Visual Analysis of Educational Data: a Case Study of Introductory Programming courses at the University of Brasília

Luiza Hansen
Dept. of Comp. Science
University of Brasília
Brasília, Brazil
0000-0002-2848-3441

Maristela Holanda
Dept. of Comp. Science
University of Brasília
Brasília, Brazil
0000-0002-0883-2579

Vinicius R. P. Borges
Dept. of Comp. Science
University of Brasília
Brasília, Brazil
0000-0003-1254-8420

Dilma Da Silva
Dept. of Comp. Science and Eng.
Texas A&M University
College Station, United States
0000-0001-6538-2888

***Abstract*—Data visualization aims to graphically represent information from a given application domain. This approach helps in the analysis and understanding of a data set through the mechanisms of interaction and generation of graphical representations, which emphasize the observation of characteristics and patterns. In this way, this technique combined with visual learning analytic enables the detection of the expected and the discovery of the unexpected. Those have been used in numerous areas such as education, where the motivation is to understand and improve the teaching and learning processes. In the literature, data visualization is used within the educational area to predict performance and identify the student profiles, as well as monitoring educational systems in order to improve the quality of teaching. In this area, introductory computing courses stand out for the high number of students who fail or drop out of these courses. On average more than 30% of students, worldwide, drop out of introductory computing courses. At University of Brasília (UnB) this ratio is greater than 50%, which makes it an appropriate scenario for the use of data analysis and visualization techniques, in order to discover patterns related to the scenario and ways to improve the situation. This paper searches and implements the most used visualization algorithms, according to the literature, in order to assist instructors and educational managers to get information about historical and demographic data related to the course. To evaluate the visualizations, the algorithms were applied in a case study of three introductory computing courses at UnB and were evaluated through a questionnaire applied to instructors and educational managers. The results show that the respondents felt more secure when using familiar algorithms, such as pie charts and bar charts. Among the selected visualizations, sankey chart, treemap, and violin chart were the least known by the respondents. Furthermore, the bar chart was the algorithm where the information was identified quickly and correctly most of the time.**

***Index Terms*—Visualization study, educational courses, introductory computing disciplines**

I. Introduction

Nowadays a great deal of information is constantly being generated, making the acquisition of knowledge and the visualization of data, a challenge. It is therefore important to understand the purposes of visualization techniques that seek to graphically represent data sets, so the visual representation generated may explore the capacity of human perception [1]. Hence, domain specialists and users of several knowledge domains can interpret and understand the spatial relations of the underlying data displayed in the visual outputs, leading to the acquisition of implicit and potentially useful knowledge.

According to the definition, visualization techniques provide several advantages to data analysis such as [2] [3] [4]: understanding information faster, due to the fact that sight is the human sense with the greatest capacity for capturing information in the shortest periods of time; viewing huge amounts of data in an intuitive and cohesive way; discovering atypical values and recognising patterns and relations; and involving the user by conveying a message, which might generate impact, work as an extension of human memory and also assist in the cognitive process.

Due to these several advantages, data visualization is already employed in various areas, such as stock market tracking [5], consulting movie databases [6] and in applications in educational areas [7] [8]. One of the fields of visualization research in an educational context is the constant dropout of students in university majors, an issue that has several consequences, including social, economic and human [9].

In this context, this paper aims to implement a methodology to analyze how visualization algorithms can be applied to an educational context and how instructors and educational managers use these visualizations. For that matter, the algorithms were applied in a case study of introductory computing courses at University of Brasília (UnB). The data used in the visualizations were selected due to the insights generated for student retention in the course. Also, the algorithms were evaluated through a questionnaire applied to instructors and educational managers.

The remainder of this paper is structured as follows: the related works are detailed in Section II. Section III shows the methodology steps followed in this work. The result of the questionnaire applied to evaluate the algorithms is in Section IV, also the discussion of the results are in Section IV. Finally, Section VI concludes the paper.

II. Related Work

In recent years, several papers in the literature have proposed approaches to analyze educational data with the support of visualization. Culligan, et al. [10] created a tool to keep track of students, allowing instructors to identify students at risk of failure or dropout and enabling early interventions. Hegde in [11] used Principal Component Analysis (PCA), a technique for dimensionality reduction, in order to predict student dropout. Zhang et al. in [12] considered Massive Open Online Course (MOOC) data to anticipate if a student would complete the course or not, as well as suggesting new courses according to each profile. Essa & Ayad [13] describe the system *Student Success System* (S3), that features holistic visual analyzes of students' academic progress.

A systematic review of the literature, which aims to obtain the papers related to the topic "visual analytics in the educational context", was carried out in [14]. This study mapped the most used algorithms in the literature, that include: bar chart, line chart, heatmap, pie chart, network, scatter plot, timeline and box plot. Also, according to this literature study, from 128 papers analyzed, only thirty of them used data from courses, in which only seven were introductory.

In this sense, this paper has as a contribution, the definition and the implementation of a methodology for the analysis of visualization algorithms, the validation of the charts with instructors and managers of the area, as also, the use of data from initial programming subjects.

III. Methodology

The methodology of this paper was structured in four phases: definition of the data set, study of the fundamentals of visual algorithms, elaboration and definition of the visualizations and, finally, analysis of the results. During the third phase, a case study was applied in order to validate the considered visualizations.

A. Phase One - Definition of the Data set

The first phase consists of the definition of the data set, in which the goal is to perform a screening of the characteristics, focusing on those that are relevant to decision making. For this purpose, anonymous data from SIGRA (Academic Systems of the University of Brasília) were used.

This data set contains information from students of UnB for the period from the second semester of 1984 to the summer semester of 2020. The data totals 181,491 items of academic and personal information from 8,052 students from the following majors: Computer Science, Computer Science Education, Civil Engineering, Computer Engineering, Electrical Engineering, Forest Engineering, Mechatronic Engineering, Mechanical Engineering, Statistics, and Mathematics.

Data processing is any operation using a data set, which may include collection, reception, reproduction, extraction, storage, among others. Thus, after extracting the information from SIGRA, it was necessary to carry out some analyzes to correct the inconsistencies in the data.

From the actions taken, it is evident: the capitalization of the values, the removal of blank spaces, and treatments for undefined values ("NaN"). The latest treatment of the data was due to a change in the curriculum of the Computer Science, Computer Science Education, Electrical Engineering, Statistics and Mathematics major, where the data reported that students were active in both the old curriculum and the new curriculum. Therefore, a new form of exit called "Curriculum change" was created and these students were declared to leave the old curriculum in the year of change, varying from course to course.

B. Phase Two - Study of the Fundamentals of Visual Algorithms

The second phase studies the fundamentals of visualization, in which the algorithms and tools to be used are delimited. It is necessary to select those that enable the interpretation of existing patterns in the data set and that can assist in the discovery of knowledge.

In this way, Lengler & Eppler in [15] grouped the main algorithms into six categories: data visualization, information visualization, concept visualization, strategy visualization, metaphor visualization and composite visualization. Among the available visualization techniques, the most commonly used in visual analysis processes in the educational context were analyzed, in accordance with Section II. Thus, in this paper two of the categories established by Lengler & Eppler are addressed: data visualization and information visualization. Other visualizations, that focused on the knowledge discovery in educational data, pointed out by the literature, were addressed as well.

The data visualization category consists of presenting the quantitative and schematic data, and aims to simplify its interpretation. This category includes the following algorithms [15]: pie chart, line chart, bar chart, area chart, histogram, scatter plot and box plot. On the other hand, the information visualization category seeks to aid in the understanding, evaluation, and analysis of data, allowing information about its internal structure, causal relationship, and dependencies to be obtained [1]. The following visualization algorithms are part of this category [15]: radar chart, parallel coordinates, timeline, venn diagram, sankey diagram, map graph and treemap. Meanwhile, new algorithms are being developed every day in order to facilitate the understanding and analysis of data [16], such as: heatmap, violin plot, bubble chart, network and density plot.

After investigating the aforementioned visualization techniques, 19 were selected in order to explore their use on educational data. The following aspects were identified in those techniques: the types of data represented in the visualizations and their characteristics, the advantages and disadvantages of each graph, and which graphs are similar, i.e., those that can convey the same message. Table I presents a summary of the studied techniques and an analysis of the appropriate scenario that they can be employed, such as the type of data (tabular, series, text), the data attributes (nominal, continuous, discrete),

among others. The link [https://github.com/luhansen/Visual-analysis-of-educational-data—algorithms] has an example of each studied algorithm.

C. Phase Three - Elaboration and Definition of the Visualizations

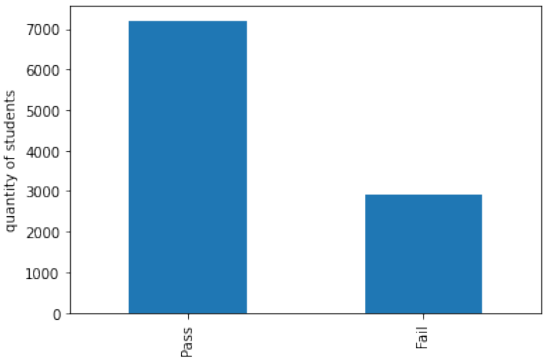
The third phase consists of the formulation, definition and specification of the visualizations. In this phase a case study of introductory computing courses at UnB was applied with the objective of validating and making the necessary adjustments. In order to achieve the main goal of this paper, the following questions were elaborated: 1) How many students pass/fail in the courses? 2) What is the evolution over the semester of the approved students compared to the total? 3) What is the evolution of the students' grades over the semester? and 4) Which major did students most frequently repeat? For the choice of visualization algorithms, some candidates were selected using the study carried out in the second phase of the methodology.

1) Question 1: How many students pass/fail in the courses?: In order to verify if the students usually pass in the courses, the grades were used to generate the visualizations. Grades equal to or greater than 5 went into the "Pass" category, while lower grades were classified in the "Fail" group. The pie chart and bar chart were chosen for their ability to present data with few categories, the visualizations are shown in Figure 1b and Figure 1a. In the charts, 71% of students passed the case study courses.

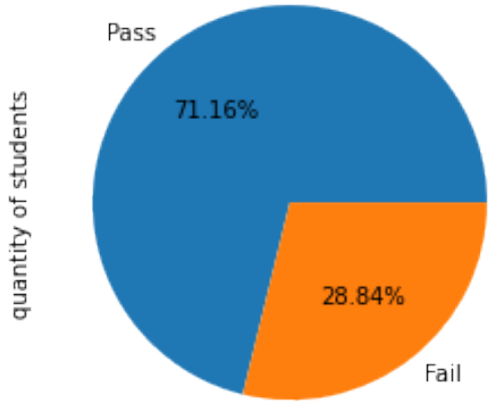
2) Question 2: What is the evolution over the semester of the approved students compared to the total?: This question intends to analyze the evolution over time and to compare the number of students who passed with the total amount who applied for the course. By using two categories (students who passed and the total amount), the line, area and bar charts were selected, as they show the evolution over time. In the visualizations in Figure 2a, Figure 2b and Figure 2c, fewer students took the course in the summer semesters, but in those semesters the proportion of students who passed is higher. The semester with the highest number of students attending the courses was in 2015-2.

3) Question 3: What is the evolution of the students' grades?: In order to visualize the evolution of the students' grades, the semester information and the respective grades were used. Line chart, area chart and bar chart were chosen for the time display. Regarding the box plot, by plotting several on a cartesian plane, one gets the idea of progress, and more information is also highlighted, such as: location, dispersion, asymmetry, tail length, and outliers. From the visualizations in Figure 3a, Figure 3b, Figure 3c and Figure 3d, the average varied, but the oscillation was not significant. The highest average was in the second semester of 2016.

4) Question 4: Which major did students most frequently repeat?: To answer this question, the following data were used: the students' major, the students' score and the number of times the student took the course. The violin chart and the box plot organize large amounts of numerical data, so they were used as candidates to compare the grades among newcomers and repeating students in the course. The parallel coordinates algorithm [17] was selected as a candidate for comparing data values that are of different types and magnitudes. Finally, the bar chart was used in conjunction with the line chart, where the lines represent the average grades of the students in each course, while the bars show the number of students, both segmented by the category "time to attend" .



(a) Bar chart.



(b) Pie chart.

Fig. 1: Generated visualizations to address Question 1.

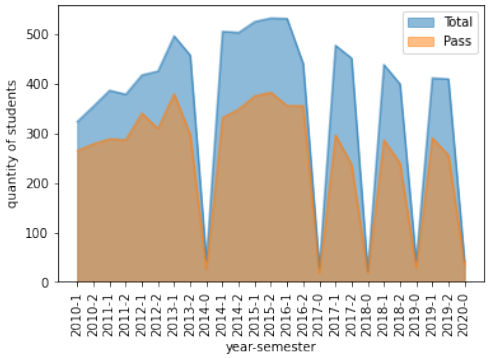
According to the analysis of the visualizations in Figure 4b, Figure 4c, Figure 4d and Figure 4a, the Mathematics major has the largest number of students taking the courses in the case study. The major with the highest grades is Civil Engineering, followed by Electrical Engineering and Mechanical Engineering. First-time students usually get higher grades than repeat students, except for Mathematics students. Furthermore, the Computer Science major has the highest percentage of students retaking the course, as well as being the course with the lowest average.

D. Phase Four - Analysis of the Results

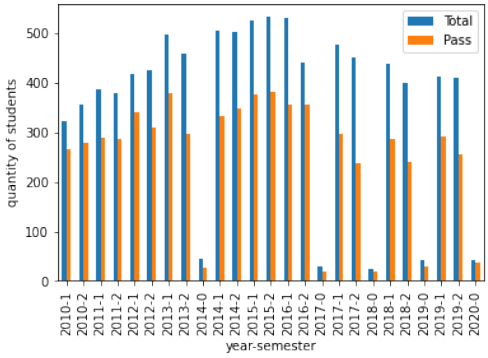
In order to determine whether the selected visualizations communicate information clearly and cohesively, a question-

TABLE I: Scenario analysis of the studied visualization algorithms.

Algorithm	Scenario analysis
Pie chart	- visualize the relationship between the values - visualize relationship of a single value to the total
Line chart	- visualize trends or changes over time
Bar chart	- compare values
Area chart	- compare volume trends over time - emphasize the change over time - draw attention to the total value between a trend.
Histogram	- show a frequency distribution
Scatter plot	- check if there is a cause-effect relationship between two variables
Box plot	- compare the range and distribution of numerical data groups
Radar chart	- compare members of a dimension in a function of several metrics
Parallel coordinates	- show comparison of data elements in a grouping
Timeline	- show a sequence of events in chronological and linear order
Venn diagram	- show relationships between data sets - organize information
Sankey diagram	- show specific quantities - find the most significant contributions to an overall flow
Data map	- compare values and show categories across geographical regions
Treemap	- work with large amounts of hierarchically structured data
Heatmap	- identify patterns
Violin plot	- compare the distribution of a variable
Bubble chart	- show the relationship between numerical values
Network	- show the components of a network and how they interact
Density Plot	- describe a numerical variable



(a) Area chart.



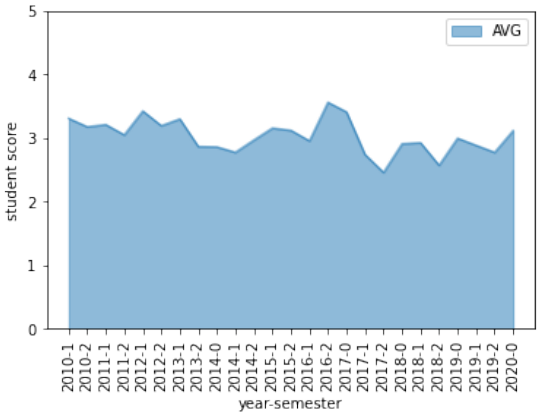
(b) Bar chart.

(c) Line chart.

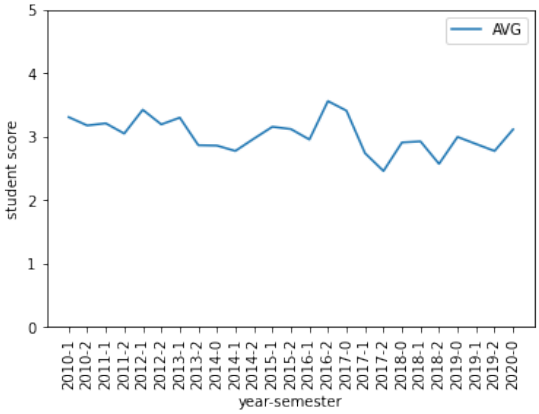
Fig. 2: Generated visualizations to address Question 2.

naire was developed to be applied to instructors and education managers using as basis the Valiati [18]. The respondents have academic degrees varying from computer/engineer, mathematics, physics and statistics. The questionnaire has two types of questions: initial questions that aim to analyze the respondent's knowledge of the visualizations; and elaborated questions, which aim to determine if the candidate visualizations communicate the right information in a clear and practical way. The elaborated questions were developed based on the questions/tasks used in the interaction tests of Valiati.

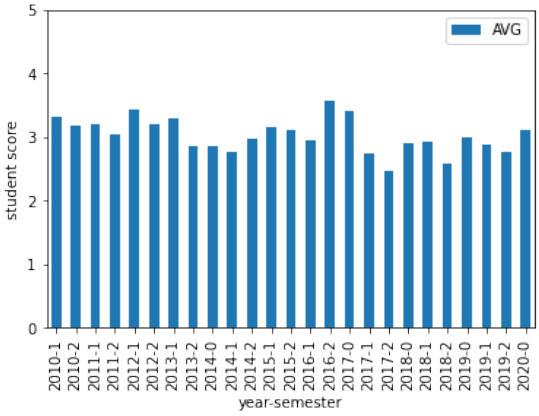
The evaluation of the visualizations consisted in the analysis of the answers to the questions, taking into consideration: the time that the respondent takes to identify the information and answer, also the correctness of the answer. These were classified as "correct" if the answer is as expected, "wrong" if the answer is different than expected, and "don't know" when the respondent chose not to answer because he could not identify the information. Regarding the solution time, the following intervals were defined: 1 to 4 seconds, 5 to 9 seconds, 10 to 15 seconds, and over 15 seconds. The results of the questionnaire application are discussed in the next section.



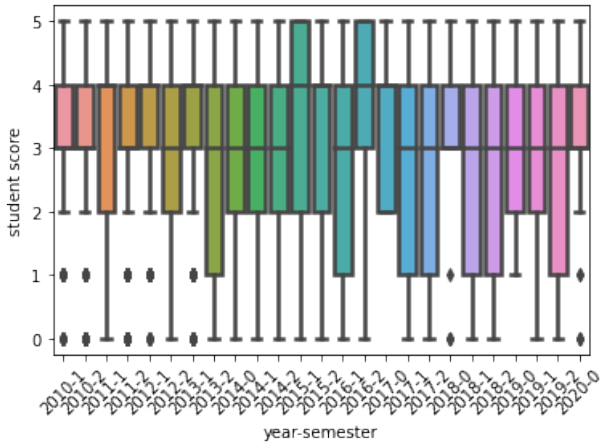
(a) Area chart.



(b) Line chart.

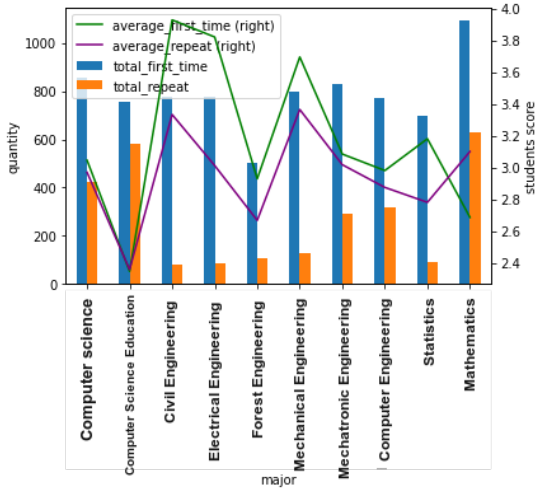


(c) Bar chart.



(d) Box plot.

Fig. 3: Generated visualizations to address Question 3.



(a) Bar chart with line chart.

IV. Questionnaire Results

To evaluate the communication of information through the visualizations, a questionnaire were developed, as specified in Section III-D. A total of 23 instructors and educational managers from computer course took part in the questionnaire, where each question was asked to eight different participants. The choice of question was random.

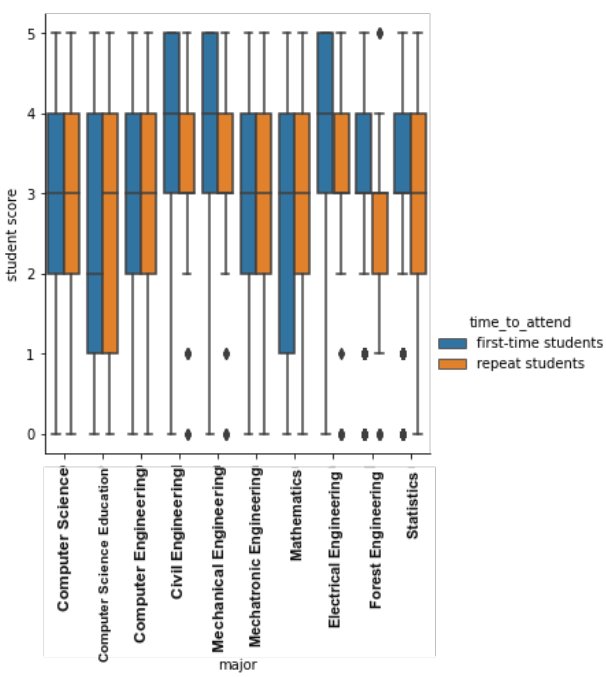
A. Initial Question

From the initial questions it was possible to evaluate the respondents' level of knowledge about the visualizations used. In this sense, the least known graph was the violin graph, followed by the parallel coordinates. However, everyone was already familiar with the bar chart, the pie chart, and the box plot. Table II presents the number of instructors/educational managers who knew, or did not know, each of the visualizations used in the questionnaires.

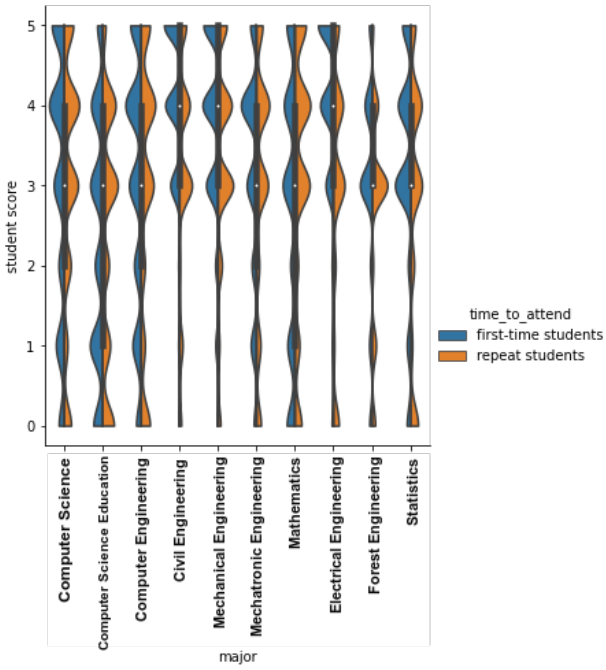
B. Elaborated Question

The elaborated questions, unlike the initial questions, vary for each domain. Thus, this section details the algorithms used in the development of each chart, as well as the results of the application of the questionnaire related to each question in phase three. These results are presented in Table III.

1) Question 1: How many students pass/fail in the course: In order to evaluate the visualizations, the elaborated question related to the first question was: "What is the difference



(b) Box plot.



(c) Violin plot.

TABLE II: Respondents' level of knowledge about visualizations.

	Known	Didn't Know	Total of respondents
Pie Chart	16	0	16
Bar Chart	23	0	23
Line Chart	14	1	15
Parallel Coordinates	7	16	23
Box Plot	15	0	15
Violin Chart	2	13	15
Area Chart	5	2	7

between the amount of students who passed and who failed the course?" All the eight instructors/educational managers got the answer right when using the bar chart, while two got it wrong when employing the pie chart. However, in both views, most respondents took longer than 15 seconds to identify the information and provide the answer.

2) Question 2: What is the evolution over the semester of the approved students compared to the total: The elaborated question was: "What semester had the highest percentage of approved students?". Where more respondents got it right when using the bar chart. However, quicker answers were obtained while looking at the line chart.

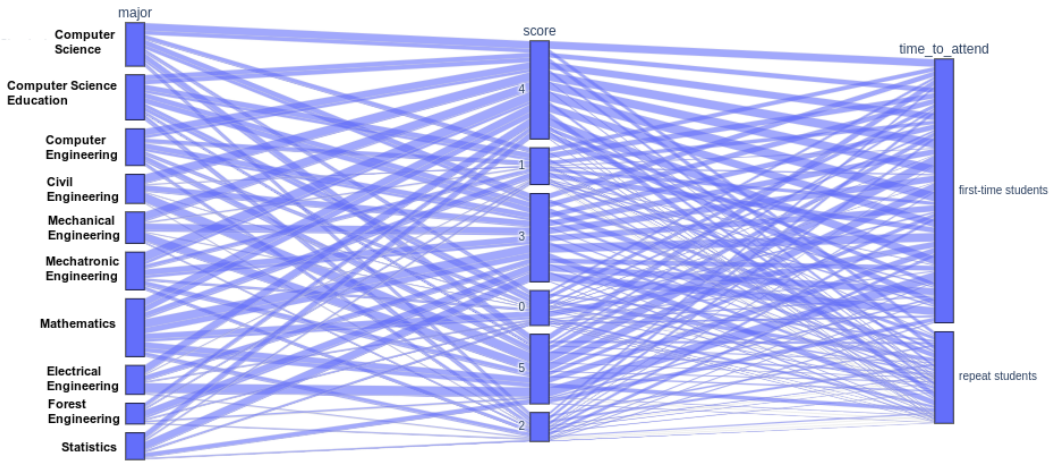
3) Question 3: What is the evolution of the students' grades: In order to evaluate the visualizations, the elaborated question related to the third question was: "Which semester has the highest average grade?". All respondents got the answer right when using the line chart and the bar chart, while the majority got it wrong when using the box plot.

4) Question 4: Which major did students most frequently repeat: The question designed for the last question was: "Students in which course have the lowest score average?". From eight people who answered the questionnaire, five could not answer this question by using the parallel coordinates and four by visualizing the violin plot. However, when looking at the line and bar charts most respondents got the question right. The violin plot was the one where users answered the question fastest, between 5 and 9 seconds.

V. Discussion

According to the study of visualization algorithms, it was observed that more than one algorithm can represent the same information. Although more complex algorithms present more information in a compact form, these are not as well known and are more difficult to interpret. However, from the results of the applied questionnaire, it was observed that the respondents felt more secure when using already known algorithms, such as pie charts and bar charts. Furthermore, the bar chart was the algorithm where the information was identified the most quickly and correctly.

The evaluation of the visualizations allowed to select the most suitable diagrams to present the data for each designed question in the study case. For the first question, the pie chart was selected due to more people answering the question correctly, even though the response time was the same



(d) Parallel coordinates.

Fig. 4: Generated visualizations to address Question 4.

TABLE III: Result for the elaborated questions.

		Response Time (seconds)				Correctness		
		1 to 4	5 to 9	10 to 15	>15	Correct answer	Wrong answer	"Don't know"
Question 1	Pie chart		3	1	4	6	2	
	Bar chart		3	1	4	8		
Question 2	Area chart		3	2	2	1	5	1
	Line chart	5	3			1	6	
	Bar chart	1	1	2	4	3	4	
Question 3	Line chart	4	3			7		
	Area chart	2	4			5	1	1
	Box plot		3	2	1	2	4	1
	Bar chart	4	3			7		
Question 4	Box plot		1	2	3	3	3	2
	Violin chart		2	1	1	3	1	4
	Bar chart with Line chart			2	5	4	3	1
	Parallel coordinates				3	3		5

compared to the bar chart. For the second question the bar chart was chosen, since it was the visualization with the most correct answers. Among the candidate algorithms for the third question, both line chart and bar chart assisted the users to answer the question quickly and correctly, in the same amount of time. The line chart associated with the bar chart was selected to represent the fourth question once it presents all the information related to the question, as well as being the graph in which more respondents got the answers to the questions right. This results are presented in Table IV.

TABLE IV: Selected algorithms.

Question	Data type	Selected chart
Question 1	categorical	pie chart
Question 2	chronological or categorical	bar chart
Question 3	chronological or categorical	line chart bar chart
Question 4	categorical	line chart

VI. Conclusion

This paper presented a study using visualization techniques applied to a case study of introductory programming courses at University of Brasília, defining and implementing a generic methodology that can be applied in other subjects. Four questions were designed in order to guide the study of the algorithms. For each question, the candidate visualizations and the attributes of the data were defined. In order to determine whether the selected algorithms communicate the information clearly and cohesively, a questionnaire was applied to 23 different instructors and education managers. This questionnaire consisted of questions that assessed the respondent's knowledge of the visualizations, as well as questions designed for each set of algorithms, in order to evaluate them.

The purpose of this paper was to provide a detailed study of several visualization techniques commonly found in the literature, in order to select the best algorithms to be applied to introductory programming courses data. In this way, instructors and educational managers can make decisions that support student learning in the courses.

To continue this work, it is proposed to analyze other charts for different questions than those elaborated in this paper, as well as to analyze other subjects. Finally, it is proposed the creation of a virtual module to be implemented in a virtual environment, with the visualizations selected from this work.

REFERENCES

[1] M. Khan and S. S. Khan, "Data and information visualization methods, and interactive mechanisms: A survey," *International Journal of Computer Applications*, vol. 34, no. 1, pp. 1–14, 2011.

[2] H. A. Do Nascimento and C. B. Ferreira, "Uma introdução à visualização de informações," *Visualidades*, vol. 9, no. 2, 2011.

[3] L. F. Estivalet, "O uso de ícones na visualização de informações," Ph.D. dissertation, Dissertação (Mestrado em Ciência da Computação)-Universidade Federal do Rio Grando do Sul, 2000.

[4] H. A. Do Nascimento and C. B. Ferreira, "Visualização de informações–uma abordagem prática," in *XXV Congresso da Sociedade Brasileira de Computação, XXIV JAI. UNISINOS, S. Leopoldo–RS*, 2005.

[5] T. Dwyer and P. Eades, "Visualising a fund manager flow graph with columns and worms," in *Proceedings Sixth International Conference on Information Visualisation*. IEEE, 2002, pp. 147–152.

[6] C. Ahlberg and B. Shneiderman, "Visual information seeking using the filmfinder," in *Conference companion on Human factors in computing systems*. ACM, 1994, pp. 433–434.

[7] M. Mcquaigue, D. Burlinson, K. Subramanian, E. Saule, and J. Payton, "Visualization, assessment and analytics in data structures learning modules," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018, pp. 864–869.

[8] T. Auvinen, L. Hakulinen, and L. Malmi, "Increasing students' awareness of their behavior in online learning environments with visualizations and achievement badges," *IEEE Transactions on Learning Technologies*, vol. 8, no. 3, pp. 261–273, 2015.

[9] W. L. Cambruzzi, S. J. Rigo, and J. L. Barbosa, "Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach." *J. Univers. Comput. Sci.*, vol. 21, no. 1, pp. 23–47, 2015.

[10] N. Culligan, K. Quille, and S. Bergin, "Veap: A visualisation engine and analyzer for press#," in *Proceedings of the 16th Koli Calling International Conference on Computing Education Research*, 2016, pp. 130–134.

[11] V. Hegde, "Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through r package," in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2016, pp. 1–6.

[12] T. Zhang and B. Yuan, "Visualizing mooc user behaviors: A case study on xuetangx," in *Springer International Conference on Intelligent Data Engineering and Automated Learning*, 2016, pp. 89–98.

[13] A. Essa and H. Ayad, "Improving student success using predictive models and data visualisations," *Research in Learning Technology*, vol. 20, 2012.

[14] L. Hansen, V. R. Borges, and M. Holanda, "A literature study of visual analysis in an educational context," in *2020 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2020, pp. 1–8.

[15] R. Lengler and M. J. Eppler, "Towards a periodic table of visualization methods for management," in *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA*, 2007.

[16] M. S. G. Canché, "Geographical network analysis and spatial econometrics as tools to enhance our understanding of student migration patterns and benefits in the us higher education network," *The Review of Higher Education*, vol. 41, no. 2, pp. 169–216, 2018.

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

[18] E. d. A. VALIATI, "Avaliação de usabilidade de técnicas de visualização de informações multidimensionais. 2008. 220f," Ph.D. dissertation, Tese (Doutorado em Ciência da Computação)-Universidade Federal do Rio grande ..., 2008.