Scientometrics (2015) 103:879–896
DOI 10.1007/s11192-015-1569-7

Assessing the profile of top Brazilian computer science researchers

Harlley Lima[1] · Thiago H. P. Silva[1] ·
Mirella M. Moro[1] · Rodrygo L. T. Santos[1] ·
Wagner Meira Jr[1] · Alberto H. F. Laender[1]

Received: 30 June 2014 / Published online: 26 March 2015

Abstract Quantitative and qualitative studies of scientific performance provide a measure of scientific productivity and represent a stimulus for improving research quality. Whatever the goal (e.g., hiring, firing, promoting or funding), such analyses may inform research agencies on directions for funding policies. In this article, we perform a data-driven assessment of the performance of top Brazilian computer science researchers considering three central dimensions: career length, number of students mentored, and volume of publications and citations. In addition, we analyze the researchers' publishing strategy, based upon their area of expertise and their focus on venues of different impact. Our findings demonstrate that it is necessary to go beyond counting publications to assess research quality and show the importance of considering the peculiarities of different areas of expertise while carrying out such an assessment.

Keywords Research performance · Scientific production · Bibliometry

✉ Mirella M. Moro
mmmoro@gmail.com

Harlley Lima
harlley@dcc.ufmg.br

Thiago H. P. Silva
thps@dcc.ufmg.br

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Wagner Meira Jr
meira@dcc.ufmg.br

Alberto H. F. Laender
laender@dcc.ufmg.br

[1] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Introduction

Quantifying the publication performance of individual researchers or groups of researchers provides a comparative measure of scientific productivity and represents a stimulus towards improving research quality. Such analyses become even more useful when the results aid governments and research agencies on new directions for research funding policies. Indeed, quantitative measures of individual performance are important for governments looking for greater efficacy in distributing resources (Lamont 2012) and universities looking for better evaluation in global rankings (Torrisi 2014). Furthermore, evaluating the productivity of researchers and nations over time may identify groups of excellence that lead their knowledge fields as well as groups that need to receive investment reinforcements (Abramo et al. 2011; Riikonen and Vihinen 2008). Overall, evaluation of scientific performance is now conducted in many nations around the world (Bosquet and Combes 2013; Delgado-Garcia et al. 2014; Ingwersen and Larsen 2014; Laender et al. 2008; Mamtora et al. 2014; Vanecek 2014).

Currently, this kind of research assessment usually depends on highly specialized committees formed by researchers' peers. For the individual perspective, committee decisions include not only firing and hiring, but also promoting, demoting, relocating and awarding researchers. Typically, such decisions involve a laborious process of manually assessing the academic curriculum of hundreds of researchers from potentially distinct fields, while aiming for a fair outcome. Here, we use one of such assessments as a case study for characterizing the performance of the top 406 Brazilian computer science (CS) researchers, ranked by the Brazilian National Council for Scientific and Technological Development (CNPq).[1] In particular, we analyze this specific group of researchers in terms of their career length, graduate student mentoring activity, and volume of publications and citations. In addition, we further analyze these researchers' publication outcome according to their areas of expertise and the impact of their targeted publication venues. The results of our thorough characterization outline the profile of the top Brazilian CS researchers and may serve as a starting point for contrasting their performance with the performance of their foreign peers. In addition, our assessment methodology may be easily adapted for other fields and other countries, and aid future manual research assessment.

In the remainder of this article, Section "Related work" provides background on related studies of scientific productivity. Section "Data acquisition and preparation" details the process of data acquisition and preparation for building a comprehensive academic profile for the top Brazilian CS researchers. Section "Global profile analysis of the top Brazilian CS researchers" presents a global analysis of the profile of these researchers in terms of career length, number of students mentored, and volume of publications and citations. Section "Impact-based stratified analysis" complements this analysis by assessing the impact of the researchers' publication outcome in light of their areas of expertise. Finally, Section "Concluding remarks" presents our conclusions.

Related work

Given the ever increasing publishing activity, several studies have analyzed the productivity of researchers for various purposes. For instance, in an analysis of 30,000 researchers working on different research areas, Kato and Ando (2013) concluded that international collaboration

[1] CNPq website: http://www.cnpq.br

improves the overall performance of researchers. Another perspective is to focus on specific countries, such as the Czech Republic (Vanecek 2014), Italy (Bonaccorsi and Daraio 2003), and Denmark (Ingwersen and Larsen 2014). Particularly, Bonaccorsi and Daraio (2003) analyze the impact of aging on the scientific production of researchers from institutions of the Italian National Research Council considering the publications of a single year (whereas we consider a spam of 10 years). It is also possible to focus on *one field* over *one country*, as studies on French economics (Bosquet and Combes 2013), Australian environmental sciences (Mamtora et al. 2014), and Spanish computer science (Ibáñez et al. 2013). Our research relates to all of those, as we focus in one field over one country: Brazilian computer science.

Regarding the Brazilian scenario, studies have characterized the growth of publications in general (Glänzel et al. 2006; Leta et al. 2006) and in specific scientific fields (Almeida and Guimarães 2013) as well as characterized the Brazilian coauthorship networks of researchers (Mena-Chalco et al. 2014). All the aforementioned works attested the undoubted growth of the Brazilian scientific production. Regarding individual fields of knowledge, Oliveira et al. (2012) evaluated the Brazilian researchers in clinical medicine. Similar to our work, such a study focused on analyzing the productivity of the top researchers in the area (according to the aforementioned ranking by CNPq). They concluded that a better instrument for defining qualitative and quantitative indicators is needed to identify researchers with outstanding scientific output.

Among a few studies on the productivity of CS groups in Brazil, Laender et al. (2008) focused on the research and education quality of top Brazilian CS graduate programs. Their results indicated the maturity of Brazilian CS graduate programs when compared to North American and European institutions. Menezes et al. (2009) corroborated such results by analyzing collaboration networks from three distinct geographical regions: Brazil, North America (Canada and US), and Europe (France, Great Britain and Switzerland). Wainer et al. (2009) compared the Brazilian CS scientific production to other countries. The results showed that Brazil has the largest production among Latin American countries, a production volume that is about one third of Spain's, one fourth of Italy's, and the same as India and Russia. Furthermore, a recent study that considers research networks in Latin America has corroborated such results (Delgado-Garcia et al. 2014). Regarding distinct research areas within CS, Wainer et al. (2013) analyzed the productivity by CS area. The results revealed that productivity and citation rates differ between some but not all CS areas and that there is no significant correlation between citation rates and the size of a CS research area.

In this article, we perform a thorough assessment of the profile of the top CS researchers in Brazil. In particular, we characterize these researchers' profile quantitatively according to dimensions such as their career length, their graduate student mentoring activity, and their publication outcome in terms of volume of publications and citations. In addition, we further analyze these researchers' publication profile by assessing the impact of their targeted venues relatively to other researchers from the same area of expertise. The results of our analysis may aid manual research assessments and serve as a basis for improving automatic researchers ranking systems, such as the one presented in our previous work (Lima et al. 2013).

Data acquisition and preparation

In this article, we start from a well defined set of researchers and evaluate their performance according to different criteria. To do so, we first created a comprehensive dataset with curated information about each researcher. In this section, we detail the four major

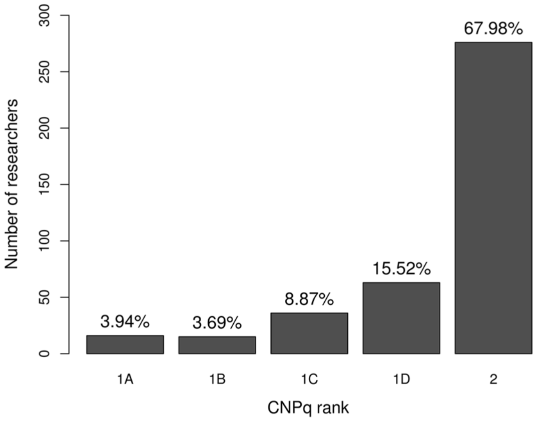


Fig. 1 Distribution of the top 406 Brazilian researchers on CS over the five CNPq ranks, according to the March 2013 classification

steps involved in building this dataset, which will be used later in the analyses presented in Sections "Global profile analysis of the top Brazilian CS researchers" and "Impact-based stratified analysis".

Step 1: Defining a target set of researchers. Currently, Brazil has around 2700 graduate programs in 46 fields, including anthropology, biology, business, computer science, engineering, medicine, and philosophy, among others. Considering all the potential researchers for this study, we decided to focus on those from the CS field who are part of the CNPq Research Fellowship Program, which provides individual grants to the top researchers in the country. All of these CNPq researchers are also professors in a graduate program in computer science or other related area (e.g., electric and systems engineering, mathematical computation and bioinformatics).

Such grants are awarded by 46 highly specialized committees that evaluate all candidates, and classify the selected ones into five ranks, in decreasing order of grant benefits: 1A, 1B, 1C, 1D and 2. To rank a candidate researcher, each committee assesses a research project proposed by the candidate and the candidate's profile in the recent years. Such a profile includes publications (journal articles, conference papers, and books); supervision of graduate students; contributions to science, technology and innovation (including patents); coordination of and participation in research projects; international insertion; participation as scientific reviewer; and other activities on scientific and academic management. In this work, we consider the 2012 edition of the program (the results of which were published in March 2013) that funded 406 CS researchers distributed over the five CNPq ranks as follows: 32 % of the researchers are in ranks 1A–1D and the remaining 68 % are in rank 2, as shown in Fig. 1. The entrance point of the program is rank 2, which concentrates most researchers since its grant value is smaller (about half of the grant awarded to ranks 1A–1D).

Step 2: Collecting the researchers' profile. In Brazil, all researchers must inform their achievements through the Lattes Platform,[2] one of the most prominent initiatives for archiving researchers' academic activities (Lane 2010). This platform provides a common interface for publishing the researcher's curriculum vitae (CV). Starting from the list of 406 CS researchers, we built their profile by extracting data from their Lattes CV, which includes their education background, professional activities, scientific production, student supervision activity, awards, and participation in events.

[2] Lattes Plataform: http://lattes.cnpq.br.

Step 3: Cleaning the researchers' profile. The data collected from Lattes must also be cleaned, in particular publication data. Specifically, the publications of each author may not be properly grouped due to name ambiguity: the same author may appear with distinct names (synonyms), or distinct authors may have similar names (polysems) (Ferreira et al. 2012), thus causing split and mixed citations (Lee et al. 2007). For dealing with the ambiguity among author names in the dataset, we use a state-of-the-art method for name disambiguation proposed by Cota et al. (2010). This method is based on the similarity of citation information such as title and venue name, and considers the authors' coauthorship network as well.

Step 4: Adding research area information. As we will discuss in the next sections, one very important piece of information for correctly assessing researchers' scientific production is their research area. In practice, we infer the research area of a researcher based upon the classification of each of the researcher's publications. This classification step is further divided into two: one for publications from conferences and another for publications from journals, as explained next. It is worth noting that considering only publications in journals is not enough for CS, as computer scientists tend to disseminate their innovative results by publishing primarily in conferences (Laender et al. 2008).

The Lattes platform defines research area at a coarse granularity and does not cover all CS areas. Therefore, we consider other sources of information. For conference papers, we used data from the SHINE project (Simple H-INdex Estimator),[3] which covers a list of venues provided by the Brazilian Computing Society's (SBC) Special Interest Groups (SIGs). There are 23 SBC SIGs focusing on different CS areas, such as computer architecture, databases, theory and so on. Specifically, we considered the 2011 version of the SHINE dataset, which contains more than 800,000 publications from approximately 1800 conferences, and 7.5 million citations, as collected in the beginning of 2011. Finally, each conference is associated with at least one area, according to the SIG that suggested it. For journal articles, we have considered data from the CAPES Qualis initiative.[4] Qualis ranks all journals in which Brazilian researchers have published in the past three years. Because of such a time frame limitation, we have also considered the list of journal titles from DBLP,[5] a computer science bibliography website. As opposed to the conference dataset, this one has no information regarding the area of each journal. Therefore, for each article, we performed a multi-label classification using LAC (Veloso et al. 2007), a lazy classification algorithm based on association rule mining. Overall, we have considered publications from 2001 to 2011 that were classified in the aforementioned 23 CS areas. Table 1 shows the number of conferences and journals for each area considered in our study, as well as summary statistics regarding the number of publications and citations per area.

Global profile analysis of the top Brazilian CS researchers

As pointed out by recent studies, the productivity of a researcher should be assessed in the perspective of his or her career background. Specifically, two relevant variables are important and relatively easy to collect: career length (Sugimoto and Cronin 2012) and students mentored (Kutlar et al. 2013; Miller et al. 2013; Torrisi 2014). Most of the time

[3] Simple H-INdex Estimator: http://shine.icomp.ufam.edu.br.

[4] CAPES Qualis Initiative: http://qualis.capes.gov.br.

[5] DBLP: http://www.informatik.uni-trier.de/~ley/db.

Table 1 Dataset information including number of conferences and journals and summary statistics (average, standard deviation, and median) for publication volume and citations of the top Brazilian researchers distributed over 23 CS areas from 2001 to 2011

Research Area	#Conf.	#Jour.	Volume			Citations		
			Avg.	SD.	Med.	Avg.	SD.	Med.
Algorithms and theory	354	188	10.65	9.97	8	95.98	151.95	41
Artificial intelligence	264	163	9.82	13.24	5	72.56	119.90	26
Collaboration systems	10	14	8.25	16.02	2	49.00	101.16	6
Computational biology	25	28	2.55	2.53	2	14.73	19.17	7
Computer graphics[a]	108	105	9.95	10.97	5	58.34	115.63	13
Computer networks[b]	297	161	13.46	18.77	6	84.95	177.22	29
Computer science education	35	37	3.25	5.04	1	10.49	23.45	1
Databases	184	127	8.55	13.99	4	73.30	182.51	12
Fault tolerant systems	32	7	2.45	3.02	1	23.35	54.80	5
Formalism	49	68	2.67	4.13	1	17.51	32.93	4
Game and entertainment	17	6	2.37	3.06	1	13.97	56.92	0
Geoinformatics	14	11	4.59	7.92	2	20.95	42.42	2
Hardware and architecture[c]	124	112	5.61	12.36	2	34.93	122.71	5
Health informatics	25	67	3.56	5.83	2	13.70	34.21	3
Human–computer interaction	21	31	2.71	3.19	1	17.48	26.15	6
Information systems	487	188	22.02	19.15	17	160.01	230.76	75.5
Music computing	15	6	2.59	3.28	2	4.21	8.21	0
Natural language processing	59	43	4.26	4.41	2	37.80	95.34	7
Neural networks	84	82	7.83	12.10	4	40.52	77.37	14
Programming languages	56	23	3.81	4.46	2	47.23	104.88	8
Robotics	56	63	3.24	3.80	2	29.71	70.27	4
Security	100	98	10.05	12.81	4	31.61	142.57	3
Software engineering	95	42	7.72	12.71	3	57.98	144.89	13

[a] Computer graphics and image processing

[b] Computer networks and distributed systems

[c] Hardware, architecture and embedded systems

(in both academia and industry), a researcher's professional career starts after acquiring a Ph.D. degree. As time goes by, the researcher establishes collaborations that go beyond working with his/her Ph.D. advisor. Likewise, the number of publications and their citations increase (Ibáñez et al. 2013). Furthermore, in academia, besides collaborating with others, a researcher may also count on the students he/she advises. Usually, the more students a researcher advises, the more work is published. Although previous studies focus on PhD students (Kutlar et al. 2013), here we consider both Ph.D. and Master's students because only one third of the Brazilian CS graduate programs grant PhD degrees.

Having created a dataset with relevant information about the top Brazilian CS researchers, we now present a thorough characterization of their profile. In particular, we aim to answer the following research questions:

Q1: How are researchers in each CNPq rank characterized in terms of career lenght?
Q2: How are researchers in each CNPq rank characterized in terms of the number of students mentored, in both Master's and PhD levels?
Q3: How are researchers in each CNPq rank characterized in terms of volume of publications and their citation counts?

We address each of these questions in the remainder of this section.

Career length

To address question Q1, we consider the year in which each researcher finished his/her Ph.D. Given the researchers grouped by their CNPq rank, Figure 2 shows the distribution of time since the researchers' received their doctorate degree. Note that the median values decrease from rank 1A towards rank 2. Hence, the more experienced a researcher is (here, experience measured from the time elapsed since their Ph.D.), the higher the rank he/she belongs to. Notice that there are also seven outliers in rank 2, which shows that the career length alone is not enough to reach higher ranks.

There are also large intervals between the medians of 1B to 1A (6.5 years) and 2 to 1D (7 years), whereas the intervals between the medians of rank 1C to 1B (2.5 years) and rank 1D to 1C (1.5 years) are smaller. These larger intervals suggest that it may be harder to be promoted from rank 2 to rank 1D and from rank 1B to 1A. Upon further analysis, these obstacles may be partially explained by the rules of CNPq to promote or demote a researcher. For instance, the career length of a researcher must be at least eight years to be promoted to rank 1D. Also, to be ranked at 1A, the researcher must have clear contributions to the national and international research communities, which generally takes longer time and effort to achieve.

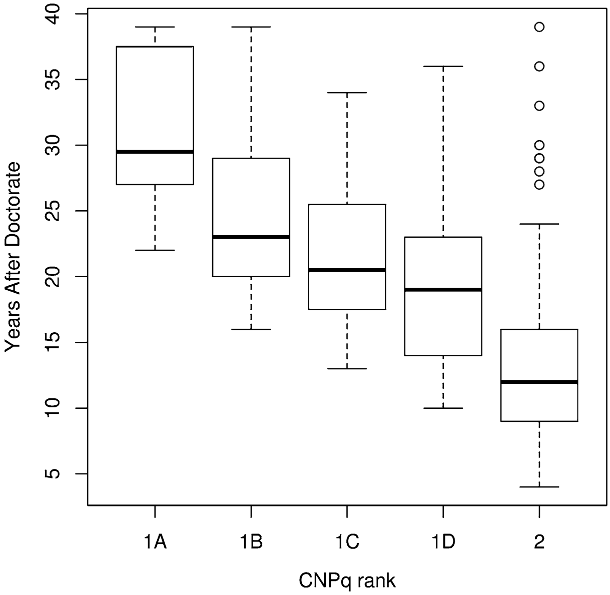


Fig. 2 Distribution of years since receiving the doctorate for the top Brazilian CS researchers over the five CNPq ranks

Student mentoring

This section addresses question Q2 by comparing the number of Master's and Ph.D. students mentored by the researchers. Figure 3 shows such distribution across the five CNPq ranks. Except between ranks 1A and 1B, there is a clear decrease in the median value of the number of graduate students from higher ranks to lower ones in both graphs. In Fig. 3a, the number of Master's students mentored by researchers with ranks 1A and 1B are similar, and the median values are 34, 35, 25, 20 and 11, respectively from rank 1A to 2. In Fig. 3b, the gap between the medians of ranks 1A and 1B is noticeable, as it is the equality between 1A and 1C. The median values for doctorates mentored are 9, 12, 9, 5 and 1, respectively from rank 1A to 2. The performance of researchers ranked 1A may be explained by administrative activities that these researchers usually assume (chancellor, provost, dean, director, funding agency committee members and so on). In other words, advising students may not be the only priority of the researchers from the top rank. Moreover, it is important to notice that when the current researchers at rank 1A started their carrers, the CS graduate programs in Brazil were starting as well with a very small number of students. In fact, most of those researchers are also founding fathers of their respective programs.

In order to eliminate any bias towards researchers with long scientific careers, in Fig. 4, the number of students mentored is normalized by the researchers' career length. The ratio of Master's students mentored per year is quite similar among the CNPq ranks, with a median value close to one student per year. Therefore, these results may not be used to distinguish researchers with different CNPq ranks. On the other hand, the ratio of doctorates mentored differs among CNPq ranks, but is close to one doctorate student per two years along the scientific career in average.

It is also important to notice that Figs. 3 and 4 show many outliers in rank 2. In order to explain such phenomenon, consider Fig. 2 as well, which shows many researchers with more than 25 years of career in rank 2. Consequently, these researchers mentored more students than others in this rank, which explains the outliers in Figs. 3 and 4.

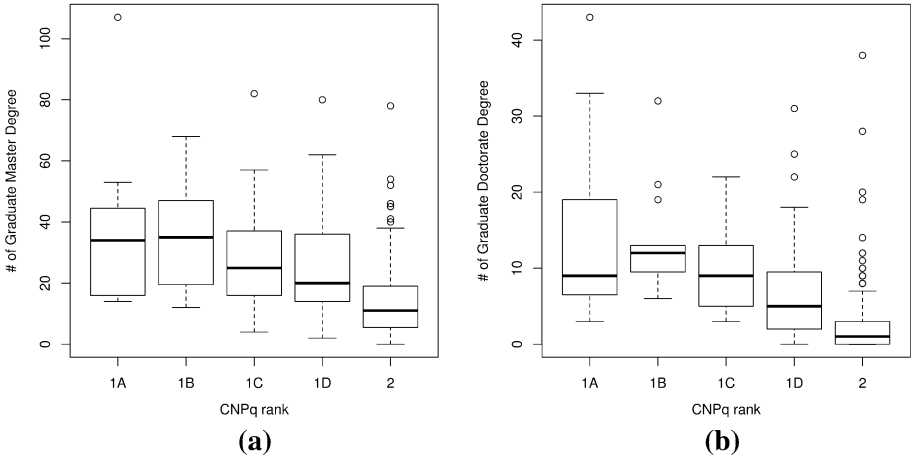


Fig. 3 Distribution of the number of **a** Master's and **b** Ph.D. students mentored by the top CS Brazilian researchers over the five CNPq ranks

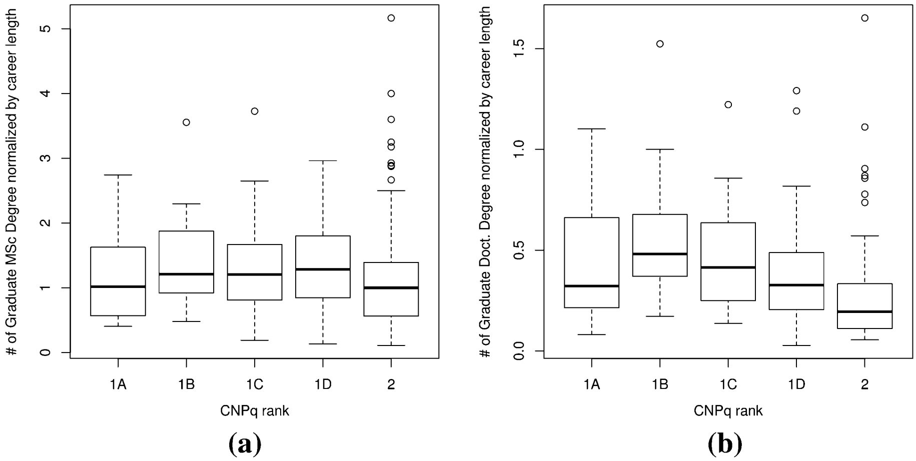


Fig. 4 Distribution of the number of **a** Master's and **b** Ph.D. students mentored by the top CS Brazilian researchers over the five CNPq ranks normalized by career length

Scientific productivity

In this section, we analyze the researchers' scientific productivity based on their publications. To address research question Q3, we compare the researchers in terms of volume of publication and citation count across the CNPq ranks. Figure 5 presents the distribution of publications per CNPq rank, considering (a) publication volume, (b) citation count and (c) h-index. Note that each point represents a researcher, and the value assigned to each researcher is normalized by the maximum value, bindind the score between 0 and 1.

In terms of volume of publication in Fig. 5a, there is a descending median score from researchers with rank 1B towards rank 2, but the median value from rank 1A is lower than 1B and has a value similar to rank 1C. One possible reason for the higher productivity of rank 1B is the higher number of students mentored by the researchers in this rank, as presented in Fig. 3. Likewise, the relatively lower performance attained by the researchers from rank 1A may be explained due to the administrative activities such researchers must undertake and the fact that they target higher impact venues (as detailed in Section "Impact-based stratified analysis"). The same pattern can be seen in terms of citation count distribution per CNPq rank. Again, the median value decreases from rank 1B towards rank 2, and the median value of rank 1A is lower than the median value from rank 1B. Even though the median value is higher for rank 1B, there is only one researcher with score greater than 0.4, whereas for rank 1A there are five.

Regarding h-index, given that it considers both publication volume and citation count, Fig. 5c shows the same pattern of Fig. 5a, b, with the median value decreasing from rank 1B to 2. Overall, the distribution of publication volume, citation count and h-index along the CNPq ranks follows the intuition of the median value decreasing for lower ranks, except for rank 1A. However, we must point out that a considerable part of the 1A researchers belong to the area of Algorithms and Theory (as detailed in Section "Impact-based stratified analysis"), whose distinguished feature is to have less volume of production with more impact. Finally, as previously discussed, for a researcher to be promoted to higher ranks, the whole career must be considered, not only the researcher's publications. Even considering only publications between 2001 and 2011, our analyses confirm

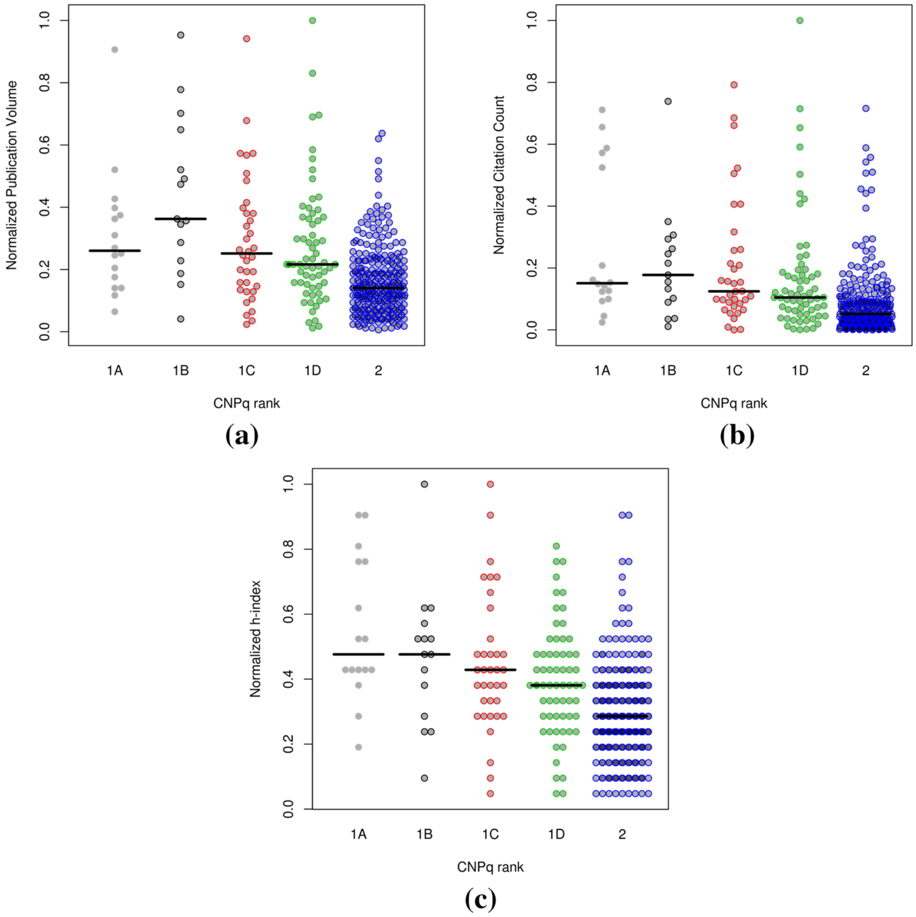


Fig. 5 Distribution of **a** publication volume, **b** citation count and **c** h-index of top Brazilian researchers over the five CNPq ranks, normalized by the maximum value across all ranks

that experienced researchers tend to have more publications and citations, resulting in a higher h-index.

Impact-based stratified analysis

In the previous section, we performed a global characterization of the profile of the top Brazilian CS researchers. In this section, we further analyze these researchers' publication profile. In particular, while the set of publications is usually the most important factor to assess and rank researchers, simply counting publications is not enough; it is also necessary to consider the quality of each publication, or the venue where it was published (Sugimoto and Cronin 2012). In this context, considering a *quality criterion* for publication venues may improve the assessment fairness, so as to prevent prolific researchers with lower impact publications from being over valued.

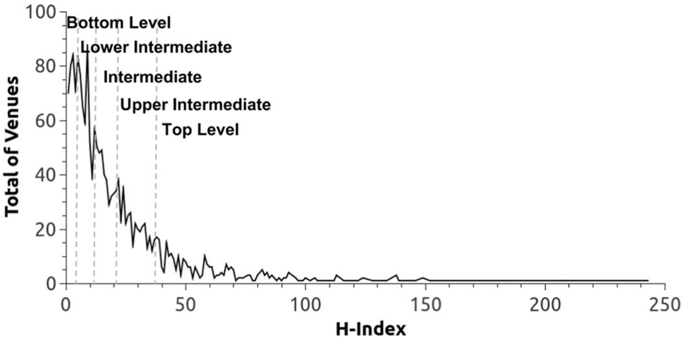


Fig. 6 Distribution of venues per h-index stratum. The venues are classified into five strata, corresponding to homogeneous percentiles, each covering 20 % of all venues: bottom, lower intermediate, intermediate, upper intermediate and top. The top stratum contains 20 % of the venues with the highest h-index value, whereas the bottom stratum contains 20 % of the venues with the lowest h-index values

In this analysis, we rely on a simple measure of venue impact, based upon the h-index of each venue. The h-index of a publication venue is equal to *n*, if *n* of its publications have at least *n* citations (Hirsch 2005). Given the skewed distribution of h-index per venue, we classify each venue in our dataset into five strata, corresponding to homogeneous percentiles, each covering 20 % of all venues: bottom, lower intermediate, intermediate, upper intermediate and top. In other words, the top stratum contains 20 % of the venues with the highest h-index value, whereas the bottom stratum contains 20 % of the venues with the lowest h-index values.

Figure 6 shows the distribution of venues per h-index stratum. Note that the majority of venues has an h-index between 10 and 40, and only few venues have an h-index higher than 50, which forms a long tail. While there are alternative schemes for classifying the impact of different publication venues, such as Thomson Reuters's Impact Factor,[6] they are not generally applicable to conferences and journals alike, which is an important requirement for assessing research productivity in the CS field (Laender et al. 2008). Given our defined measure of impact, we aim to answer the following research questions:

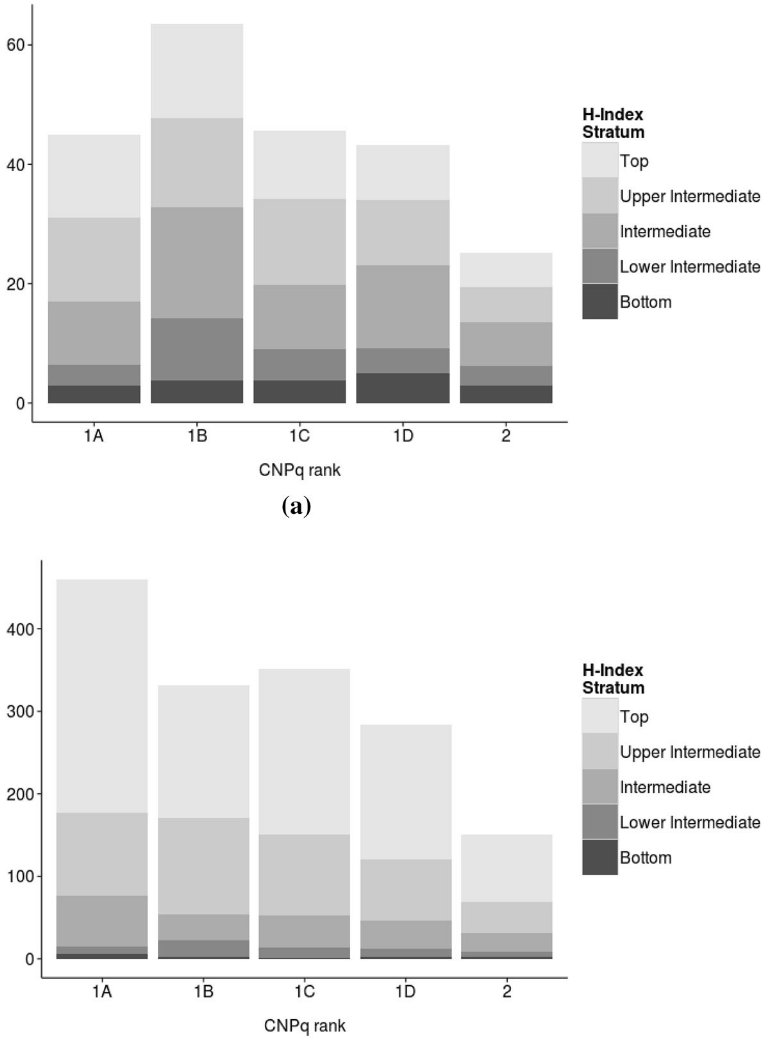
Q4: What is the impact of the venues targeted by Brazilian CS researchers?
Q5: How does this impact vary across different CS areas?

The remainder of this section addresses these two questions.

Analysis of the impact of venues

In order to assess the quality of the top Brazilian CS researchers' productivity and address question Q4, we analyze the impact of these researchers' publications according to the aforementioned h-index strata. Figure 7 shows the average publication volume and citation count along the h-index strata for each CNPq rank. From the figure, we observe that the researchers do not always publish in venues that belong to the top stratum. An increasing focus towards such top venues is only remarkable for researchers ranked 1A, whereas the researchers ranking lower tend to spread their publishing activity in venues of the other strata.

[6] Impact factor: http://thomsonreuters.com/journal-citation-reports/.



(a)

(b)

Fig. 7 Distribution of **a** publication volume and **b** citations along venue h-index stratum for each CNPq rank. Each value is the average number of either publications or citations per stratum for the researchers in each CNPq rank

By contrasting Fig. 7a, b, we observe that researchers ranked 1B publish the most in terms of total volume, followed by researchers ranked 1A, 1C, 1D, and 2. However, this higher volume of publications does *not* necessarily translate into a higher citation rate. Indeed, researchers ranked 1A and 1C present almost the same publication volume, but those ranked at 1A present higher citation rates than those at all the other ranks. In fact, the citation rates of researchers ranked 1B are even lower than those ranked 1C, and close to researchers ranked 1D. Finally, the difference between researchers ranked 2 and the others in terms of both publication volume and citation count becomes more pronounced as we consider venues at higher h-index strata.

Although this analysis presents aspects of quality of the researchers' publications, we cannot state that the researchers from one rank are indeed *better* than the others based only on the venue h-index strata. It would be very easy to say that, based on our results, top publications reflect the work of top researchers. However, that claim does *not* account for new venues that have the potential to grow in number of citations (e.g., recently released ACM journals), venues specialized in new areas of expertise that have been more explored recently (for example *nanotechnology*), and venues kept by relatively smaller communities (for example, *music computing*). Such issues regarding the profile of each area demand a more detailed analysis and are further discussed in the next section.

Impact-based analysis per CS area

According to Glänzel and Schubert (2003), Oliveira et al. (2012), Lima et al. (2013) and Wainer et al. (2013), the relative performance of researchers may significantly vary depending on their field of work. This is also true for areas withing one field. For example, regarding CS, experimental evaluation in the area of *human–computer interaction* usually takes longer than in other CS areas when arranging and assessing users' feedback is necessary (Barbosa and de Souza 2011). In contrast, CS areas such as *databases* and *computer graphics* do not usually face the same problem as their experimental evaluations often depend on the outcome of an *automatic* process, such as a query evaluation or a graphics rendering engine. In order to account for these peculiarities and address research question Q5, we now characterize the distribution of researchers and publications per CNPq rank and h-index strata across the 23 CS areas previously described in Section "Data acquisition and preparation".

Figure 8 shows the distribution of publications per venue h-index strata across the 23 considered CS areas. In particular, each bar represents the average number of publications per researcher in each area in logarithmic scale. From Fig. 8 we observe that different areas have different venue h-index distribution patterns, which corroborates the works cited in the previous paragraph. In general, researchers from all areas tend to publish in venues belonging to the top stratum, except for few areas such as *games and entertainment*, *health informatics* and *music computing*, which are new areas without a portfolio of consolidated publication venues (i.e., with enough time of existence to have a higher h-index value). Other areas have an increasing focus on venues of higher h-index strata, such as *algorithms and theory* and *information systems*. It is not possible to claim that some researchers have better publications than others based solely on the venue h-index strata distribution, because each area has its own peculiarities, such as lack of venues with high impact, time-consuming experimental evaluation, and so on. For example, *computational biology* concentrates its publication in the top stratum. In fact, many publication venues from this area have high impact (e.g., Nature, Science and Cell). Therefore, researchers from this area may publish fewer papers, but with higher impact. On the other hand, researchers from *geoinformatics* concentrate their publications in the intermediate stratum, but this area is affected by the lack of venues with higher impact.

Figure 9 shows the distribution of researchers per CNPq rank and CS area, derived from the publications authored by each researcher and their assigned area. For instance, a publication assigned to three distinct CS areas and authored by a researcher ranked 1A contributes to the presence of 1A researchers in each of these three areas with a score of 1/3. Once again, given the skewed distribution across the 23 considered areas, we use a logarithmic scale. From Fig. 9, we observe that *games and entertainment* has no researcher ranked 1A or 1B. It may be because this area is a novel field of knowledge in CS and there

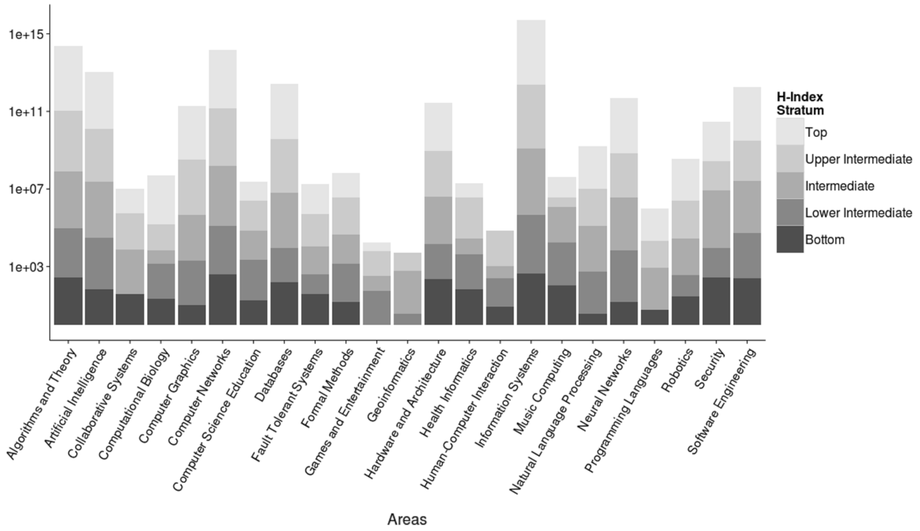


Fig. 8 Distribution of the average number of publications per venue h-index stratum and 23 CS areas, per researchers in each area, using logarithmic scale

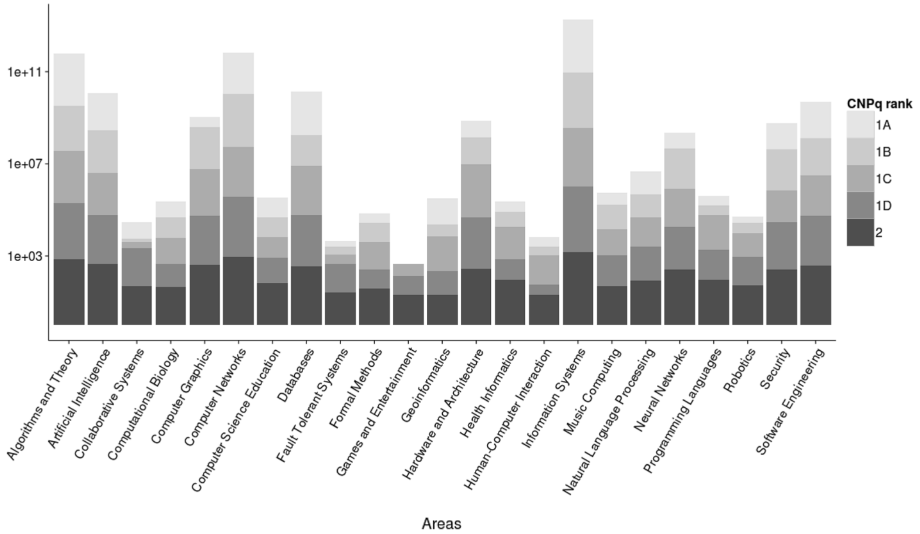


Fig. 9 Distribution of researchers per CNPq rank and 23 CS area derived from the publications authored by each researcher and their assigned area, using logarithmic scale

is no researcher who meets all the requirements to belong to these ranks. Likewise, the *collaborative systems* and *fault tolerant systems* areas follow the same pattern, with a substantial presence in ranks 1D and 2.

In general, we observe that most researchers tend to publish in venues from the top stratum, where the exceptions may be explained by the peculiarities of each area. We also note that some areas, such as *algorithms and theory* and *information systems* concentrate more publications than others, which happens because a large part of the venues are classified in these two areas. On the other hand, areas such as *games and entertainment* and *geoinformatics* have fewer publications, probably because of the size and age of their communities. Indeed, relatively newer areas of knowledge in CS tend to suffer with minimum or no representation at the top stratum. In turn, the lack of researchers ranked 1A or 1B may indicate that these areas receive less funding, which may impact their researchers' performance. On the other hand, areas such as *algorithms and theory*, *computer networks*, *databases*, *information systems* and *software engineering* have researchers in all CNPq ranks, which indicates that such areas are more consolidated in Brazil.

Concluding remarks

We have analyzed the scientific profile of the top Brazilian CS researchers. We considered 406 researchers who participate in the CNPq Researcher Fellowship Program. The characterization study focused on three aspects: career length (years since the researcher received the doctorate degree), student supervision (number of students mentored) and scientific productivity (publication volume and citation count). Regarding scientific productivity, we further stratified the publication outcome of each researcher based upon the impact of their targeted venues and contrasted the stratified productivity of researchers across different CS areas. The main results from our evaluation may be summarized as follows:

1. In terms of career length, researchers ranked higher by CNPq tend to be more experienced, although there are some researchers ranked at 2 whose careers are longer than 25 years. Furthermore, we observed that it usually takes longer for a researcher to be promoted from rank 2 to 1D, as a consequence of specific time requirements and the financial impact of the promotions, and from 1B to 1A, as a consequence of the international recognition requirement, compared to promotions between other ranks.
2. The ratio of Master's students mentored during the researchers' career is close to one student per year, and this ratio is quite similar for all CNPq ranks. However, the ratio of doctorates mentored differs across the CNPq ranks. The median values of doctorates during the researchers' career decrease from ranks 1A–1D to rank 2, except for rank 1A, whose median value is very similar to rank 1D. The number of doctorate programs has been increasing consistently in the last years in Brazil and more experienced researchers, who usually participate in the older programs, have been advising PhDs longer, explaining such differences. However, we notice again that when most current researchers at rank 1A started theirs careers, the CS graduate programs in Brazil were at early days and offered only a Master's degree.
3. Regarding the distribution of publication volume and citation count per CNPq rank, we observe an expected distribution, in which the median values decrease towards lower ranks. Again, the exception is rank 1A, which has a median value similar to rank 1C for both criteria.
4. The CNPq researchers do not always publish in venues that belong to the top stratum. Except for researchers ranked at 1A, who publish more on venues belonging to the top stratum, all the remaining ranks devote most of their efforts to venues from the intermediate and upper intermediate strata, which means that there is some room for improving the quality of theirs publications.

5. Although most of the CS areas tend to concentrate their publications in venues from the top stratum, peculiarities of each area (e.g., small community size, lack of venues with high impact, lack of experienced researchers) influence the distribution of publications towards venues with lower impact.

Such results are very important in many aspects. First, besides counting publications and citations, considering career length, number of students mentored and the impact of publications may provide new insights to governments and funding agencies. Also, manual research assessments may take advantage of our methodology in order to identify researchers who are outliers, i.e., their performance is considerably better than those in the same ranks (see Fig. 5). As other studies have shown—e.g., (Cerchiello and Giudici 2014), our results do emphasize that higher volume of publications does not imply equaly high number of citations (see Fig. 5). Indeed, it is also clear that researchers ranked at 1A have more citations even for a recent publication interval (our study considers publications between 2001 and 2011). Likewise, such researchers may have similar performance to the other ranks regarding number of students mentored, publication volume and citations; however, they distinguish themselves by focusing their publications on the top venues (considering our venue h-index strata in Fig. 7). We have also taken previous bibliometric studies (as those in the "Related work" section) one step forward by showing how important the subareas of expertise are when assessing research profiles (Figs. 8, 9). Finally, a previous study has shown the negative association between productivity indicators and the average age of researchers (Bonaccorsi and Daraio 2003). However, the authors consider a single year of publication. Here, our results are based on 10-year data and show that, despite a slightly lower performance by senior researchers in terms of publication count, they keep their productivity standard by targeting higher impact venues.

Overall, our results show that CNPq ranks are coherent for the computer science area, except for a few outliers which are usually explained by historical and personal reasons (e.g., absence of leave for industry). It is also worth mentioning that no single criterion is enough for explaining the CNPq rank and we foresee opportunities for other criteria to be added.

As a methodological improvement, future studies could consider the researchers' international collaboration, and how it may impact their academic productivity. Furthermore, the simple impact metric used in our analysis could be extended to better capture additional characteristics of the publication venues, such as longevity and periodicity.

Acknowledgments The authors would like to thank Isac Sandin (UFMG) for his invaluable help with processing data from the Lattes Platform, as well as the team behind the SHINE project, notably Prof. Altigran S. da Silva (UFAM). The authors would also like to acknowledge the financial support from InWeb—National Institute of Science and Technology for the Web, and their individual grants from CNPq and FAPEMIG.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). National research assessment exercises: the effects of changing the rules of the game during the game. *Scientometrics*, *88*(1), 229–238. doi:10.1007/s11192-011-0373-2.

Almeida, E., & Guimarães, J. (2013). Brazil's growing production of scientific articles—how are we doing with review articles and other qualitative indicators? *Scientometrics*, *97*(2), 287–315. doi:10.1007/s11192-013-0967-y.

Barbosa, S. D. J., & de Souza, C. S. (2011). Are HCI researchers an endangered species in Brazil? *ACM Interactions Magazine*, *18*(3), 69–71. doi:10.1145/1962438.1962454.

Bonaccorsi, A., & Daraio, C. (2003). Age effects in scientific productivity. *Scientometrics*, *58*(1), 49–90. doi:10.1023/A:1025427507552.

Bosquet, C., & Combes, P. P. (2013). Are academics who publish more also more cited? Individual determinants of publication and citation records. *Scientometrics*, *97*(3), 831–857. doi:10.1007/s11192-013-0996-6.

Cerchiello, P., & Giudici, P. (2014). On a statistical h index. *Scientometrics*, *99*(2), 299–312. doi:10.1007/s11192-013-1194-2.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, *61*(9), 1853–1870. doi:10.1002/asi.21363.

Delgado-Garcia, J.F., Laender, A.H.F. Jr., W.M. (2014). Analyzing the Coauthorship Networks of Latin American Computer Science Research Groups. In *Proceedings of the 9th Latin American Web Congress*, Ouro Preto, Brazil, pp 77–81.

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, *41*(2), 15–26. doi:10.1145/2350036.2350040.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. doi:10.1023/A:1022378804087.

Glänzel, W., Leta, J., & Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, *67*(1), 67–86. doi:10.1007/s11192-006-0055-7.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16,569–16,572. doi:10.1073/pnas.0507655102.

Ibáñez, A., Larrañga, P., & Bielza, C. (2013). Cluster methods for assessing research performance: exploring Spanish computer science. *Scientometrics*, *97*(3), 571–600. doi:10.1007/s11192-013-0985-9.

Ingwersen, P., Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–2012. *Scientometrics*, *101*(2), 1325–1344. doi:10.1007/s11192-014-1291-x.

Kato, M., & Ando, A. (2013). The relationship between research performance and international collaboration in chemistry. *Scientometrics*, *97*(3), 535–553. doi:10.1007/s11192-013-1011-y.

Kutlar, A., Kabasakal, A., & Ekici, M. (2013). Contributions of Turkish academicians supervising phd dissertations and their universities to economics: An evaluation of the 1990–2011 period. *Scientometrics*, *97*(3), 639–658. doi:10.1007/s11192-013-0973-0.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., & Ziviani, N. (2008). Assessing the research and education quality of the top Brazilian computer science graduate programs. *SIGCSE Bulletin*, *40*(2), 135–145. doi:10.1145/1383602.1383654.

Lamont, M. (2012). Toward a comparative sociology of valuation and evaluation. *Annual Review of Sociology*, *38*(21), 201–221. doi:10.1146/annurev-soc-070308-120022.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, *464*(7288), 488–489. doi:10.1038/464488a.

Lee, D., Kang, J., Mitra, P., Giles, C. L., & On, B. W. (2007). Are your citations clean? *Communications of the ACM*, *50*(12), 33–38. doi:10.1145/1323688.1323690.

Leta, J., Glänzel, W., & Thijs, B. (2006). Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics*, *67*(1), 87–105. doi:10.1007/s11192-006-0051-y.

Lima, H., Silva, T. H. P., Moro, M. M., Santos, R. L. T., Meira, W, Jr, & Laender, A. H. (2013). Aggregating productivity indices for ranking researchers across multiple areas. *Proceedings of Joint Conference on Digital Libraries*. Indiana, USA: Indianapolis, pp. 97–106. doi:10.1145/2467696.2467715.

Mamtora, J., Wolstenholme, J. K., & Haddow, G. (2014). Environmental sciences research in northern Australia, 2000–2011: A bibliometric analysis within the context of a national research assessment exercise. *Scientometrics*, *98*(1), 265–281. doi:10.1007/s11192-013-1037-1.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M., & Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, *65*(7), 1424–1445. doi:10.1002/asi.23010.

Menezes, G.V., Ziviani, N., Laender, A.H., Almeida, V. (2009). A Geographical Analysis of Knowledge Production in Computer Science. In: *Proceedings of International World Wide Web Conference*, Madrid, Spain, pp. 1041–1050. doi:10.1145/1526709.1526849.

Miller, J., Coble, K. H., & Lusk, J. L. (2013). Evaluating top faculty researchers and the incentives that motivate them. *Scientometrics*, *97*(3), 519–533. doi:10.1007/s11192-013-0987-7.

Oliveira, E. A., Colosimo, E. A., Martelli, D. R., Quirino, I. G., Oliveira, M. C. L., Lima, L. S., et al. (2012). Comparison of Brazilian researchers in clinical medicine: are criteria for ranking well-adjusted? *Scientometrics*, *90*(2), 429–443. doi:10.1007/s11192-011-0492-9.

Riikonen, P., & Vihinen, M. (2008). National research contributions: A case study on finnish biomedical research. *Scientometrics*, *77*(2), 207–222. doi:10.1007/s11192-007-1962-y.

Sugimoto, C. R., & Cronin, B. (2012). Biobibliometric profiling: An examination of multifaceted approaches to scholarship. *Journal of the Association for Information Science and Technology*, *63*(3), 450–468. doi:10.1002/asi.21695.

Torrisi, B. (2014). A multidimensional approach to academic productivity. *Scientometrics*, *99*(3), 755–783. doi:10.1007/s11192-013-1149-7.

Vanecek, J. (2014). The effect of performance-based research funding on output of r&d results in the Czech Republic. *Scientometrics*, *98*(1), 657–681. doi:10.1007/s11192-013-1061-1.

Veloso, A., Meira, W, Jr, Gonçalves, M., & Zaki, M. (2007). Multi-label lazy associative classification. In: *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Warsaw, Poland, pp. 605–612.

Wainer, J., Xavier, E. C., & de Lima, Bezerra F. (2009). Scientific production in computer science: a comparative study of Brazil and other countries. *Scientometrics*, *81*(2), 535–547. doi:10.1007/s11192-008-2156-y.

Wainer, J., Eckmann, M., Goldenstein, S., & Rocha, A. (2013). How productivity and impact differ across computer science subareas. *Commun ACM*, *56*(8), 67–73. doi:10.1145/2492007.2492026.