Visual Analisys of Educational Data: A Gender Study in Computer Courses in University of Brasilia

Luiza Hansen
Dept. of Computer Science
University of Brasilia
Brasilia, Brazil
luizaahansen@gmail.com

Vinicius Borges
Dept. of Computer Science
University of Brasilia
Brasilia, Brazil
viniciusrpb@unb.br

Aleteia Araujo
Dept. of Computer Science
University of Brasilia
Brasilia, Brazil
aleteia@unb.br

Maristela Holanda
Dept. of Computer Science
University of Brasilia
Brasilia, Brazil
mholanda@unb.br

***Abstract*—The presence of women in technology-related courses is declining every year, reaching in 2016 less than 20% of the total student body in the Department of Computer Science in UnB (*Universidade de Brasília*). This paper uses visualization techniques to analyze and identify profile patterns in girls on undergraduate courses in the computing field. Dimensionality reduction technique (PCA), HeatMap and Parallel Coordinates were used for the visual data analysis process, considering the students' situation in relation to UnB (active, drop out or graduated). In this work, the existing correlations between variables were evidenced, with more in-depth analysis of the association between entrance period and the nature of the university departure form and period of the university attendance. Also, the students participating in the quota schema were analysed and the study suggests that there is no correlation between students enrolled under the quota system and the form of departure from the course.**

***Index Terms*—women, computer, visualization, PCA, HeatMap, Parallel Coordinates**

I. Introduction

Women in Computer Science is an important research topic, since only 18% of female students in Brazil completed programs in different computer courses in the year 2016 [1] according to SBC (*Sociedade Brasileira de Computação*) which has analysed INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*) data, obtained from the Census of 2016 Higher Education. In UnB (*Universidade de Brasília*) the Computer Science, Computer Engineering and Licenciate in Computer courses have a small percentage of women compared to men, less than 20% in 2016.

Ada Lovelace, ENIAC girls, sister Mary Kenneth Keller and Grace Hopper are great women who influenced the development of the technology used nowadays. Compared to other courses, Computer Science was in the early days a promising area for women, having more female participation than courses such as physics and medicine [2]. However, since the 1990s, the participation of women in computer programs has been decreasing.

To modify this scenario, there are some important initiatives, such as IEEE resources to support women in tech and offer many initiatives to empower their career development and salary potential [3]. CLEI LAWCC (*Latin American Women in Computing Congress*) is a congress to highlight the research, interests, and achievements of women in the various areas of computing, with the aim of encouraging the active participation of more women [4]. Also, in Brazil, there is the CSBC (*Congresso da Sociedade Brasileira de Computação*) which has been conducting WIT (*Women in Information and Technology*) workshop since 2007. The SBC is still involved with the project "Meninas Digitais" the goal of which is to discuss the low participation rate of women in the field of computing [5], developing several projects all over Brazil. In addition, government organisations such as MCTI (*Ministério de Ciência, Tecnologia e Inovação*) have released guidelines for specific research projects for the education of girls in Exact Sciences and Computer courses [6]. In 2005, the "Instituto Unibanco" released "Gestão Escolar para Equidade: Elas nas Exatas" (School Management of Equality: Female Students for STEM) which aims to reduce the impact of gender disparity in Exact Sciences [7]. In 2018, CNPq released "CNPq/MCTIC Nº 31/2018 - Meninas nas Ciências Exatas, Engenharias e Computação"(*"CNPq / MCTIC number 31/2018 - Girls in the Exact Sciences, Engineering and Computing"*).

Data mining techniques have been used to discover patterns and implicit knowledge in educational databases. In addition, they have been used in studies that analyze the profiles and characteristics of women in exact courses, assisting education specialists. The use of visualization techniques can benefit these tasks, since the human visual system identifies and interprets relevant patterns in graphic representations more quickly than descriptive or exploratory analysis [8].

The aim of this article is to analyse the data of women on computer courses at UnB, using visual techniques based on HeatMap, parallel coordinates and scatterplot, considering the results obtained in Principal Component Analysis (PCA), employed to reduce the dimensionality of the database.

The remainder of this paper is structured as follows: Section II briefly describes the visualization techniques; Section III presents work related to this research; Section IV describes the methodology employed for this research; Section V presents the main outcomes of this research; Finally, in Section VI conclusions and future work are presented.

II. Data Visualization

This section focuses on the main visual techniques used in this paper, HeatMap, Parallel Coordinate and *Principal*

Component Analysis (PCA).

A. HeatMap

Visualization of a scatter plot may be overlapping when data sets have many instances, making interpretation and extraction of information difficult. To work around this limitation, this paper studied the use of HeatMap technique, which represents the geographic density of dot elements on a map, using colored areas to represent them [9].

HeatMap has been used to help detect the degree of correlation of the attributes of data sets. By identifying the principal data attributes, it is possible to analyse and remove some of these, so that in the final analysis there are fewer requisites and in that way, an attribute doesn't affect another one.

B. Parallel Coordinates

Parallel Coordinates consist in multivariate data analysis, where N positive integer, a coordinate system to m-Dimensional Euclidiano R^m space is built. In the parallel coordinate graph, each line represents a variable and the data instances are polylines that intersect these axes in a certain attribute value related to these axes [10]. This graph is appropriate to identify discrepant values or patterns based on metric-related factors and to find crossover points.

C. PCA

The main views generated in this paper use *Principal Component Analysis* (PCA) techniques. These consist of mapping data from a high-dimensional space to a space of reduced size, where the variance is optimum [11]. First, data is centered on the origin of the coordinate system, and then the covariance matrix. Then, a spectral decomposition occurs in the correlation matrix, resulting in achieving eigenvalues and eigenvectors. To get the reduced space with p dimensions, the eigenvectors associated with the highest p eigenvalues are selected. To view PCA data in a bi-dimension plot, consider $p = 2$.

III. Related Work

There have been several studies addressing the gender issue in Computer Science. In this section articles about the decreasing number of women in Computing Major are reviewed.

Stout et al. in [12] and Cheryan et al. in [13] provide studies about stereotyping in Computer majors, also arguing that there is a higher ratio of men than women in this field in the US (*United States*). Likewise, Mercier et al. present surveys, drawings, and interviews which were used to examine the perception of US middle school students about characteristics of knowledgeable computer users [14]. These results showed cultural stereotypes of a computer user: 89% were male and 94% wore glasses.

Keinan et al. show data that the ratio of graduated women from Bachelor programs in Computer Science was almost 40% in 1984, dwindling to 20% by 2006 in the US [15]. Vardi has similar results, and adds that in 2013 and 2014, only 14.7% of those graduates were women [16].

Papastergiou used descriptive statistics, principal component analysis and analysis of variance in [17] to investigate 358 Greek high school students' intentions and motivation for pursuing academic studies in CS (*Computer Science*). This study looked into several factors, such as the influence of family and academic environment on their career choices, their perception of a professional career in CS, and their self-efficacy beliefs regarding computers. The analysis showed that a lack of exposure to and use of a computer at home and in school from early stages in the students' lives seems to be the main factor in discouraging them from studying CS, especially in the data for girls.

Anderson et al. applied means, Mann-Whitney U test comparison, and non-parametric statistics in [18] to study possible factors related to low rates of female participation in education pathways leading to information and communications technology (ICT) professions, considering data from a three-year period. The survey, which had binary options, such as "I am very interested in computers" and "I am not interested in computers", was presented to 1,453 high school girls in their Senior year. The study identified two factors associated with a woman's aversion to ICT careers: the perception that the subject is boring, and an intense dislike of computers.

Maia describes a similar situation in [19], presenting a study on female participation in university majors in Computing in Brazil, based on the Higher Education Census data from the Ministry of Culture and Education between the years of 2000 and 2013. One of the issues raised was that while the number of male graduates increased 98% in the period, that of female graduates decreased 8%.

Lagesen [20] analysed four inclusion ways to recruit more female students to Computer Science. These strategies were: achieve a critical mass, educational reform, redefine the gender symbolism of computer science and change the content of the discipline. The results show that increasing the number of recruits is the main strategy to change the course perception from "male" to "neutral". Besides that, the increase of women in the course seems to cause an improvement in the learning environment, since problems associated with being in a minority group (too much visibility, undesired attention) become less present.

Nunes et al. [21] map some Brazilian initiatives that encourage woman to enter or remain in computer courses. They used a similar methodology to Systematic Mapping (SM), with the following steps - definition of the research questions and search for the initiatives, definition of the inclusion and exclusion criteria and classification of the initiatives and their analysis.

Chopra et al. [22] apply deep learning, text mining and statistical methods to unique academic datasets. Over 30,000 applications to undergraduate engineering programs at a large North American institution were analysed. Their goal was to determine whether female applicants express different reasons for applying to an engineering program, and whether female applicants have different technical and extracurricular backgrounds.

Hansen et al. [23] aimed to identify relevant patterns using a visual technique named Andrews Plot considering three profiles of girls: active students, disconnected students and graduated students, concluding that those who left without graduating have similar features with those who graduating. Holanda et al. [24] also study woman profile in UnB, presenting a summary of the profile analyzing the three indicators: entering the courses; student dismissals; and academic achievement. In academic performance, boys and girls have a close average of grades, but looking at all the subjects of the courses female students have, on average, slightly better performance than male students.

Borges et al. [25] describes a method for students' performance prediction using principal components analysis (PCA) and classifiers Support Vector Machines and Naive Bayes. Experiments were performed to compare the prediction performances between the high (original) and the reduced dimensional datasets. The students' data reveal some personal and academical variables that might influence their performances, such as failures, period grades, mother's education and alcohol consumption. Also, results indicated that the dataset with reduced dimensionality retained the most relevant information and relations among attributes.

Sax et al. [26] present a study in the USA (*United States of America*) with survey data on college students during 4 decades to: (a) document trends in aspirations to major in computer science among undergraduate women and men; (b) explore the characteristics of women and men who choose to major in computer science and how this population has evolved over time; and (c) identify the key determinants of the gender gap in the selection of computer science majors. The result revealed across all years, that men were more likely than women to major in computer science. The representation of women reached its peak of 44% in 1980. Women's share of computer science majors dropped in the early 1980s and has been on a nearly constant decline since then. By the mid-2000s, the number of women who were computer science majors had dropped to 12%, and this number rose only slightly to 15% by 2011.

Given this alarming decline and the widening disparity between male and female representations in the field of Computer Science, this paper aims to investigate female students in Computing Majors at UnB in Brazil. Visualization techniques have been used to analyze and identify profile patterns in female undergraduate students.

IV. Methodology

With the suggested goals in mind, the execution of this research included the adapted phases Knowledge Discovery in Databases (KDD) [27]. The original methodology involves the steps: selection, pre-processing, transformation, data mining, interpretation/evaluation, as shown in Figure 1. In this paper, the data mining step was replaced by data visualization. These steps are performed interactively and iteratively, as they involve the cooperation of the person responsible for data analysis and the fact that this process is not applied sequentially but requires repeated parameter selections [28].

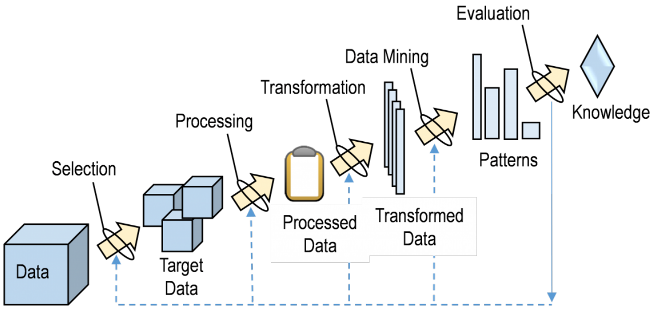

Figure 1. Methodology for knowledge discovery process.

The selection step, as a way to include the main university computing courses, define that Computer Science, Computer Licentiate, Mechatronics Engineering, Communication Network Engineering, Software Engineering and Computer Engineering would be analysed.

The cleaning phase, also known as the pre-processing phase, includes searching for data discrepancy. Thus, examination in a programming language called R ensured there was no lack of information, redundancy or discrepancy and noise [29].

Also in the pre-processing phase, a few corrections of code mismatch were made considering that the data was in Portuguese. Also, it was necessary to merge "Engenharia Mecatrônica" (Mechatronics Engineering) and "Engenharia de Controle e Automação" (Control and Automation Engineering), since they are considered the same course. Moreover, under the category for "drop out" the following situations were aggregated: new entrance examination, the different forms of disconnection, change of course, transfer, entrance examination for another qualification, shift change and three consecutive reprimands in the same compulsory discipline, reducing the scope of variables of the form of exit of the course.

A relational database was built in the transformation phase as shown in Figure 2, with the information category split into three tables: the "Student" table has information about enrolled students, like gender, race, school attended, etc; the "Subject" table has information like name and number of credits in each subject; Finally, the "Student_Subject" table has information about the student score in that subject, the average in the semester, etc. The analysis data in this first phase is related to the "Student" table, although the "Student_Subject" table is also important and should be considered for further analysis.

After the transformation phase, female students were filtered, continuing to the visualisation phase, since this paper focuses on the profile of female students in computer courses. At the end, there was a total of 789 female students, 117 from Computer Licentiate, 205 from Computer Science, 118 from Mechatronics Engineering, 196 from Communication Network Engineering, 90 from Software Engineering and just 63 from Computer Engineering.

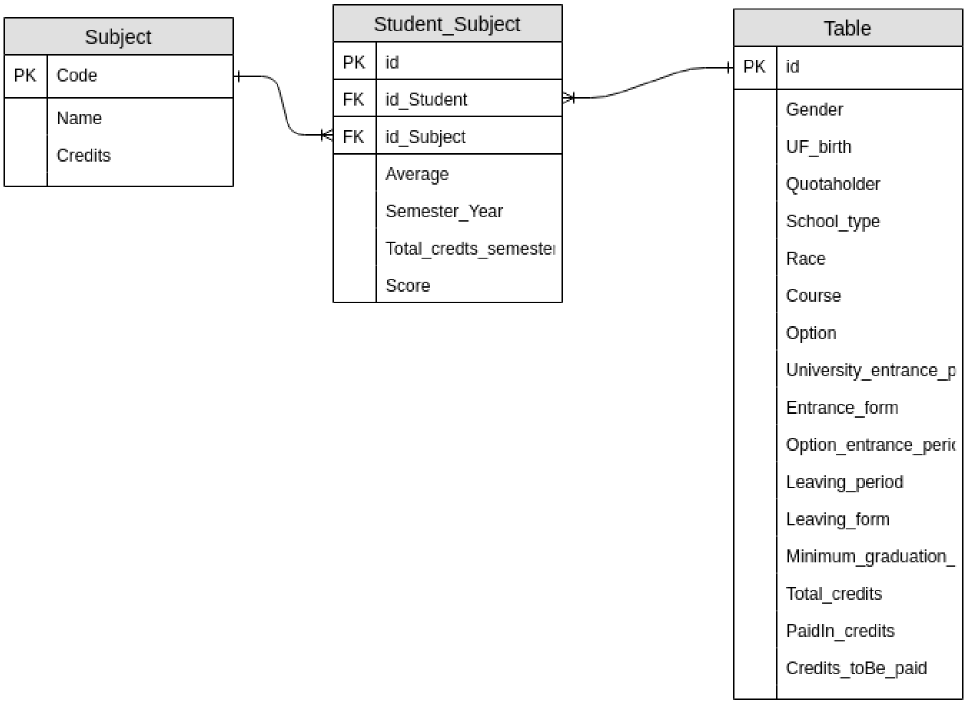

Figure 2. Database model.

At the visualization phase, PCA, Parallel Coordinates and HeatMat techniques were used. HeatMap is important because it emphasizes the relation between contexts and data sources. This algorithm helps extract numerical and unique information bites, even in a large database. Also the Parallel Coordinate technique aims to facilitate the understanding of complex trends and identify small gaps more successfully. The advantage of scatterplot, like PCA, is that it directly represent bivariate relationships (eg dependency, association, *outliers*) between distinct variables, where gaps can be identified.

V. Results

Is summarised, in this chapter, the results obtained from the data, focus on the resolution of the hypotheses raised in Chapter I. For this, traditional visualization techniques such as HeatMap and Parallel Coordinates and visualization techniques based on dimensionality reduction such as PCA. These techniques were used to explore the data and find implicit patterns, since the data present has many dimensions.

In PCA technique, eigenvalues and eigenvectors are important information for evaluating existing patterns in the data and their attributes. Each eigenvector is associated to an eigenvalue, thus, the Principal Components (PCs) space is represented by eigenvectors associated to the highest data variance eigenvalues. Figure 3 illustrates that the first major component concentrates the largest variance, in other words, it is responsible for the maximum variability of the data. The second, third and fourth components, in the case of the evaluated data, concentrate slightly different variances, but explain a large part of the data variances.

The variance field represents the variability of the data and cumulative variance represents the sum of these variances, both in percentages. Table I and Table II show that the variance accumulated in the first 7 major components accounts for 94% of the variability of the data. If we chose to use these components, we would be reducing the number from 11 to 7 latent variables, losing less than 6% of the data variability information.

Table I
SEVEN FIRST PRINCIPAL COMPONENTS.

PC	standard deviation	variance (%)	cumulative variance (%)
1	3.323	30.214	30.214
2	1.823	16.57	46.788
3	1.575	14.32	11.062
4	1.2168	11.062	72.171
5	0.935	8.502	80.67
6	0.901	8.197	88.87
7	0.607	5.523	94.39

Table II
FOUR LAST PRINCIPAL COMPONENTS.

PC	standard deviation	variance (%)	cumulative variance (%)
8	0.458	4.166	98.56
9	0.153	1.391	99.95
10	0.0052	0.047	100
11	2.5e-31	2.3e-30	100

The perceptual map is represented by a diagram that describes the relationship between elements of the same category with different characteristics [30]. It is worth emphasizing that the visualization of the perceptual map can approximate some

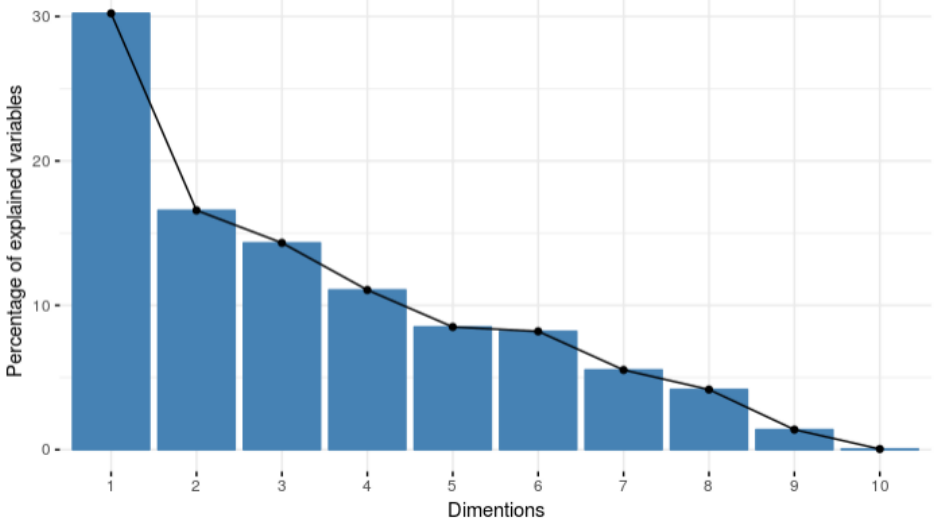

Figure 3. Percentage of explained variance by each principal component.

attributes taking into account the correlation between them, as well as their contribution and variation pattern in association with the components.

Figure 4 represents the perceptual map, which illustrates the contribution of each variable in dimensions 1 and 2 of the Principal Components, the horizontal axis being dimension 1 (Dim 1), and the vertical axis being dimension 2 (Dim 2). It is possible to observe four quadrants in the graph, where the first quadrant has the dimensions 1 and 2 with positive values and the third quadrant with negative values. The second quadrant presents dimensions 1 and 2, respectively, positive and negative, whereas the fourth quadrant represents the inverse.

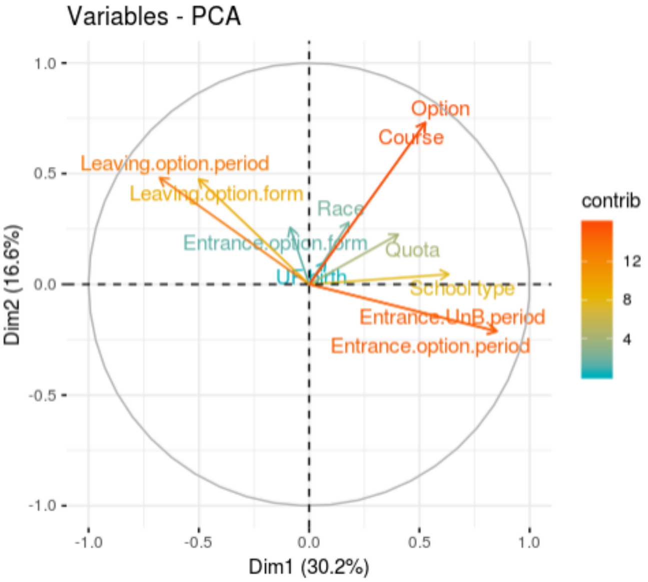

Figure 4. Graph of variables using the Perceptual Map technique.

In the map, the variables that have the same direction have positive correlation, whereas the variables that have opposite directions have negative correlation. Considering two positively correlated variables X and Y, if the values of X increase, the values of Y also increase. On the other hand, if the values of X decrease, the values of Y follow. Now if X and Y are two negatively correlated variables, if the values of X increase, then the values of Y decrease and vice versa. Thus, among the analyzed variables, those that have a positive correlation are:

- period of university entrance and the period of option entrance;
- the course and the option;
- option leaving form and option leaving period.

There are both of these variables - entrance university period and entrance option period - since the student has the possibility to change the course, where the enrollment is maintained, for this reason it is important to have both periods. The variables are equal if the students did not change the course, but they can be different if the students start the university on another course. Therefore, the period of entry into the option and the university have negative correlation with the period of exit of the option.

In order to obtain more detailed information on the variables, the HeatMap technique was used to illustrate the level of correlation between each variable, either positive or negative, and to locate the variables with the highest degree of relationship. N groups of variables were grouped correlated, thus generating the graph in Figure 5, where the colors of each correlation indicate the strength and the sign of correlation, with negative correlations being red and blue being positive.

In UnB there is a quota system that reserves 5% of vacancies for black students of both types of schooling (public or private); and 50% for public school students. Within the last 50%, 25% are intended for students with a gross family

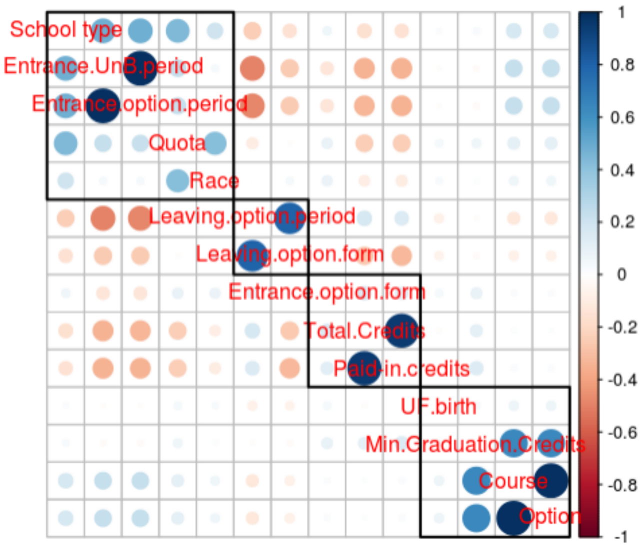

Figure 5. HeatMap grouped by the most correlated variables.

income per capita equal to or less than a salary and a half, and the last 25%, for students with a higher income. Within the division of the quotas of public schools, a part of the vacancies is reserved for the students who declare themselves brown, black or indigenous. This percentage is defined based on the total sum of the population that integrates these groups in each unit of the Federation, according to the last IBGE (*Instituto Brasileiro de Geografia e Estatística*) census.

Observing the ethnic groups, it is verified that the quota variable (whether or not a student has quota) has a positive correlation with the period of admission to the university, once the system of quotas for black people in UnB came into force 15 years ago, which corroborates the correlation obtained between these attributes. Other variables that vary together because they have a positive correlation are related to attributes like quota, race, type of school and the period of entry into the option.

It is observed in the heat map that the course has a positive correlation with the type of school (public or private school), thus indicating that the choice of course is influenced by the high school that the student attended. Another positive correlation observed refers to the minimum credits for graduation with the course, since the student must complete the minimum amount of credits, which varies according to the course, to graduate.

As one of the hypotheses of this work is to investigate the reasons for the female students leaving the computing courses, it is interesting to analyze the attribute related to the leaving form of these students. According to Figure 5, it can be noted that the leaving form has positive correlation with the exit period and negative correlation with both - type of school and period of entry into university and into the option. In this sense, the data referring to the form of exit of the course were classified as: “Formed”, the students who finished the course, “Active”, the students who are still attending, and “Leaving” the students who left the course for some reason other than graduation.

Figure 6 presents the result of the application of the Parallel Coordinates technique, used to analyze the relationship between the period of entry into the option and the final situation of the students, with the option’s leaving form attribute. There is a pattern in the polyline of the chart associated with female students who left the course in two periods: between the first half of 1992 and the second half of 1995, and between the first half of 2001 and the second half of 2016. Data from the first semester of 1996, the first semester of 2000 and the second semester of 2007 until the summer vacations of 2017 are observed. Finally, the active ones are those that do not have an exit period, since the polylines connect to the value zero in the exit period of the option. Figure 7 shows these years in greater detail.

In Figure 8 the relationship of the previously analyzed variables (leaving form and leaving period) with the entry period is evident, both in the UnB and in the option. In this way, it can be seen that the active students entered the option, from 2007 to 2016, since those who graduated entered the option between the first semester of 1991 and the second semester of 2013. Curiously, there are some cases of graduated students who entered the second half of 2015. Considering that computer courses have a minimum of 8 semesters, it is assumed that these students came from other courses and already had credits, enabling graduation in 5 semesters.

VI. Conclusion

This paper presents a visual analysis of undergraduate students in technology, in courses at UnB. It was necessary to collect the data, clean them and apply techniques that can extract important information to develop a greater understanding of the main difficulties faced by women in these courses.

Inside the information visualization, the exploratory view can be understood as a process of hypothesis generation, allowing the user to obtain information to acquire relevant information. For the purpose of this paper, two traditional visual techniques - parallel coordinate and HeatMap - along with one multivariate data visualization technique - *PCA* were applied.

In addition, an analysis of the data of the students on different Computing courses at UnB was carried out. The approach considered the characteristics and the student groups by leaving form, and obtained relevant information about these girls profiles. The data visual analysis employed the technique *PCA* which generate different kinds of graphic representations, such as scatterplot, HeatMap and histogram.

The visualization emphasized the correlation between variables, and grouped the girls by leaving form (active, graduated, detached). This way it was possible to determine some relevant factors about the girls’ profiles, such as the importance of reserved access to those who fit in the profile request, and the importance of the school when choosing a major.

By these means, this paper intends to provide the Computing departments of UnB with the opportunity to take decisions

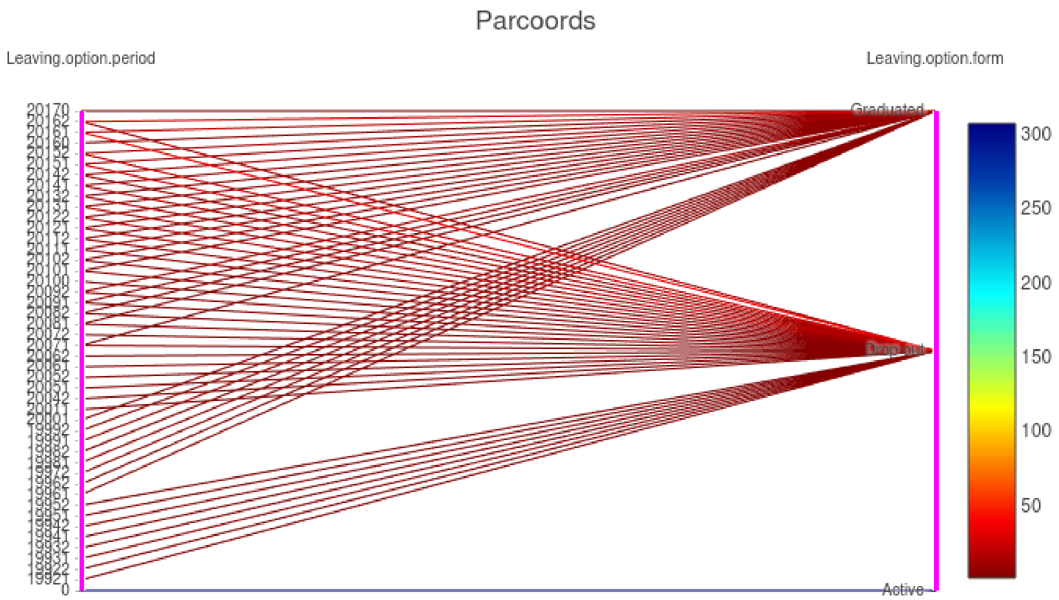

Figure 6. Relation between leaving form and leaving period.

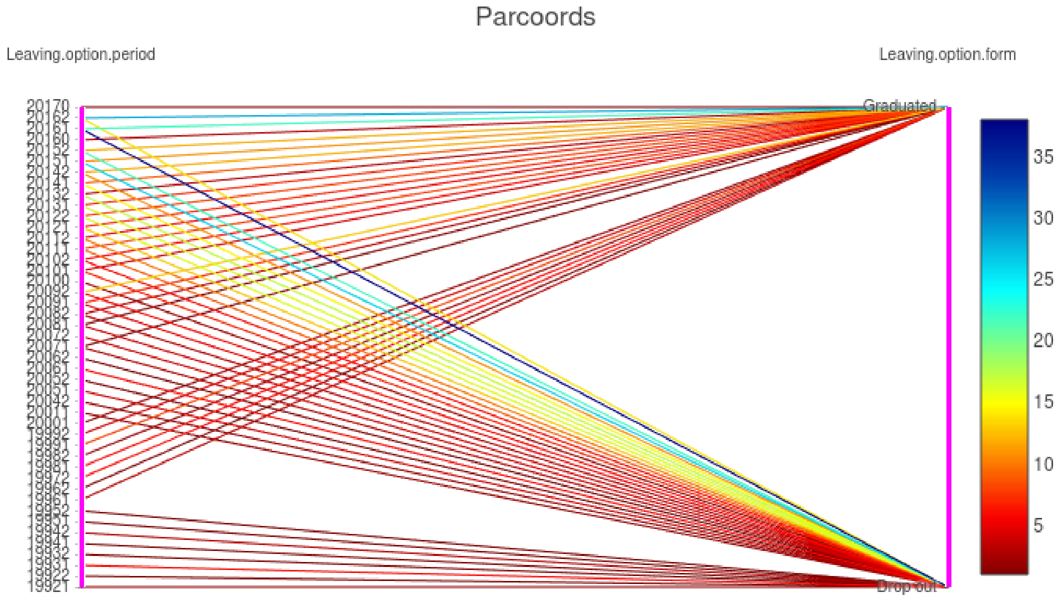

Figure 7. Filter periods

through public politics that provide/develop policies aimed at providing an attractive teaching environment, affordable and inclusive to girls. In this way, the number of female graduates and women in the labour market may be increased

In order to continue this work, the following studies are suggested: The study of other non-linear techniques, with a process of exploratory view to extract implicit and relevant information, focus on correlated variables with leaving form; use the data of each student related to the subjects to have a more accurate analysis, evaluating not just the profile, but

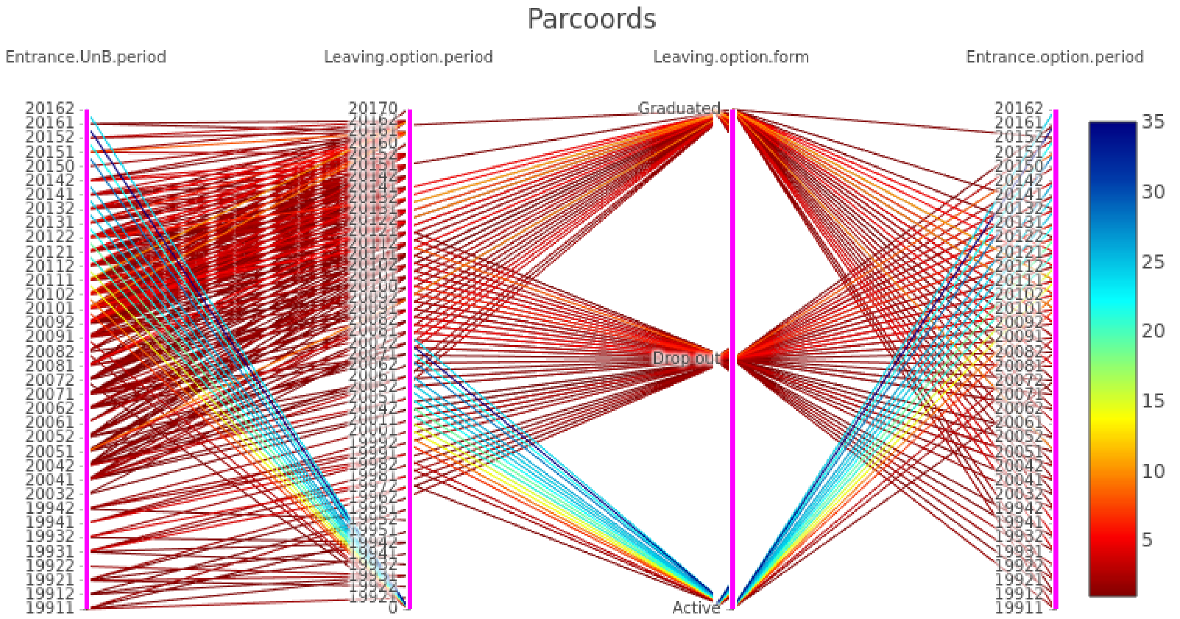

Figure 8. Relationship between leaving form, leaving period, and entrance periods.

also the behavior in the university, as grade in each subject and average of the period; the use of visualization as a support to data mining, might predict and analyse the performance of the computer students, distinguishing between graduated and non-graduated.

References

[1] D. Nunes, “Educação superior em computação, estatısticas 2016”, *Sociedade Brasileira de Computação-SBC.*, vol. 7, 2016.

[2] S. Henn, “When women stopped coding”, *NPR Planet Money*, vol. 21, 2014.

[3] IEEE, Women in Technology. Available in: https://www.computer.org/communities/women-in-computing, 2019.

[4] CLEI, *Clei-lawcc latin american women in computing congress*, CLEI 2019. Available in: http://clei2019.utp.ac.pa/en/#eventos$_e speciales$, 2019.

[5] C. Maciel, S. A. Bim, and K. da Silva Figueiredo, “Digital girls program-disseminating computer science to girls in brazil”, in *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*, IEEE, 2018, pp. 29–32.

[6] CNPq, Edital 18/2013 MCTI/CNPq/SPM-PR/Petrobras – Meninas e Jovens Fazendo Ciência Exatas, Engenharias e Computação, 2013.

[7] I. Unibanco, Edital Gestão Escolar para Equiadae: Elas nas Exztas, 2015. Disponível em: http://www.fundosocialelas.org/elasnasexatas/. Accessed 09-April-2017, 2015.

[8] R. S. Archela, “Imagem e representação gráfica”, *GEOGRAFIA (Londrina)*, vol. 8, no. 1, pp. 5–11, 1999.

[9] W. Deng, Y. Wang, Z. Liu, H. Cheng, and Y. Xue, “Hemi: A toolkit for illustrating heatmaps”, *PloS one*, vol. 9, no. 11, e111988, 2014.

[10] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press, 2015.

[11] I. T. Jolliffe, “Principal component analysis and factor analysis”, in *Principal component analysis*, Springer, 1986, pp. 115–128.

[12] J. G. Stout, V. A. Grunberg, and T. A. Ito, “Gender roles and stereotypes about science careers help explain women and men’s science pursuits”, *Sex Roles*, vol. 75, no. 9, pp. 490–499, 2016, ISSN: 1573-2762. DOI: 10.1007/s11199-016-0647-5.

[13] S. Cheryan, V. C. Plaut, C. Handron, and L. Hudson, “The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women”, *Sex Roles*, vol. 69, no. 1, pp. 58–71, 2013, ISSN: 1573-2762. DOI: 10.1007/s11199-013-0296-x.

[14] E. M. Mercier, B. Barron, and K. M. O’Connor, “Images of self and others as computer users: The role of gender and experience”, *Journal of Computer Assisted Learning*, vol. 22, no. 5, pp. 335–348, 2006, ISSN: 1365-2729. DOI: 10.1111/j.1365-2729.2006.00182.x.

[15] E. Keinan, “A New Frontier: But for Whom? An Analysis of the Micro-Computer and Women?s Declining Participation in Computer Science”, Claremont McKenna College, Tech. Rep. 1466, 2017. [Online]. Available: http://scholarship.claremont.edu/cmc%5C_theses/1466.

[16] M. Y. Vardi, “What can be done about gender diversity in computing?: A lot!”, *Commun. ACM*, vol. 58, no. 10, pp. 5–5, Sep. 2015, ISSN: 0001-0782. DOI: 10.1145/2816937.

[17] M. Papastergiou, “Are Computer Science and Information Technology still masculine fields? High school students perceptions and career choices”, *Computers & Education*, vol. 51, no. 2, pp. 594–608, 2008, ISSN: 0360-1315. DOI: http://dx.doi.org/10.1016/j.compedu.2007.06.009.

[18] N. Anderson, C. Lankshear, C. Timms, and L. Courtney, “‘Because it’s boring, irrelevant and I don’t like computers’: Why high school girls avoid professionally-oriented ICT subjects”, *Computers & Education*, vol. 50, no. 4, pp. 1304–1318, 2008.

[19] M. M. Maia, “Limites de gênero e presença feminina nos cursos superiores brasileiros do campo da computação”, *cadernos pagu*, vol. 46, pp. 223–244, 2016.

[20] V. A. Lagesen, “The strength of numbers: Strategies to include women into computer science”, *Social Studies of Science*, vol. 37, no. 1, pp. 67–92, 2007.

[21] M. Nunes, C. S. Louzada, E. M. Salgueiro, B. T. Andrade, P. Lima, and R. Figueiredo, “Mapeamento de iniciativas brasileiras que fomentam a entrada de mulheres na computação”, in *Anais do XXXVI Congresso da Sociedade Brasileira de Computação-X Women in Information Technology (WIT 2016)*, 2016, pp. 2697–2701.

[22] S. Chopra, M. Mirsafian, A. Khan, and L. Golab, “Gender differences in science and engineering: A data mining approach”, 2019.

[23] L. A. Hansen, L. M. Chagas, V. R. Borges, M. Holanda, *et al.*, “Análise visual de dados educacionais: Um estudo de gênero nos cursos de computação da universidade de brasılia”, in *12º Women in Information Technology (WIT 2018)*, SBC, vol. 12, 2018.

[24] M. Holanda, M. Dantas, G. Couto, J. M. Correa, A. P. F. de Araújo, and M. E. T. Walter, “Perfil das alunas no departamento de computação da universidade de brasılia”, in *11º Women in Information Technology (WIT 2017)*, SBC, vol. 11, 2017.

[25] V. R. P. Borges, S. Esteves, P. de Nardi Araújo, L. C. de Oliveira, and M. Holanda, “Using principal component analysis to support students’ performance prediction and data analysis”, in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 29, 2018, p. 1383.

[26] L. J. Sax, K. J. Lehman, J. A. Jacobs, M. A. Kanny, G. Lim, L. Monje-Paulson, and H. B. Zimmerman, “Anatomy of an enduring gender gap: The evolution of women’s participation in computer science”, *The Journal of Higher Education*, vol. 88, no. 2, pp. 258–293, 2017.

[27] C. N. Costa, J. V. Coutinho, L. H. de Magalhães, and M. A. Arbex, “Descoberta de conhecimento em bases de dados”, *Revista Eletrônica: Faculdade Santos Dumont*, vol. 2,

[28] Â. M. J. Corrêa and H. Sferra, “Conceitos e aplicações de data mining”, *Revista de ciência & tecnologia*, vol. 11, pp. 19–34, 2003.

[29] S. Navega, “Princıpios essenciais do data mining”, *Anais do Infoimagem*, 2002.

[30] G. Santos and V. Silva, “Mapa perceptual como ferramenta para a análise da imagem de destinos turísticos”, *Revista de Turismo Contemporâneo*, vol. 3, no. 2, 2015.