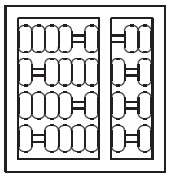

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Brazilian Computer Science research: gender and regional distributions

Denis Arruda *Fábio Bezerra*
Vânia Almeida Neris *Patricia Rocha de Toro*
Jacques Wainer

Technical Report - IC-07-32 - Relatório Técnico

October - 2007 - Outubro

Brazilian Computer Science research: gender and regional distributions

Denis Arruda Fábio Bezerra Vânia Almeida Neris
Patricia Rocha de Toro Jacques Wainer

October 3, 2007

Abstract

This paper analysis the distribution of some characteristics of computer scientists in Brazil according to regions and gender. Computer scientist is defined as the faculty of a graduate level computer science department. Under this definition, there were 886 computer scientists in Brazil in November 2006. We studied the self-declared research areas, the production of journal and conference papers of these scientists, and whether they receive a recognition fellowship by one of Brazilian grant agents, and analyse the differences regarding the geo-political region in Brazil and the gender of the scientists.

1 Introduction

There is a growing interest in both gender and national aspects of science production.

1.1 Related research

Regarding gender and the Sciences, an important line of research follows [1] "productivity puzzle" findings, which states that females have lower productivity than their male counterparts.

[2] analyses Spanish male and female researchers in Material Sciences, and finds no difference in the productivity between these two groups. This study also suggests that women are making a remarkable effort to increase their presence in science and to adjust to patterns of international scientific excellence: women publish in journal with higher impact than men.

There is an intense research regarding gender and computing. For example, [3] reviews the literature up to 2002 on gender and computer science, but with an emphasis on students, from preschoolers (for example [4]), to undergrads (for example [5]), to graduate students. Another important work is [6] which showed the decreasing proportion of females in CS from high school to graduate school to faculty positions. [7, 8, 9, 10, 11, 12] are examples of studies of differences of men and women in the IT workplace, with emphasis on the barriers women face in accepting and keeping an IT job.

Regarding CS research, [13] examines the productivity of male and female researchers in the area of Information Systems (IS). Similar to our work, that work aims at comparing

the proportion of published articles from female researchers in relation to male researchers, compared to the *baseline* - the proportion of female to male researchers. As their universe, the authors select the top 251 most prolific researchers in 12 of the top journals in the area. In this universe, the productivity of women and men are not significantly different.

Regarding national aspects of Computer Science production, there have been a series of publications, some comparative and some not, that analyses details of each country pattern of publication. According to [14], Chinese researchers prefer to publish their research results in domestic journals, while Indian researchers prefer to publish their research results in journals published in the advanced countries of the West (during 1971-2000). During the period under study, India's research output was significantly higher than Chinese research output. However, China's output has increased significantly.

[15] is a comparative study of Computer Science research performance in the USA, Germany, UK, Japan, India and China during the period from 1993 to 2002. This study shows that China's output has been growing fast, while the other five countries all have a declining trend. The scientists in Western countries are more likely to work independently or only prefer to work with one collaborator, while Asian scientists preferred to work in big groups that usually have more than three scientists.

Brazilian science has been the topic of many publications, although none regarding Computer Science. Brazilian science, in general, has been analyzed in [16, 17], for example. Brazilian science has also been studied in larger context of Latin-Caribbean countries [18]. Ibero-American countries [19], or third world countries [20]. In reference to particular disciplines, [21] analyses the distribution between domestic and international publication in Psychiatry for the period of 1981 to 1995. [22] discusses the increase in publication from Brazilian scientist in the area of Limnology (a sub area of Ecology) from 1970 to 2004, and the lower than average number of citation these article receive.

An analysis of scientific production of Brazilian men and women in astronomy, oceanography and immunology during 1997-2001 is presented in [23]. This study shows that there is effectively no difference between them in potential impact. The authors found evidence that Brazilian women tend to receive fewer "fellowships grants" (see section 2.1).

Finally, there are not much research on regional differences of science production within a single country. [24] discusses the patterns of inter-region scientific collaboration in China. [25] studies the distribution of R&D production, the flow of R&D personnel across different regions, and the patters of co-authorship in scientific publications in and between the different Swedish regions.

2 Background

2.1 The Lattes curriculum system

Brazil has developed an interesting system to record the production of its scientists. It is called the Lattes curricula system, or Lattes CV for short. The current version of the system allow researchers to update their Lattes CV using a Web system. We will discuss below that researchers have many incentives to keep their CV updated, and thus, the Lattes system is a very valuable resource regarding the Brazilian science.

In Brazil there are two main science funding agencies, CAPES (Coordination for the Improvement of Higher Education Personnel, a section of the Ministry of Education) and CNPq (National Council for the Development of Science and Technology, a part of the Ministry of Science and Technology). CNPq is the funding agency that evaluates and funds researchers, and CAPES evaluates and support graduate courses. CNPq funds research based on the peer evaluation of the merits of the proponents and of their research proposals. During many years CNPq has pushed forward the idea of a standardization of the researchers' curriculum, to facilitate the analysis and comparison of the researchers' scientific production. This effort culminated in the Lattes system, a Web based data entry system that centralizes all Brazilian researchers curricula, in a standard format, and with a standard language. The data for each researcher is publicly available at the CNPq site.

CNPq also provides a particular form of funding for researches, called *productivity fellowships* that pays a "salary" to the researches, that is, a research fund that is discretionary and need not to be reported back to the CNPq. The fellowships are divides into classes 1A, 1B, 1C, 1D and 2, by decreasing order of value and prestige. The fellowships of level 2 pays the only the salary, fellowships of class 1 also pays an extra grant money (almost the same value of the salary) that need to be spent in research related activities and need to be reported.

The number of fellowships, both level 1 and 2 are limited. Researchers submit a fellowship proposal that is evaluated by a committee of peers. A researchers with no fellowship can only be accepted into level 2, if their proposal and their research record are approved by peer evaluators and by the committee, and if there are openings for that fellowship. A researcher which already receives a fellowship has to renew his fellowship every two or three years (depending on the level) and as result may retain the fellowship at his current level, or be promoted to the level above, or demoted to the level below.

The level of a researcher's fellowship is a recognized measure of scientific prestige. For example, some grant proposal can only be requested by level 1 researchers, and level 1 researchers are the only ones that can be members of both CAPES and CNPq evaluating committees, and that can vote for candidates for these committees.

CAPES evaluates all the graduate courses in Brazil. Courses are evaluated by a grade from 1 to 7, where grade 3 and above grants the course a "recognition of quality" by CAPES. CAPES also distributes to each graduate course a number of student scholarships, based on the course evaluation. Most graduate level students receive these "institutional" scholarships, although some faculty may have research grants (from other sources) that support their students.

The course evaluation is performed by a committee of researches selected from the respective scientific community. The evaluation of the courses is based strongly on the scientific production of the faculty, but other factors are also taken into consideration, such as average time to graduation, broadness of topics covered in graduate courses and in the faculty research, and so on. Recently, CAPES and CNPq unified their systems, so that the data regarding the faculty productivity is automatically extracted from the Lattes CV.

Thus, the Lattes CV became the central repository of information regarding the Brazilian researchers, and there are strong incentives, both personal and institutional, to keep one's Lattes CV updated.

Sub-Area	Specialities
Computer Mathematics	Symbolic Mathematics Analytical and Simulation Models
Computer Methodology and Techniques	Database Software Engineering Programming Languages Graphics Processing Information Systems
Computer Systems	Computer Systems Architecture Hardware Basic Software Teleinformatics
Computer Theory	Algorithms and Complexity Analysis Computability and Computer Models Formal Languages and Automata Logics and Semantics of Programs

Table 1: Computer science knowledge areas (adapted from CNPq).

2.1.1 Research topics

One of the fields in the Lattes CV is the researcher's self-declared research interests or research topics. The Lattes system, has a four level hierarchy to specify a research topic. The higher level, called *grand area*, is divided into Engineering, Exact Sciences, Social Sciences, and so on. The next level is called *area*; Computer Science, is one of the areas of Exact Sciences. The next level is called *sub-area*, and for Computer Science there are four areas: Computer Mathematics, Computer Methodology and Techniques, Computer Systems, and Computer Theory, and each sub-area is further divided into *specialities*. The list of sub-areas and specialities is displayed in Table 1.

We could not find the source or the rationale for this particular division of CS, nor a clear definition of each of the categories. But the choices of subareas is clearly limited, and does not contemplate known subareas such as Artificial Intelligence, Human Computer Interfaces, among others. In the first versions of the system, the researcher would have to choose among those alternatives, but later versions allow the researcher to enter a set of keywords as their sub-areas, and specialities instead of the predefined choices. Furthermore, the researcher can declare many research interests, in different grand areas, if needed.

2.2 Brazilian geopolitical regions

Brazil is currently divided into five geopolitical regions, grouped by territorial proximity and geographic characteristics. That division was proposed by IBGE[26] with the objective of organizing statistical and economic analysis. A summarized description of economic vocation for each geopolitical region is presented bellow:

North : It is comprised of seven states which constitute the largest area of Brazilian Amazon. Its economy is based mainly on mineral extraction (gold, aluminum, iron, manganese and nickel) and vegetable extraction (wood). The industrial economy is very incipient, and it is mainly based on improvement of agricultural products.

Northeast : It is comprised of nine states which constitute the largest coast of Brazil. Its economy is based on agriculture, industry, tourism and petroleum extraction. In agriculture, sugarcane is the most expressive product, while in industry, electronic products and software are the prominence.

Midwest : It is comprised of three states and one federal district, where it is placed the capital of Brazil. Its economy is based mainly on animal husbandry and agriculture. The vegetable and mineral extractions also constitute the economy of region, but in a smaller scale if compared to husbandry and agriculture. However, the industry is little expressive.

Southeast : It is comprised of four states and it represents the most developed region of Brazil. Its economy is very strong, diversified and corresponds to 55% of Brazilian's GDP (Gross Domestic Product). The sugarcane, orange and soy are prominent products of agriculture. In husbandry, the region has the second largest cattle, while raising birds and the egg production are the largest of country. In high tech industry, the prominence is the Brazilian's "Silicon Valley", which is comprised of four cities (Sao Paulo, Sao Jose dos Campos, Sao Carlos and Campinas) and shelters one of the best universities of high tech research from Brazil (USP, UNICAMP and ITA).

South : It is comprised of three states and it represents the region with the best HDI (Human Development Index) of Brazil. Its economy is based mainly on agriculture and industry, despite the husbandry be an important component of the economic system of region. In industry, the expressiveness sectors are the automobile industry and the textile industry.

Figure 1 shows the regions, and Table 2 shows some relevant socio-economic data on each region. The GDP (Gross Domestic Product) and HDI (Human Development Index) values were obtained from IBGE 2004 [26] and PNUD [27].

Region	GDP	Average HDI 2000	Population (millions)
North	5%	0.725	12.9
Northeast	14%	0.676	47.7
Center-West	8%	0.792	11.6
Southeast	55%	0.791	72.4
South	18%	0.807	25.1

Table 2: Socio-economic characterization of the Brazilian geopolitical regions.

Figure 1: Brazilian Geopolitics Regions

3 Methodology

We selected all graduate courses in Computer Science that were classified with a grade 3 or above by CAPES, which adds up to 44 courses (in October 2006). The list of courses (by the name of the university) and their corresponding regions are displayed in Table 3.

The list of faculty for each of the 44 courses was found in each of the respective departments' websites. We considered as researcher all faculty listed in the website, including the faculty listed as external members. About 2% of the names were repeated, that is, the same faculty is listed in two or more courses. In that case, we classified the researcher in his self-declared primary course.

We accessed the Lattes CV of each of the names above. 1% of the researchers had no Lattes CV, and were discarded. In the end, we collected and classified the Lattes CV of 886 people, which we consider the set of CS researches in Brazil.

The data collection regarding the production of each of the 886 researchers was performed from the middle of October to the middle of November, 2006. To guarantee that data would be collected following a unique process, the four people received training about how data should be collected and inserted in database. Each people collected approximately 240 curricula. Around 10% of records were analyzed by a second person. Such analysis worked as a validation for the data collection method.

We collected all publication mentioned in their Lattes CV from 2000 to 2006. The Lattes CV has different categories for journal and conference publications and we used these categories for our data collection. We made no distinction on whether the publication was in a national or an international journal, or if it was in a national, regional, or international conference. But some researchers did classify publications in Springer's Lecture Notes series as journal publications. In this case, we disconsidered the entry.

Institution	State	Region	Institution	State	Region
FEESR	SP	Southeast	FESP/UPE	PE	Northeast
IME	RJ	Southeast	FUNECE	CE	Northeast
PUC/MG	MG	Southeast	UFBA	BA	Northeast
PUC-RIO	RJ	Southeast	UFC	CE	Northeast
UFES	ES	Southeast	UFCG	PB	Northeast
UFF	RJ	Southeast	UFPB/J.P.	PB	Northeast
UFMG	MG	Southeast	UFPE	PE	Northeast
UFRJ	RJ	Southeast	UFRN	RN	Northeast
UFRJ	RJ	Southeast	UNIFACS	BA	Northeast
UFSCAR	SP	Southeast	UNIFOR	CE	Northeast
UFU	MG	Southeast	UFAM	AM	North
UFV	MG	Southeast	UFPA	PA	North
UNESP/SJRP	SP	Southeast	PUC/PR	PR	South
UNICAMP	SP	Southeast	PUC/RS	RS	South
UNIMEP	SP	Southeast	UCPEL	RS	South
UNIRIO	RJ	Southeast	UEM	PR	South
UNISANTOS	SP	Southeast	UFPR	PR	South
USP	SP	Southeast	UFRGS	RS	South
USP/SC	SP	Southeast	UFSC	SC	South
UFG	GO	Midwest	UFSM	RS	South
UFMS	MS	Midwest	UNISINOS	RS	South
UNB	DF	Midwest	UNIVALI	SC	South

Table 3: Graduate programs in Computer Science

3.1 Research topic

Since the pre-defined subareas and specialities in computer science are uninformative, as discussed above, we defined 12 "research topics" which, we believe, better define the different sub-areas of Computer Science, as seen by the community. These research topics are: artificial intelligence, bioinformatics, collaborative systems, computer in education, data bases, hardware and computer architecture, human-computer interfaces, image processing and computer vision, networks and distributed systems, security, software engineering, and theory.

We collected all sub-areas and specialities under the Computer Science area, as entered by the researchers, and grouped in the 12 research topics. Bellow the keywords used to define each topic:

artificial intelligence : computational intelligence, artificial intelligence, knowledge representation, neural networks, data mining, fuzzy logic, machine learning, automatic reasoning, knowledge based system, natural language processing.

bioinformatics bioinformatics, computational biology, computational molecular biology, biotechnology

collaborative systems computer supported cooperative work, computer supported collaborative learning, groupware, workflow, CSCW.

computer and education distance education, educational informatics, informatics applied to education, artificial intelligence applied to education

data base data base, distributed data base, temporal data base, data integration, data integration in the Web, information integration, XML and semi-structured data bases

hardware and computer architecture computer architecture, computer systems architecture, tool for integrated circuit design, hardware, microelectronics, instruction level parallelism, parallel processing

human computer interfaces interface design, human-computer interface, human machine interface, usability and accessibility, user interface.

image processing and vision graphic computing, image processing, image analysis, computer vision.

network and distributed systems computer networks, management of computer networks, grid computing, mobile computing, computer protocols, Middleware, distributed systems, parallel and distributed processing, wireless networks, sensor networks, teleinformatics, fault tolerant systems.

security biometry, cryptography, computer security, information security, systems security, computer systems security, computer network security

software engineering : software development, formal methods, object oriented programming/development, aspect oriented programming/development, software components

theory algorithms, distributed algorithms, algorithm complexity, cryptography, quantum computing, computational geometry, formal languages and automata, optimization, combinatorial optimization, continuous optimization, non-linear optimization, graphs.

3.2 Statistical analysis

When comparing some data regarding the regional distribution we used a Chi-squared goodness-of-fit test, as implemented in the `chisq.test` function in R[1]. The distribution of researchers for each region is displayed in table 4 is considered as the target distribution. The fitness of any other regional data to the target distribution is verified with the Chi-squared test, and if the resulting p-value is smaller than 0.05 (95% of confidence), the regional data is considered not to fit the target distribution.

When comparing gender data, we use the exact binomial test for goodness of fit of a binomial distribution, as implemented by the function `binom.test` in R. Given any gender data, the exact binomial test will verify its fitness to a target distribution, which we take to be the total proportion of male and females taken from table 5. The exact binomial test

[1] www.r-project.org

is more precise than a chi-squared test for binomial distributions, but again the p-value of less than 0.05 indicates that the data does not fit the target distribution (with 95% of confidence).

4 Results

4.1 Region

Table 4 shows the distribution of CS graduate courses and researchers by the geopolitical regions.

Region	Graduate Courses	Researchers	Researchers per course
North	2	28	14.0
Northeast	10	155	15.5
Midwest	3	41	13.7
Southeast	19	454	23.9
South	10	208	20.8

Table 4: Distribution of CS graduate courses and researchers by region.

Table 5 shows the distribution of CS researchers by region and by gender.

Region	Male Researchers	Female Researchers
North	25	3
Northeast	128	27
Midwest	30	11
Southeast	338	116
South	164	44
Total	685	201

Table 5: Distribution of researchers by gender and region

4.2 Research Areas

Table 6 presents the distribution of the declared research topics by gender and by the geopolitic regions.

4.3 Publications

In Brazil for the period from 2000 to 2006, for each paper published in Journals there are 4.27 papers published in Conference proceedings. From 2000 to 2006, the Brazilian researchers considered in this paper published 4470 papers in Journals and 19081 papers in Conference proceedings. Table 7 shows the distribution of publications according to region, and table 8, the distribution according to gender.

	Male	Female	North	Northeast	Midwest	Southeast	South
Artificial intelligence	90	43	6	20	5	66	36
Bioinformatics	14	5	0	1	2	12	4
Collaborative systems	7	7	1	2	0	10	1
Computers in education	19	19	2	6	2	16	12
Data base	69	25	6	19	5	104	43
Hardware and comp. arch.	138	12	2	21	6	63	58
Human-computer interfaces	11	11	1	3	0	9	9
Image processing and vision	78	22	1	15	3	52	29
Networks and dist. systems	160	27	8	47	5	79	48
Security	16	4	0	3	1	8	8
Software engineering	140	50	8	48	8	91	29
Theory	161	42	3	38	15	104	43

Table 6: Distribution of Ad hoc research interests by gender and by region.

	Journals	Proceedings
North	68	511
Northeast	547	3305
Midwest	142	357
Southeast	2637	8922
South	1076	5986
Total	4470	19081

Table 7: Publications in journals and conferences by political regions between years 2000 and 2006.

4.4 Productivity Grants

Among the 886 researchers considered in the paper, 206 receive a productivity grant. Tables 9 and 10 display the distribution of CNPq fellowships by region and gender, respectively.

Gender	Journal	Conferences
Male	3494	14636
Female	976	4445

Table 8: Average of publications in journals and conferences by gender.

Fellowship	North	Northeast	Midwest	Southeast	South	Total
1A	0	1	0	11	0	12
1B	0	0	0	13	1	14
1C	0	4	0	11	8	23
1D	0	3	0	29	5	37
2	0	12	6	76	24	120
Total	2	20	6	140	38	206

Table 9: Distribution of fellowships per region.

Fellowship	Male researchers	Female researchers
1A	12	0
1B	11	3
1C	21	2
1D	28	9
Sub Total 1	72	14
2	82	38
Total	144	52

Table 10: Fellowships by gender.

5 Discussion

5.1 Regional differences

The distribution of courses and researchers by region (table 4) cannot be considered both data from the same distribution (χ^2 = 34.1981, df = 4, p-value = 6.786e-07) and thus the ratios of researchers per course are significantly different. There is a linear correlation between the number of courses and the number of researchers per course - the more graduate courses there are in a region, the higher the number of researchers each course has.

The distribution of male and female researchers by the regions (table 5)is a standard contingency table. The Chi-squared test for independence reveals that the one cannot reject the hypothesis that the table is independent (X-squared = 7.543, df = 4, p-value = 0.1098), that is, that there are no significant differences in the gender distribution by the regions.

The distribution of researchers in most research areas is not significantly different than the distribution of researchers in the political regions, as shown in table 11. The exceptions are the areas of software engineering, which has a lower than expected concentration of researchers in the South, and a higher than expected in the North and Northeast; network and distributed systems, with a higher concentration on the North and Northeast, and hardware and computer architecture with a much higher than expected concentration of researchers in the South.

At least the results regarding software engineering and hardware can be explained by historical specialization. The largest Northeastern Computer Science Department (UFPE) has foster the opening of a large number of startup companies in software development, and

Research topics	p-value
Artificial intelligence	0.7515
Bioinformatics	0.3878
Collaborative systems	0.3952
Computers in education	0.7006
Data base	0.1424
Hardware and computer architecture	0.0027
Human-computer interfaces	0.3414
Image processing and vision	0.4880
Networks and distributed systems	0.0485
Security	0.4883
Software engineering	0.0291
Theory	0.3299

Table 11: Significance of the differences in the distribution of research interests by region.

this has probably lead a lot of local researchers to focus in software engineering, which would not only result in many PhD's in the area that may be hired by the department as faculty, but would also attract software engineering faculty candidates to apply to that department. On the other hand, the largest Southern CS department (UFRGS) had a strong group in hardware and PhD out of that group are probably being hired by the department as faculty.

The productivity of researchers in both journal and proceedings are significantly different in the regions, resulting in X-squared = 39.0387, df = 4, p-value = 6.84e-08, for the goodness-of-fit of the journal production, and X-squared = 53.6579, df = 4, p-value = 6.206e-11 for the goodness of fit test for the proceeding production. Table 12 display productivity for articles in journals and papers in conference proceedings for the researchers in each region. Researchers in the Southeast seems to be concentrating their production in journal articles, at the expense of their production in conference proceedings, where as researchers in the South seems to have achieved a more balanced distribution of their efforts. The lower productivity for the North, Northeast and Midwest researchers may be explained by two factors, the lower supply of graduate students in these regions, and, possibly, a lower critical mass of researchers in each department. Some of the Southern and Southeastern universities and computer science departments are more prestigious and thus attract a lot of graduate students, some from the North, Northeast and Midwest regions. It is likely that researchers from these two regions have more graduate students than average. A second explanation is that, as seen in table 4, there are more researchers per department in the South and Southeast; with less colleagues, the researchers in the lower productivity regions have less chances to collaborate and have a bigger share of the administrative burden.

Finally, the distribution of fellowships (total of both levels) by regions does not follow the distribution of researchers in the regions (p-value = 0.0003396) but it seems to follow the distribution of journal publications (p-value = 0.149). That seems to be consistent with the direction that the CNPq committee have been following of assigning more importance to publications in journals than in conferences.

	Journals	Proceedings
North	2.43	18.25
Northeast	3.53	21.32
Midwest	3.46	8.71
Southeast	**5.81**	19.65
South	5.17	**28.78**

Table 12: Productivity in journals and conferences by political regions (number of papers per researcher from 2000 to 2006).

5.2 Gender differences

Research topics	p-value
Artificial intelligence	0.01248
Bioinformatics	0.7836
Collaborative systems	0.02305
Computers in education	0.0002661
Data base	0.3882
Hardware and computer architecture	3.262e-06
Human-computer interfaces	0.0045
Image processing and vision	1
Networks and distributed systems	0.006574
Security	1
Software engineering	0.2263
Theory	0.5575

Table 13: Significance of the differences in the distribution of research interests by gender.

Table 13 list the p-value of the goodness-of-fit of the proportion of male and female researchers in the research topics, with the baseline of the total number of male and female researchers. The table shows that there is no significant differences between the distribution of the gender of the researchers in the areas software engineering, image processing, theory, security, bioinformatics, and data base. The areas of artificial intelligence, computers in education, collaborative systems, and human computer interfaces have a statistically significant higher proportion of females than the general ratio of females to males. And the areas of network and distributed systems, hardware and computer architectures are "male" research areas.

Regarding production, there is no significant difference in the distribution of the journal article production given the proportion of the gender of the researchers (p-value = 0.1805), but there is a small but significant difference in conference proceedings (p-value = 0.04495). Thus female researchers have a 22.11 productivity regarding the number of papers published in the 7 years considered in this research, while male researchers had a 21.37 papers per researcher.

But the distribution of females are not evenly distributed when we consider the re-

searchers ranked by their production (articles in journals). For example, among the 30 most prolific researchers, only two are females (or 6.7%). Among the 100 most prolific researchers, 17 are females; among the top 250, 47 (or 18.9%) are females. Only among the top 400 is the female proportion similar to the proportion on the whole population of researchers.

Finally regarding the fellowships, there is no significant differences in the distribution of the fellowships of level 1 or the total number of fellowships across genders (p-value = 0.1969 and p-value = 0.2011 respectively), but for fellowships of level 2 the differences are significant (p-value = 0.02199). That is, while 12% of the male researchers have a level 2 fellowship, 19% of the female researchers have them, and the difference is statistically significant.

6 Conclusions

The field of computer science in Brazil seems to be reasonable egalitarian regarding gender, given the low number of female computer scientists. Female scientists tend to concentrate in the areas of artificial intelligence, collaborative systems, computer in education, and human-computer interfaces, areas in which the "human component" is more salient. And females tend to avoid areas such as hardware and networks, in which the "technological component" is more relevant. That seems consistent with research such as [28] which show that men prefer more technological activities, although that work is not about research choices.

Women are as recognized as "major researchers" as men, at least using the fellowship measure - in fact women slightly but significantly more likely to receive such a grant than man, contrasting an opposite result for astronomy, oceanography and immunology reported in [23].

The "productivity puzzle" also is not apparent if we take the whole CS community - women are at least or even more productive than men. But if we take the top 30 or the top 100 most prolific CS researchers in Brazil, women are not as present in these groups as their proportion in the whole CS community. It would be interesting to find out reasons for such disparity, beyond the ones raised by [1] regarding the demands of family versus work life. A possible explanation is that the most prolific researchers are more likely the older ones, people that had the time to create large research groups, and maybe women are newcomers to the Brazilian CS community. This hypothesis can be evaluated using the Lattes systems.

Regarding regional differences, there are some significant differences in productivity in the different regions, and some differences in the concentration of researchers in a few research topics. We do not believe that the two disparities are related - that the research topics of favor in the less productive regions are "more difficult" and thus result in less publications. We believe that there are other, more likely explanations. We raised the hypothesis that the Southern and Southeastern departments attract more graduate students, reducing the availability of graduate students in the other regions. A second hypotheses is that because of a lower number of researchers per department, faculty in the lower productivity regions have a higher administrative burden, and are not as productive as their Southeastern and

Southern counterparts. Finally, a third hypothesis, related to the productivity puzzle, is that older researchers are in general more productive - and some of the computer science departments in the most productive regions are among the oldest in Brazil, and thus more likely to harbor the older researchers. As we mentioned, this hypothesis can be evaluated using the Lattes system, and is a important future continuation of this work.

Finally, the regional differences regarding emphasis in some research topics is an interesting phenomena, which deserves some further research.

6.1 Limits of this research

In this research we used a limited definition of "computer science researcher", as faculty associated to departments that grant a graduate degree in computer science and which are accredited by CAPES with a grade 3 or more. There are also computer science researchers in other departments, for example, computer engineering, electronic engineering, applied mathematics and maybe mecatronics. Despite working on these other departments some of these researchers would classify themselves as computer scientists and would be recognized as such by the CS community. Expanding our research to encompass them would be an important future step of this work.

It is also possible that researchers in Computer Science are also present in non-graduate granting departments, or in government and industry research centers. It is very likely that recent PhDs have been hired by colleges and departments that do not grant a master or PhD degrees. Such recent faculty will probably try to maintain their research efforts even in colleges and departments in which research is not as valued. Finally there are a few government research centers, which have research in more applied aspects of computer science.

References

[1] J. R. Cole and H. Zuckerman. The productivity puzzle: persistence and change in patterns of publication of men and women scientists. In P. Maehr and M. W. Steinkamp, editors, *Advances in Motivation and Achievement*, volume 2, pages 217–258. JAI Press, 1984.

[2] Elba Mauleon and Maria Bordons. Productivity, impact and publication habits by gender in the area of material sciences. *Scientometrics*, 66(1):199–218, 2006.

[3] Denise Gurer and Tracy Camp. An acm-w literature review on women in computing. *SIGCSE Bull.*, 34(2):121–127, 2002.

[4] J Bernhard. Gender-related attitudes and the development of computer skills: A preschool intervention. *The Alberta Journal of Educational Research*, 38(3):177–188, 1992.

[5] J. McGrath Cohoon. Toward improving female retention in the computer science major. *Commun. ACM*, 44(5):108–114, 2001.

[6] Tracy Camp. The incredible shrinking pipeline. *Commun. ACM*, 40(10):103–110, 1997.

[7] M. K. Ahuja. Women in the information technology profession: a literature review, synthesis and research agenda. *European Journal of Information Systems*, 11(1):20–34, 2002.

[8] Wendy Cukier. Constructing the IT skills shortage in Canada: the implications of institutional discourse and practices for the participation of women. In *SIGMIS CPR '03: Proceedings of the 2003 SIGMIS conference on Computer personnel research*, pages 24–33, New York, NY, USA, 2003. ACM Press.

[9] K. D. Joshi and Kristine Kuhn. Examining the masculinity and femininity of critical attributes necessary to succeed in IT. In *SIGMIS CPR '05: Proceedings of the 2005 ACM SIGMIS CPR conference on Computer personnel research*, pages 32–35, New York, NY, USA, 2005. ACM Press.

[10] Mary Sumner and Kay Werner. The impact of gender differences on the career experiences of information systems professionals. In *SIGCPR '01: Proceedings of the 2001 ACM SIGCPR conference on Computer personnel research*, pages 125–131, New York, NY, USA, 2001. ACM Press.

[11] Cynthia K. Riemenschneider, Deborah J. Armstrong, Myria W. Allen, and Margaret F. Reid. Barriers facing women in the IT work force. *SIGMIS Database*, 37(4):58–78, 2006.

[12] Andrea Hoplight Tapia. Hostile work environment.com. In *SIGMIS CPR '03: Proceedings of the 2003 SIGMIS conference on Computer personnel research*, pages 64–67, New York, NY, USA, 2003. ACM Press.

[13] Michael J. Gallivan and Raquel Benbunan-Fich. Examining the relationship between gender and the research productivity of is faculty. In *SIGMIS CPR '06: Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research*, pages 103–113, New York, NY, USA, 2006. ACM Press.

[14] Suresh Kumar and K. C. Garg. Scientometrics of computer science research in India and China. *Scientometrics*, V. 64(n. 2):p. 121–132, 2005.

[15] Jiancheng Guan and Nan Ma. A comparative study of research performance in computer science. *Scientometrics*, V. 61(n. 3):p. 339–359, 2004.

[16] Wolfgang Glanzel, Jacqueline Leta, and Bart Thijs. Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, 67(1):67–86, 2006.

[17] Jacqueline Leta, Wolfgang Glanzel, and Bart Thijs. Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics*, 67(1):87–105, 2006.

[18] M. Krauskopf, Maria Ines Vera, Vania Krauskopf, and A. Welljams-Dorof. A citationist perspective on science in Latin America and the Caribbean, 1981-1993. *Scientometrics*, 34(1):3–25, 1995.

[19] Amalia Mirta Calvi no. Assessment of research performance in food science and technology: Publication behavior of five Iberian-American countries (1992-2003). *Scientometrics*, 69(1):103–116, 2006.

[20] Farideh Osareh and Concepcion S. Wilson. Third World Countries (TWC) research publications by disciplines: A country-by-country citation analysis. *Scientometrics*, 39(3):253–266, 1997.

[21] Ivan Figueira, Raphael Jacques, and Jacqueline Leta. A comparison between domestic and international publications in Brazilian psychiatry. *Scientometrics*, 56(3):317–327, 2003.

[22] Adriano S. Melo, Luis Mauricio Bini, and Priscilla Carvalho. Brazilian articles in international journals on Limnology. *Scientometrics*, 67(2):187–199, 2006.

[23] Jacqueline Leta and Grant Lewison. The contribution of women in brazilian science: A case study in astronomy, immunology and oceanography. *Scientometrics*, V. 57(n. 3):p. 339–353, 2003.

[24] Liming Liang and Ling Zhu. Major factors affecting China's inter-regional research collaboration: Regional scientific productivity and geographical proximity. *Scientometrics*, 55(2):287–316, 2002.

[25] Rickard Danell and Olle Persson. Regional R&D activities and interactions in the Swedish Triple Helix. *Scientometrics*, 58(2):203–218, 2003.

[26] IBGE. Brazilian institute of geography and statistics. http://www.ibge.gov.br/home/, 2007.

[27] PNUD. United nations development programme. http://www.pnud.org.br/, 2007.

[28] T. Klief and W. Faulkner. Boys and their toys: Men's pleasures in technology. `http://www.rcss.ed.ac.uk/sigis/public/backgrounddocs/BOYS_AND_THEIR_TOYS-ZIF8.rtf`, 2002.