average SJR per researcher (left), average points from *Qualis* (center), and average number of journal publications per year after PhD completion (right).

Fig. 3 shows the time evolution of the percentage of publications in international journals listed on the SJR, the percentage of domestic journals listed on the SJR, and the proportion of articles that are not listed on the SJR. As can be seen, international reach is also heterogeneous across the subject-area groups. The proportion of publications in international journals is lower for the applied social sciences, humanities, and linguistics and arts. Moreover, this proportion is stable over the period. We see a different picture when looking at the proportions for the other subject-area groups. These present a more significant proportion of publications in the SJR, and this trend seems to increase over time. In particular, the biological sciences show exponential growth of publications in journals listed on the SJR. Thus, heterogeneity of areas justifies our methodological choice of analysis by considering publication quantity and impact for each subject-area group separately. Incidentally, such heterogeneity is an important characteristic any funding body should take into account.

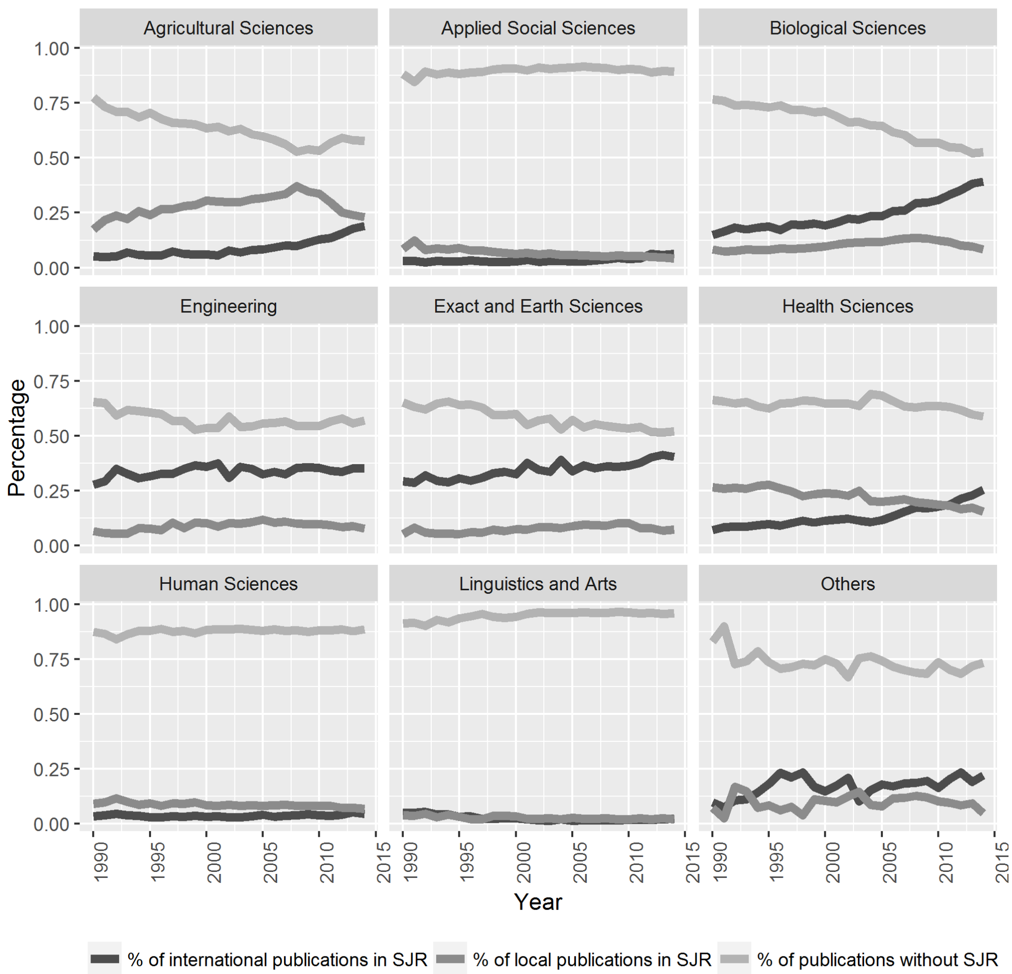

Fig. 3 Time evolution of the percentage of publications in international journals listed on the SJR, the percentage of domestic journals listed on the SJR, and the proportion of articles that are not listed on the SJR.

2.4 *Methodology for statistical assessment*

We estimated a generalized linear regression model (GLM) (Nelder and Bake, 1972) to test the explanatory power of the variables discussed above. Following Laurance et al. (2013) and Bin et al. (2015), we considered a Gamma distribution GLM with a log-link function. This modeling approach is justified in our case given that we have three dependent variables that are positive and continuous. The econometric model is given by:

$$E(p_i \mid X) = \exp(X) \quad (1)$$

$$X = \alpha + \beta_1 g + \beta_2 f + \beta_3 t + \beta_4 y + \beta_5 y^2 \quad (2)$$

where p_i is a measure of research productivity for researcher i; g is a dummy variable for gender (male = 1; female = 0); f is a dummy for PhD origin (foreign = 1; domestic = 0); t is time taken in years for finishing PhD coursework and thesis; y is the number of years involved in research since PhD completion (year 2015 minus year of completion). The inclusion of gender and PhD origin has already been justified. The inclusion of variable t assumes that the years a researcher spends doing his or her PhD coursework and thesis can predict his or her future productivity. Variable y measures the research age of an individual and can also be a proxy for his or her age itself (Bonaccorsi and Daraio, 2003). Following Rorstad and Aksnes (2015), variable y is also considered as a squared, separate term in order to capture a possible inverted U-shaped function for researcher age. A researcher's productivity first is expected to increase soon after PhD completion, reach a peak and then decline (Rorstad and Aksnes, 2015). The parameters in Eq. 1 are estimated by maximum likelihood.

3. Results

3.1 *Quantiles*

We first compute the quantiles of two of the alternative metrics of researcher productivity considered in this work: average SJR and average number of points from *Qualis*. Then, we compute the quantiles for the third metrics: number of papers published per year since PhD completion. Researchers falling in the leftmost, first quantile are the least productive, whereas those falling in the rightmost, fifth quantile are the most productive. Fig. 4 shows gender participation in each quantile, whereas Fig. 5 displays the quantiles for PhD origin. Given the heterogeneity across the disciplines, all quantiles intervals were built within each field.

Fig. 4 suggests the relationship between researcher productivity and gender is mixed. However, a closer look at engineering, exact and earth sciences shows a predominance of males among the most productive researchers. In contrast, the proportion of females increases with the average SJR in humanities, a result at odds with previous findings in literature (Prpi, 2002). Moreover, the proportion of males is higher among those with the highest number of papers published per year since PhD completion, notably in the biological and health sciences.

Unlike the role of gender in researcher productivity, Fig. 5 suggests a clear pattern for the effect of holding a PhD from abroad on researcher productivity. This pattern emerges for all the subject-area groups. Fig. 5 (left, center) shows the proportion of researchers with a PhD from abroad is higher among the groups with highest SJR and *Qualis*. Moreover, as productivity rises across the quantiles, so does the proportion of researchers holding a PhD from abroad.

However, in striking contrast with the metrics of the SJR and *Qualis*, researchers with the highest number of publications hold domestic PhDs (right hand side in Fig. 5). This is particularly noticeable for the applied social sciences, linguistics and arts. This result posits a trade-off between research impact and quantity (number of publications). Researchers holding a PhD from domestic institutions tend to publish more papers, but in journals of lower impact.

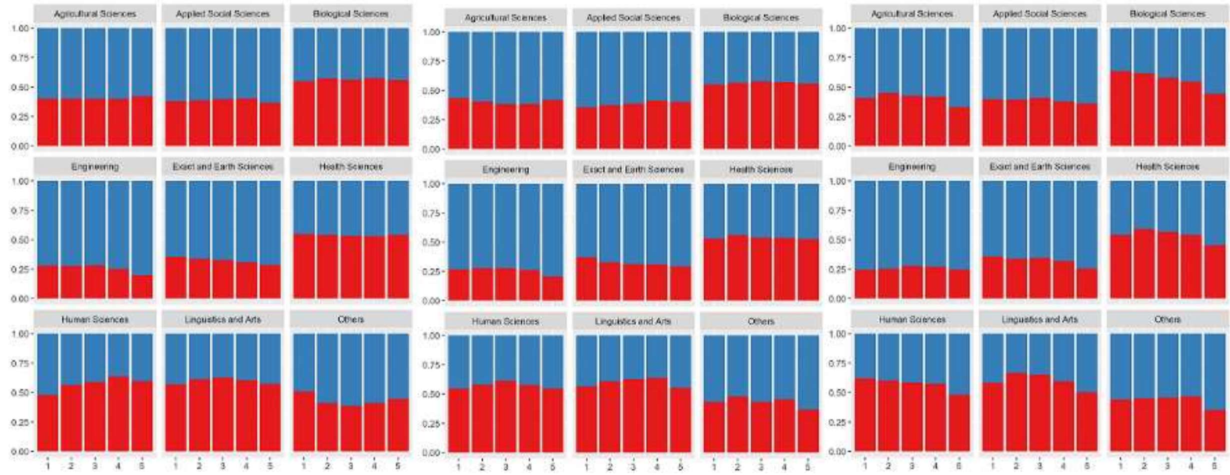

Fig 4 Metrics of researcher productivity and gender (red = female; blue = male): quantiles for average SJR (left), average *Qualis* (center), and number of publications per year since PhD completion (right).

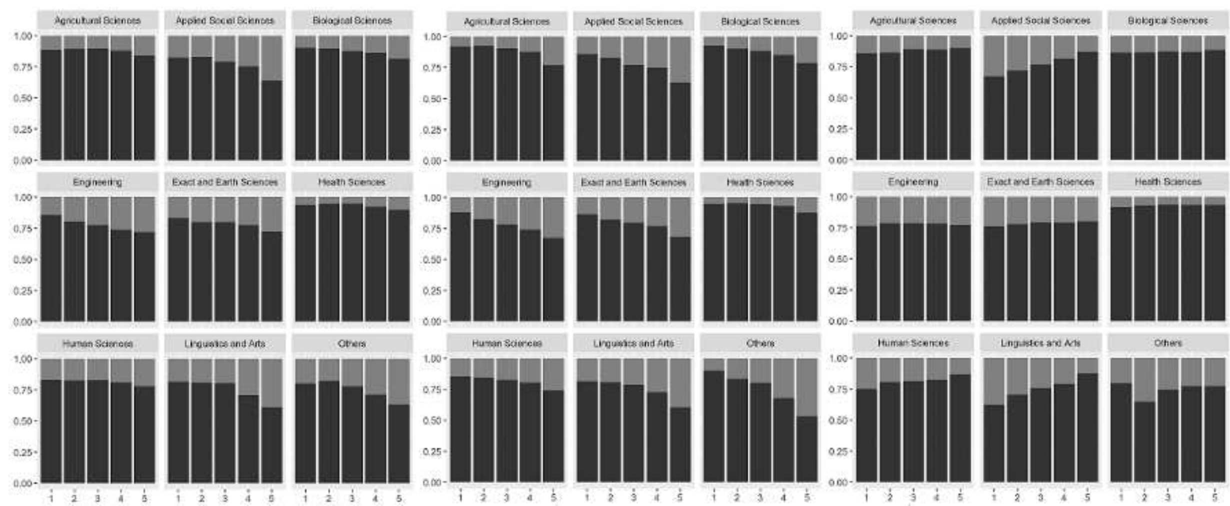

Fig 5 Metrics of researcher productivity and PhD origin (darker columns = domestic; lighter columns = foreign): quantiles for average SJR (left), average *Qualis* (center), and number of publications per year since PhD completion (right).

3.2 *Regressions*

We now present the estimation results of the GLM in Eq. (1). Table 3 shows the results considering the average SJR as the dependent variable, whereas Table 4 considers the average *Qualis* as the dependent variable.

Table 3 confirms our previous results in the graphical analysis of quantiles for gender (Fig. 4). There is a mixed effect of gender on the average SJR. Indeed, parameter β_1 is positive and statistically significant in four of the subject-area groups, but negative in five, with only one case of statistical significance (agricultural sciences). However, it emerges a positive correlation between the SJR and gender in the aggregate, which means males are more productive. Nevertheless, considering the findings for gender reported earlier, such a result is to be taken with caution.

The most robust result in Tables 3 and 4 refers to the impact of holding a PhD from a foreign institution on the average SJR as well as on the average *Qualis*. For all the subject-area groups, we find a positive and statistically significant linear relationship between these variables. The null hypothesis $\beta_2 = 0$ is rejected at the one percent level for all the cases. The implication is then straightforward: When controlling for other factors, holding a PhD from abroad can boost future performance in terms of the SJR. When looking at the values of the estimated parameters and comparing them to the unconditional value of the average SJR per researcher (the constant) and to the value of β_2, we can see the relative effect of holding a PhD from abroad is especially strong for the health sciences and linguistics and arts. We observe a relative increase of the average SJR of, respectively, 36 percent (0.79/2.18) and 38 percent (0.55/1.43). Interestingly, we also observe the time taken for finishing a PhD is statistically significant and positively related to the average SJR in eight out of the nine subject-area groups. Those who take more time to finish their PhD job tend to publish in higher impact journals. Moreover, the findings for the effect of research age of an individual (considering both y and y^2) on the average SJR do suggest the existence of an inverted U-shaped function. Thus, there is a point in time where a researcher reaches maximum performance.

Table 3. Estimation of Eq. (1) considering p_i as the average SJR (× 100).

	Agricultural sciences	*Applied social sciences*	*Biological sciences*	*Engineering*	*Exact and earth sciences*	*Health sciences*	*Humanities*	*Linguistics and arts*	*Others*
Intercept	3.76***	2.44***	3.07***	3.89***	3.09***	2.18***	3.86***	1.43***	3.66***
	(0.03)	(0.07)	(0.02)	(0.03)	(0.03)	(0.10)	(0.04)	(0.20)	(0.23)
g	-0.01	-0.05	0.03**	0.10***	-0.02**	0.14**	0.18***	-0.10	-0.03
	(0.01)	(0.04)	(0.01)	(0.01)	(0.01)	(0.06)	(0.02)	(0.11)	(0.11)
f	0.35***	0.36***	0.25***	0.25***	0.23***	0.79***	0.27***	0.55***	0.36***
	(0.02)	(0.05)	(0.02)	(0.02)	(0.02)	(0.07)	(0.02)	(0.12)	(0.12)
t	0.02***	0.03**	0.05***	0.01**	0.06***	0.09***	0.02***	0.10**	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.04)	(0.05)
y	-0.03***	-0.06***	-0.02***	-0.03***	-0.02***	-0.04***	-0.03***	-0.07***	-0.05***
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)	(0.02)
y^2	0.00***	0.00***	0.00***	0.00***	0.00***	0.00*	0.00***	0.00***	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
AIC	126983.98	36261.05	198319.70	111263.04	190146.11	212565.70	63529.62	8206.05	4088.13
BIC	127037.61	36307.06	198375.55	111314.44	190201.39	212622.76	63580.16	8244.20	4116.89
Log-likelihood	-63484.99	-18123.5	-99152.85	-55624.52	-95066.05	-106275.85	-31757.81	-4096.03	-2037.6
Deviance	8631.34	9475.84	14531.95	10658.56	14659.32	20685.28	13738.76	2497.95	518.53
Observations	15703	5286	21565	11419	19883	25619	10091	1720	450

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 5 shows the results considering the average number of papers published per year since PhD completion as the dependent variable. We find the effect of gender on the number of publications is positive and statistically significant. For all the subject-area groups, β_1 is greater than zero and significant at 5 percent, implying that male researchers tend to have more publications per year when compared to their female counterparts. However, in terms of size, such an impact is relatively small. The average value of α/β_1 across all the subject-area groups is 16 percent. This means a male researcher is likely to publish 16 percent more papers per year than a female within each subject-area group.

As for the effect of holding a PhD from abroad, there are no clear-cut results for β_2 in Table 5. Only two subject-area groups show negative and significant values for β_2. Overall, holding a PhD from abroad is not related to a researcher's quantity of papers published. Of note, β_3 is usually negative and statistically significant for all the subject-area groups. Thus, the time taken for doing a PhD is negatively associated with the number of publications. The more years a researcher takes to finish his or her doctorate, the more likely for him or her to publish less in the subsequent years. Moreover, β_4 is negative and β_5 is positive in Table 5. This suggests the existence of a nonlinear function relating researcher age and number of papers published, a result already found by Rorstad and Aksnes (2015).

Table 4. Estimation of Eq. (1) considering p_i as the average *Qualis*.

	Agricultural sciences	*Applied social sciences*	*Biological sciences*	*Engineering*	*Exact and earth sciences*	*Health sciences*	*Humanities*	*Linguistics and arts*	*Others*
Intercept	4.00***	3.81***	3.73***	4.02***	3.62***	3.66***	3.94***	3.90***	3.74***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.03)	(0.07)
g	-0.01***	-0.01	0.01	0.02***	-0.01	-0.03***	0.02***	0.01	-0.00
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.03)
f	0.13***	0.05***	0.12***	0.10***	0.14***	0.15***	0.14***	0.07***	0.23***
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)
t	0.00*	0.02***	0.02***	-0.00	0.02***	0.01***	0.01**	0.00	0.04**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
y	-0.00***	-0.00	-0.00***	-0.00***	-0.00***	0.00	-0.01***	0.01***	-0.01*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
y^2	0.00***	0.00***	0.00***	0.00***	0.00***	0.00	0.00***	-0.00*	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
AIC	126983.98	36261.5	198319.7	111263.04	190146.11	212565.70	63529.62	8206.05	4088.3
BIC	127037.61	36307.06	198375.5	111314.44	190201.39	212622.76	63580.16	8244.20	4116.8
Log-likelihood	-63484.99	-18123.5	-99152.8	-55624.52	-95066.05	-106275.8	-31757.81	-4096.03	-2037.0
Deviance	8631.34	9475.84	14531.95	10658.56	14659.32	20685.28	13738.76	2497.95	518.53
Observations	15703	5286	21565	11419	19883	25619	10091	1720	450

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

3.3 *Discussion*

Taken together, the results in Tables 3, 4 and 5 allow one to draw some implications. We replicate the finding that male researchers outperform their female counterparts in quantity of papers published (Abramo et al., 2015; Rorstad and Aksnes, 2015). However, the role gender plays on research is not that clear-cut. While males surpass females in terms of quantity (number of publications), research impact (as measured by the SJR and *Qualis*) is not significantly affected by a researcher's gender. After connecting the latter finding with the fact that female participation is catching up (Fig. 1), one may speculate that the next generation of females will take the lead in scholarly publishing in Brazil (van Arensbergen et al., 2012).

Table 5. Estimation of Eq. (1) considering p_i as the average number of papers published per year since PhD completion.

	Agricultural sciences	*Applied social sciences*	*Biological sciences*	*Engineering*	*Exact and earth sciences*	*Health sciences*	*Humanities*	*Linguistics and arts*	*Others*
Intercept	1.29***	1.16***	1.60***	0.90***	1.56***	1.52***	0.91***	1.12***	1.23***
	(0.03)	(0.03)	(0.02)	(0.04)	(0.03)	(0.05)	(0.05)	(0.08)	(0.21)
g	0.28***	0.22***	0.25***	0.20***	0.22***	0.14***	0.05**	0.21***	0.22**
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)	(0.10)
f	0.01	-0.03	0.02	0.01	0.01	-0.16***	0.04	-0.27***	0.01
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.05)	(0.11)
t	-0.11***	-0.08***	-0.09***	-0.08***	-0.12***	-0.10***	-0.10***	-0.05***	-0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.05)
y	-0.03***	-0.05***	-0.04***	-0.03***	-0.04***	-0.06***	-0.02***	-0.06***	-0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)
y2	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
AIC	126983.98	36261.0	198319.70	111263.04	190146.11	212565.70	63529.62	8206.05	4088.13
BIC	127037.61	36307.0	198375.55	111314.44	190201.39	212622.76	63580.16	8244.20	4116.89
Log-likelihood	-63484.99	-18123	-99152.85	-55624.52	-95066.05	-106275.5	-31757.81	-4096.03	-2037.0
Deviation	8631.34	9475.84	14531.95	10658.56	14659.32	20685.28	13738.76	2497.95	518.53
Observations	15703	5286	21565	11419	19883	25619	10091	1720	450

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Here, we also replicate the finding of Roos et al. (2014) in a larger scale. Researchers holding a PhD from a foreign institution tend to publish in journals with higher SJR than those with domestic PhDs. However, in contrast with Roos et al. (2014), our results do not allow one to infer for all the subject areas that, on average, researchers holding PhDs from abroad publish less frequently.

One possible explanation for the fact that researchers with a foreign PhD publish in journals of higher impact is that Brazilian doctoral programs still do not rival the primary institutions of the United States and Europe. In such higher education institutions, there is an established research environment, top thesis advisors, superior research support in financial terms and facilities, and closer contact with research leaders. Add to these the fact that debut papers come in co-authorship with supervisors, a fact that certainly provides a rewarding start for the novice researcher.

An important policy implication from our results is that Brazilian research will gain more impact as potential researchers continue to be sent to foreign institutions to earn their PhDs. Here, scholarships and domestic-foreign partnerships (such as the current "sandwich" mode doctorates) will continue to add value. Sandwich doctorates refer to a special program in which a doctoral candidate from a Brazilian university spends at least 6 months in a foreign institution in order to develop its research under the supervision of an experienced researcher in the field.

We also confirm the findings of Tuesta et al. (2015) in a much larger scale, concerning the years a researcher spends doing the PhD job. We find this variable can explain expected research impact. Those who take more time for doing their PhDs and, as a result, have a longer and closer relationship to their supervisors tend to publish in scholarly journals of higher impact. However, they also tend to publish relatively less. This result can inform university staff in the selection processes of novice researchers.

4. Conclusion

This study looks into the factors explaining scholarly publishing in Brazil. Its main contribution relates to the quality and reach of the data considered. We take advantage of a massive database—the *Lattes* platform—and assemble a sample of 180,000 CVs of researchers from most of the subject areas. We collect the particular information conveyed in *Lattes* regarding journal destination of the Brazilian scholarly publishing. Researcher publication productivity is evaluated through three metrics related to published articles: average SCImago Journal Rank, average points from *Qualis* (which provides coverage of Brazilian domestic journals), and average number of journal publications per year after a researcher's PhD completion. The effect on productivity is evaluated for selected characteristics of a researcher: gender, PhD origin (whether domestic or foreign), time taken for doing the PhD job, and number of years doing research since PhD completion.

As for the effect of gender on research impact, we find that male researchers publish in journals with higher impact in the subject-area groups of engineering and exact and earth sciences. However, females have more impact in the humanities. Moreover, the proportion of males is higher among those with the highest number of papers published per year since PhD completion, notably in the subject areas from the biological and health sciences. In the aggregate, males are more productive, but the effect of gender on the metrics of researcher productivity is mixed as one considers individual groups of subject areas. Gender matters for the number of publications, but there is no robust correlation between gender and the average SJR. Thus, though male researchers publish more on average than females do, the impact of male work is not significantly higher than that of females. As for the participation of females in research, a time series perspective of the data since 1980 shows females are catching up over the years. Apart from engineering, and the exact, earth and health sciences, the proportion of female researchers reached nearly 50 percent in 2014.

For all the subject-area groups, we find an unambiguous effect of holding a PhD from a foreign institution on a researcher's impact. The proportion of researchers with a PhD from abroad is higher among the groups with the highest average SJR and *Qualis*. Holding a PhD from abroad increases the chances for a researcher to publish in journals of higher impact, while at the same time it does not significantly affect publication rate. Indeed, researchers holding domestic PhDs tend to publish more papers, but in journals of lower impact. Therefore, it emerges a trade-off between expected research impact (measured by the SJR and *Qualis*) and quantity (number of publications). Moreover, the proportion of PhD graduates from abroad has been declining since 1980, a fact that can be explained by the expansion of domestic PhD programs in Brazil, rather than by the reduction in absolute terms of PhD graduates from abroad. An important public policy implication from these results is that Brazilian research will gain more impact as potential researchers continue to be sent to foreign institutions for doing their PhD job.

We also find the years a researcher spends doing his or her PhD coursework and thesis can predict his or her future productivity. The time taken for doing a PhD is negatively related to the number of publications. The more years a researcher takes to finish his or her doctorate, the more likely he or she will publish less in the subsequent years. We explain this effect as a result of the longer and closer relationship between student and PhD supervisor.

We also replicate the literature finding of a nonlinear, inverted U-shaped function relating research age and number of papers published. The number of years involved in research since PhD completion is also related to the average SJR. For these metrics of productivity, researcher performance first increases soon after PhD completion, reaches a peak, and then declines.

Finally, it is important to point out some caveats. We use the SJR as a measure of expected impact of a researcher's work. While convenient for dealing with a large sample such as ours, such a choice is not free of criticisms since it assumes a paper impact will follow the historical impact of the journal where it was published. A more robust strategy would be to use the current citation records of the publication, which were not considered in this study. Moreover, our study considers only scholarly articles published in journals. Books, book chapters, conference papers, and working papers are left out, and this omission is justified by the fact that assessing the quality of such output using a large dataset such as ours is unfeasible. Our study, thus, cannot fully appreciate the factors explaining researcher impact because for some subject-area groups, books and book chapters may be arguably more relevant for gauging productivity than scholarly articles. This omission certainly deserves further scrutiny by future research.

References

Abramo G., Cicero T, Dangelo CA (2015) Should the research performance of scientists be distinguished by gender? *Journal of Informetrics 9*(1), 25-38.

Almeida EC, Guimarães JA (2013) Brazil's growing production of scientific articles: How are we doing with review articles and other qualitative indicators? *Scientometrics 97*(2), 287-315.

Analytics R, Weston S (2014) doParallel: For each parallel adaptor for the parallel package. URL http://CRAN.R-project.org/package=doParallel, r package version 1.0.8.

Barrios M, Villarroya A, Borrego A (2013) Scientific production in psychology: A gender analysis. *Scientometrics 95*(1), 15-23.

Beets SD, Lewis BR, Browe HH (2016) The quality of business ethics journals: An assessment based on application. *Business & Society 55*(2), 188-213.

Bin A, Salles-Filho S, Capanema LM, Colugnati FAB (2015) What difference does it make? Impact of peer-reviewed scholarships on scientific production. *Scientometrics 102*(2), 1167-1188.

Bollen J, Van de Sompel H, Hagberg A, Chute R (2009) A principal component analysis of 39 scientific impact measures. *PLoS One 4*(6), e6022.

Bonaccorsi A, Daraio C (2003) Age effects in scientific productivity. *Scientometrics 58*(1), 49-90.

Bordons M, Fernández MT, Gómez I (2002) Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics 53*(2), 195-206.

Bourke P, Butler L (1996) Publication types, citation rates and evaluation. *Scientometrics 37*(3), 437-494.

Charyton C, Elliott JO, Rahman MA, Woodard JL, DeDios S (2011) Gender and science: Women Nobel laureates. *The Journal of Creative Behavior 45*(3), 203-214.

Chung HK, Cox RAK (1990) Patterns of productivity in the finance literature: A study of the bibliometric distributions. *Journal of Finance 45*(1), 301-309.

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

Da Silva S (2009) Going parochial in the assessment of the Brazilian economics research output. *Economics Bulletin 29*(4), 2826-2846.

Dahl DB (2014) xtable: Export tables to LaTeX or HTML. URL http://CRAN.R-project.org/package=xtable, r package version 1.7-4.

Diniz-Filho JA, Fioravanti MCS, Bini LM, Rangel TF (2016) Drivers of academic performance in a Brazilian university under a government-restructuring program. *Journal of Informetrics 10* (1), 151-161.

Dowle M, Srinivasan A, Short T, with contributions from R Saporta SL, Antonyan E (2015) data.table: Extension of Data.frame. URL http://CRAN.R-project.org/package=data.table, r package version 1.9.6.

Dundar H, Lewis, DR (1998) Determinants of research productivity in higher education. *Research in Higher Education 39*(6), 607-631.

Falagas ME, Kouranos VD, Arencibia-Jorge R, Karageorgopoulos DE (2008) Comparison of SCImago journal rank indicator with journal impact factor. *FASEB Journal 22*(8), 2623-2628.

Frey BS., Rost K (2010) Do rankings reflect research quality? *Journal of Applied Economics 13*(1), 1-38.

Genova G, Astudillo H, Fraga A (2016) The scientometric bubble considered harmful. *Science and Engineering Ethics 22*(1), 227-235.

Guimarães B (2012) Qualis as a measuring stick for research output in economics. *Brazilian Review of Econometrics 31*(1), 3-18.

Gonzalez-Pereira B, Guerrero-Bote V, Moya-Anegon F (2010) A new approach to the metric of journals scientific prestige: The SJR indicator. *Journal of Informetrics 4*(3), 379-391.

Hall CM (2011) Publish and perish? Bibliometric analysis, journal ranking and the assessment of research quality in tourism. *Tourism Management 32*(1), 16-27.

Huang DW (2016) Positive correlation between quality and quantity in academic journals. *Journal of Informetrics 10*(2), 329-335.

Isfandyari-Moghaddam A, Hasanzadeh M, Ghayoori Z (2012) A study of factors affecting research productivity of Iranian women in ISI. *Scientometrics 91*(1), 159-172.

Kyvik S (2003) Changing trends in publishing behaviour among university faculty, 1980-2000. *Scientometrics 58*(1), 35-48.

Lane J (2010) Let's make science metrics more scientific. *Nature 464*(7288), 488-489.

Larivire V, Vignola-Gagn E, Villeneuve C, Glinas P, Gingras Y (2011) Sex differences in research funding, productivity and impact: An analysis of Quebec university professors. *Scientometrics 87*(3), 483-498.

Laurance WF, Useche DC, Laurance SG, Bradshaw CJ (2013) Predicting publication success for biologists. *BioScience 63*(10), 817-823.

Leifeld P (2013) texreg: Conversion of statistical model output in R to LATEX and HTML tables. *Journal of Statistical Software 55*(8), 1-24.

Leite P, Mugnaini R, Leta J (2011) A new indicator for international visibility: Exploring Brazilian scientific community. *Scientometrics 88*(1), 311-319.

Leta J, Lewison G (2003) The contribution of women in Brazilian science: A case study in astronomy, immunology and oceanography. *Scientometrics 57*(3), 339-353.

Lotka AJ (1926) The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*(12), 317-324.

Moed HF (2005) Citation analysis in research evaluation. Springer, Dordrecht.

Morganson VJ, Jones MP, Major DA (2010) Understanding women's underrepresentation in science, technology, engineering, and mathematics: The role of social coping. *The Career Development Quarterly 59*(2), 169-179.

Nelder JA, Baker RJ (1972) Generalized linear models. *Encyclopedia of Statistical Sciences*.

Piro F, Aksnes DW, Rørstad K (2013) A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology 64*(2), 307-320.

Prpi K (2002) Gender and productivity differentials in science. *Scientometrics 55*(1), 27-58.

Puuska HM (2009) Effects of scholars gender and professional position on publishing productivity in different publication types: Analysis of a Finnish university. *Scientometrics 82*(2), 419-437.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Roos DH, Calabró L, De Jesus SL, Souza DO, Barbosa NV, Da Rocha JBT (2014) Brazilian scientific production in areas of biological sciences: A comparative study on the modalities of full doctorate in Brazil or abroad. *Scientometrics 98*(1), 415-427.

Rorstad K, Aksnes DW (2015) Publication rate expressed by age, gender and academic position: A large-scale analysis of Norwegian academic staff. *Journal of Informetrics 9*(2), 317-333.

Ruiz-Castillo J, Costas R (2014) The skewness of scientific productivity. *Journal of Informetrics 8*(4), 917-934.

Seglen PO (1997) Why the impact factor of journals should not be used for evaluating research. *British Medical Journal 314*(7079), 498-502.

Seglen PO, Aksnes DW (2000) Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research. *Scientometrics 49*(1), 125-143.

Shin JC, Cummings WK (2010) Multilevel analysis of academic publishing across disciplines: Research preference, collaboration, and time on research. *Scientometrics 85*(2), 581-594.

Tuesta EF, Delgado KV, Mugnaini R, Digiampietri LA, Mena-Chalco JP, Pérez-Alcázar JJ (2015) Analysis of an advisor–advisee relationship: An exploratory study of the area of exact and earth sciences in Brazil. *PloS One 10*(5), e0129065.

van Arensbergen P, van der Weijden I, van den Besselaar P (2012) Gender differences in scientific productivity: A persisting phenomenon? *Scientometrics 93*(3), 857-868.

Wickham H (2009) ggplot2: Elegant graphics for data analysis. Springer, New York, URL http://had.co.nz/ggplot2/book.

Wickham H, Francois R (2015) dplyr: A grammar of data manipulation. URL http://CRAN.R-project.org/package=dplyr, r package version 0.4.3.

Wickham H, James DA, Falcon S (2014) RSQLite: SQLite interface for R. URL http://CRAN.R-project.org/package=RSQLite, r package version 1.0.0.

Appendix. Descriptive data from the *Lattes* platform for the subject areas with at least 100 active researchers.

Lattes label for a subject-area group	*Subject area with at least 100 researchers*	*Number of researchers*	*Number of publications*	*Average number of publications*	*Male researchers, %*	*PhD from a foreign university, %*
Agricultural sciences	agronomy	6,222	160,235	25.75	65.93	12.79
Agricultural sciences	veterinary	3,209	89,456	27.88	50.79	10.53
Agricultural sciences	zootechny	2,251	59,468	26.42	62.37	11.46
Agricultural sciences	food science and technology	1,884	43,723	23.21	34.55	9.55
Agricultural sciences	forest engineering	958	23,101	24.11	69.21	17.43
Agricultural sciences	agricultural engineering	787	20,152	25.61	80.69	12.71
Agricultural sciences	fishing engineering	383	8,872	23.16	67.10	17.75
Applied social sciences	business administration	1,840	40,143	21.82	65.98	17.66
Applied social sciences	economics	1,608	28,010	17.42	74.88	30.04
Applied social sciences	information science	380	6,268	16.49	35.00	18.68
Applied social sciences	communication	354	7,983	22.55	48.02	18.93
Applied social sciences	law	336	7,909	23.54	58.33	24.70
Applied social sciences	architecture and urbanism	280	3,865	13.80	46.43	27.14
Applied social sciences	urban and regional studies	139	2,158	15.53	51.08	35.25
Applied social sciences	social services	102	1,451	14.23	12.75	17.65
Biological sciences	biochemistry	3,308	70,341	21.26	43.77	12.33
Biological sciences	genetics	2,728	61,425	22.52	41.35	13.31
Biological sciences	ecology	2,432	50,446	20.74	53.25	20.02
Biological sciences	microbiology	2,108	44,388	21.06	34.01	11.95
Biological sciences	zoology	1,733	41,455	23.92	57.99	17.14
Biological sciences	botany	1,701	34,660	20.38	40.56	12.35
Biological sciences	physiology	1,527	36,346	23.80	44.66	9.36
Biological sciences	morphology	1,244	28,890	23.22	43.49	6.75
Biological sciences	immunology	1,212	28,114	23.20	35.64	9.41
Biological sciences	pharmacology	1,206	35,667	29.57	41.04	8.21
Biological sciences	biology	970	13,606	14.03	36.70	17.11
Biological sciences	parasitology	930	23,822	25.62	37.63	12.04
Biological sciences	biophysics	439	9,905	22.56	59.45	16.63
Engineering	electrical engineering	2,445	31,333	12.82	88.83	23.15
Engineering	metallurgical engineering	1,665	40,022	24.04	67.93	19.58
Engineering	chemical engineering	1,617	27,873	17.24	54.79	16.08
Engineering	mechanical engineering	1,514	21,967	14.51	88.31	25.83
Engineering	civil engineering	1,319	18,120	13.74	73.31	22.97
Engineering	production engineering	950	17,918	18.86	72.32	17.68
Engineering	sanitary engineering	701	11,235	16.03	57.63	22.68
Engineering	biomedical engineering	367	8,869	24.17	71.39	29.97
Engineering	nuclear engineering	261	5,570	21.34	64.75	21.46
Engineering	aerospace engineering	217	3,391	15.63	82.95	35.48
Engineering	mining engineering	131	1,763	13.46	75.57	24.43
Engineering	traffic engineering	122	1,525	12.50	66.39	36.07
Exact and earth sciences	chemistry	6,324	135,232	21.38	52.06	13.55
Exact and earth sciences	physics	4,914	131,056	26.67	81.62	22.18
Exact and earth sciences	computer science	2,627	29,360	11.18	76.82	27.56
Exact and earth sciences	geoscience	2,523	46,961	18.61	67.90	24.53
Exact and earth sciences	mathematics	1,894	23,391	12.35	73.07	27.82
Exact and earth sciences	statistics	755	17,804	23.58	63.58	20.93
Exact and earth sciences	oceanography	519	10,559	20.34	60.31	37.76
Exact and earth sciences	astronomy	320	8,721	27.25	77.19	27.50
Health sciences	medicine	10,740	362,537	33.76	62.09	5.68
Health sciences	dentistry	4,161	148,683	35.73	49.46	4.76
Health sciences	public health studies	2,394	62,790	26.23	34.17	12.16
Health sciences	nursing	2,239	60,165	26.87	10.27	2.95
Health sciences	pharmaceutical science	2,053	42,254	20.58	35.22	10.18
Health sciences	physiotherapy	1,252	26,958	21.53	28.83	7.91
Health sciences	physical education	1,175	38,640	32.89	65.02	18.98
Health sciences	nutrition	1,155	22,773	19.72	12.38	6.41
Health sciences	phonoaudiology	422	9,930	23.53	7.82	5.69
Humanities	psychology	2,849	52,549	18.44	30.43	14.88
Humanities	education	2,843	50,287	17.69	33.35	14.39
Humanities	history	1,150	19,633	17.07	53.22	15.74
Humanities	sociology	955	16,048	16.80	52.98	25.76
Humanities	anthropology	714	11,701	16.39	43.84	26.61
Humanities	political science	619	10,308	16.65	65.75	27.14
Humanities	geography	430	7,802	18.14	59.53	21.63
Humanities	philosophy	409	8,161	19.95	80.20	31.54
Linguistics and arts	literature	756	14,662	19.39	37.43	18.92
Linguistics and arts	linguistics	555	10,826	19.51	32.79	25.95
Linguistics and arts	arts	409	4,688	11.46	55.50	35.94
Others	environmental sciences	219	2,934	13.40	48.86	18.72