Visualizing Regional Diffferences in Higher Education in Brazil: An Investigation of ENADE Micro-Data.

Felipe Nunes Walmsley*
Centro de Informática - UFPE

ABSTRACT

Over the years, several attempts have been made to assess the quality of higher education in Brazil. Currently, the ENADE exam is one of the most important tools in this assessment, and detailed data about exam results are made available to the public. We propose a visualization tool aimed at revealing regional differences in students performance, focusing on mean scores, and relative percentages of top and bottom scorers. Our results show clear regional differences, and suggest that factors within a state can heavily influence the performance of students, particularly when compared to neighboring states.

Index Terms: Human-centered computing—Visualization—Visualization techniques;

1 INTRODUCTION

In education, there is a constant concern with measuring the performance of students, and in particular there is a desire to use these measures to assess the quality of the education being provided to students. In Brazil, there are several initiatives designed to measure the performance of students at each level of education, such as the mandatory exams:

- The *Prova Brasil*, designed to measure the performance of primary school students;
- The National Secondary Education Exam (from the Portuguese *Exame Nacional do Ensino Médio*, ENEM), aimed at secondary students;
- The National Students Performance Exam (from the Portuguese *Exame Nacional de Desempenho dos Estudantes*, ENADE), which assesses tertiary education.

This last exam is the focus of this work. The (INEP), an institute subject to the Brazilian Ministry of Education publishes a data set containing the results of the exam, down to the level of individual students. This provides us with a unique opportunity to perform a number of analyses, which can help us paint a detailed picture of the current state of tertiary education in Brazil, and whether not the students are achieving the (formação) expected of them, and what factors are contributing to or detracting from their education.

While there are previous works dealing with ENADE, these consist mostly of research in the field of education. The works in [1] and [4] deal with ENADE from a historical and conceptual perspective. There are very few works that deal with the micro-data itself. In [6], the authors perform data mining to discover patterns on the 2014 exam data set, but the analysis is restricted to Computer Science programs in the state of Rio Grande do Sul. To the best of our knowledge, there is no work directly comparable to ours using ENADE data in the field of visualization.

*e-mail: fnw@cin.ufpe.br

In this work, we propose a visualization tool, which aims to help the user tease out interesting patterns in the distribution of students' grades around the country. The tool focuses on highlighting regional differences, allowing the user to observe how the students of each state perform, and also how unequal is the performance of students within each state.

2 BACKGROUND AND RELATED WORKS

2.1 The ENADE exam

In Brazil, the Ministry of Education requires all first and last year university students to sit in the ENADE. This exam aims to assess the knowledge of the students in both general topics and in their area of study, as a means to ascertain the quality of tertiary education in Brazil. The exam is applied in a rotating fashion to students from different areas of knowledge.

The individual (but anonymized) results for each student are made publicly available [2]. The data set contains detailed information about each student, such as their answers to individual items in the exam, their grades in the general and specific portions of the exams, and their overall grade. The 2016 ENADE micro-data data set consists of 216044 entries, each with 141 variables.

Furthermore, the students are required to answer a comprehensive questionnaire, which queries the students in relation to topics such as their perception of the exam itself, their opinion about the quality of their college and program, and also questions about their socioeconomic background. Below, some examples of the types of questions presented in the questionnaire, translated from the Portuguese original.

- *Perception of the exam:* How hard was this exam in the general knowledge portion?
- *Opinion about their college/program:* Are the labs, equipment and materials available to your program adequate to the needs of the program?
- *Socioeconomic background:* Was your admission to university due to affirmative action policies?

Given the plethora of data available, a multitude of analysis are possible. In fact, the data set is so large and complex that any single work would be incapable of performing a satisfactorily comprehensive analysis of the data. In light of this, we focus on evaluating the grades of the students in the exam, as a stepping stone for future works.

2.2 Related Works

Most of the works relating to the ENADE exam are developed within the field of educational research. In [1], the authors deal with the conception of the National System for the Evaluation of Higher Education (from the Portuguese, *Sistema Nacional de Avaliação do Ensino Superior*, SINAES). The ENADE is a tool within the SINAES, and the authors also deal with the conception and eventual implementation of ENADE. In [4], the authors also deal with the same topics, but also perform a comparison between ENADE and its predecessor exam, known as "Provão", and in particular the authors

evaluate how well ENADE managed to respond to the criticism levelled at the Provão exam.

In [6], the authors apply data mining techniques to the 2014 ENADE data, focusing on the Computer Science programs in the state of Rio Grande do Sul, and focus on finding groupings of the data at the institution level.

The closest work we could find to our own actually deals with ENEM data [5], which is also publicly available. This work also investigates regional differences in ENADE performance, but does so at the level of macro-regions (there are five officially recognized macro-regions in Brazil, North, Northeast, Southeast, South and Center-West), instead of looking at the level of individual states.

3 Proposed Method

The system proposed in this work is a visualization tool aimed at helping the user visualize regional differences in ENADE results. In this work, we use the results for the 2016 edition of the exam, the latest results available as of July 2018. As previously explained, the ENADE micro-data data set consists of both a large number of entries and variables. Therefore, in this work, we focus on a small number of variables and groupings. In particular, we focus on the overall grades of students, the most basic proxy for the quality of education in a particular region. We do not perform any analysis pertaining to the student questionnaire, due to the complexity of the data, and the necessity of specific methods to deal with self-reported data and socioeconomic variables.

We proposed two different visualizations, both presented as choropleths. The first is focused on grades at the level of the states of the federation. We provide the user with the following three options in this visualization.

1. **Average Grade:** The average grade of the students in the state;
2. **Fraction of the best students:** The fraction F_{top} of the students from the state in the top 1% of grades;
3. **Fraction of the worst students:** The fraction F_{bot} of the students from the state in the bottom 1% of non-zero grades.

Where F_{top} is given by Equation 1:

$$F_{top} = \frac{\frac{N_s^{top}}{N_{top}}}{\frac{N_s}{N}} \quad (1)$$

Where N_s^{top} is the number of students from the state in the top 1% of grades, N_{top} is the number of students in the top 1%, N_s is the number of students from the state, and N is the total number of students in the data set. Note that N_{top} is simply equal to $0.01N$. The same normalization is performed for F_{bot}, but replacing N_s^{top} with N_s^{bot} and N_{top} with N_{bot}.

The second visualization offered to user is concerned with the difference between individual higher education institutions. However, in this case, since there is a large number of such institutions, we avoided presenting the exact same options as we did above, and instead focused on showing the disparities between institutions within a state. In order to do so, we utilized the Gini coefficient, which measures the inequality in a distribution.

Once again, we offer the user the same three options as above. In this case, however, we have the value of each of the above variables for each institution in a state. For example, let A_i be the average grade for an institution i in a state. In the average grade visualization, we present the user with the Gini coefficient for A_i.

Let F_{top}^i be the fraction of students from an institution in the top 1% (similar to F_{top}). It is defined by Equation 2:

$$F_{top}^i = \frac{\frac{N_i^{top}}{N_{top}}}{\frac{N_i}{N}} \quad (2)$$

Where N_i^{top} is the number of students from the institution in the top 1% of grades, N_{top} is the number of students in the top 1%, N_i is the number of students from the institution, and N is the total number of students in the data set. A similar normalization is performed for the fraction of students in the bottom 1%, F_{bot}^i.

Again, instead of presenting the individual values for F_{top}^i and F_{bot}^i, we present the Gini coefficient for these variables.

We use HTML and JavaScript to produce the web page containing the visualization. The map displayed is created using the Leaflet library. In order to read the data and perform reductions and sub-setting, we employ the D3.js and Crossfilter libraries. The visualization tool can be accessed at the address `https://fnw.github.io/dados-enade-visualizacao-2018-1/`.

4 Results

4.1 Our Hypotheses

Coming into this investigation, we expected to see large regional differences, as Brazil has a high degree of regional inequality in distribution of income, which we would expect to greatly influence the performance of the students. Furthermore, we expected a few institutions in each state (and even in the country itself) to concentrate the majority of high-performing students.

Nevertheless, we aimed to use our tool to uncover unexpected results, which go against common sense. We posit that these results may lead to the uncovering of outliers in the Brazilian tertiary education system, and that future studies of these outliers may result in the creation of policies that will benefit the system as a whole.

4.2 Discussion

Our tool allowed us to identify some facts of interest about the performance of states in the ENADE exam.

The first interesting case is that of Espírito Santo, which seems to be one of the better performers, in spite of being poorer than other states in the federation (in terms of GDP per capita) [3] and having a relatively small population. The state has both the highest mean score and highest relative percentage of top performers, F_{top}, as shown by Figures 1 and 2

Meanwhile, on the northeast macro-region of Brasil, which is comprised of Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe and Bahia, only two states have a value of F_{top} greater than one. These states are Ceará and Rio Grande do Norte, with values of 1.52 and 1.22, respectively. Figure 3 shows these two states in the context of the macro-region.

Another interesting case is that of Paraná, which performs unexpectedly poorly when compared to the other two states of the south macro-region of Brazil. While the three states are all in the top 6 of richest states of the federation (again, in terms of GDP per capita), only Paraná has a relative percentage of bottom scores greater than one, as shown in Figure 4.

5 Conclusions

In this work, we proposed a visualization tool for the ENADE micro-data data set, with the goal of providing users with the tools to investigate regional differences in student peformance. More specifically, we proposed two choropleth based visualizations that allowed the users two observe the average scores and percentage of top and bottom students for each state, and also visualize how unequal was the performance of institutions within a state.

Our results show that particular factors, which we were not able to tease out with our analysis, can heavily influence the performance of students in a state, particularly when compared to similarly

Figure 1: Mean scores for the states, with Espírito Santo highlighted.

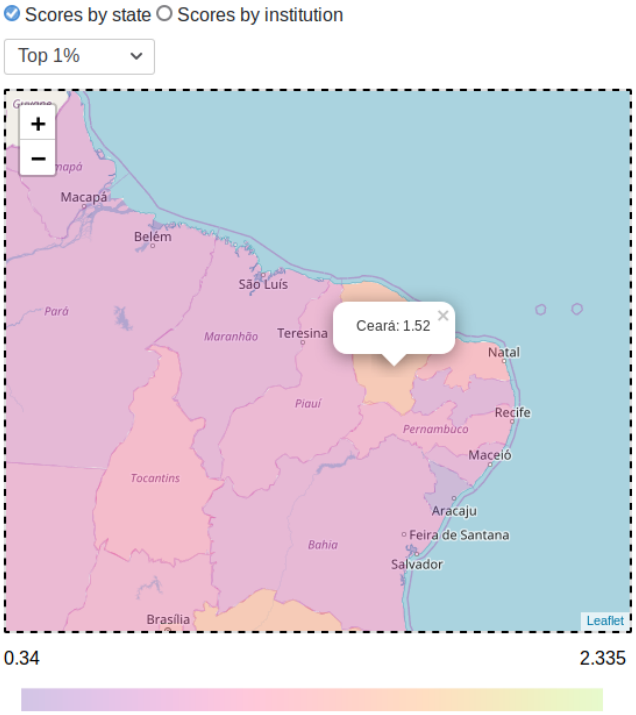

Figure 3: Percentages of top scores for the states, with Ceará highlighted.

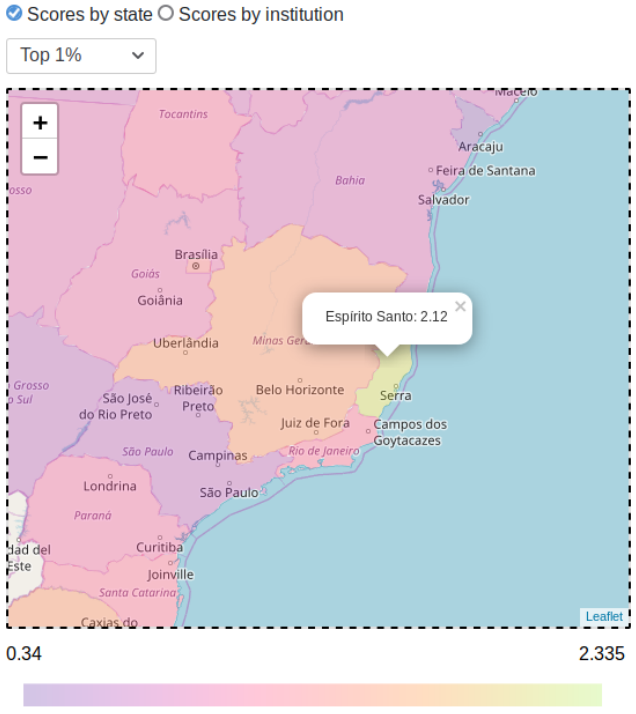

Figure 2: Percentages of top scores for the states, with Espírito Santo highlighted.

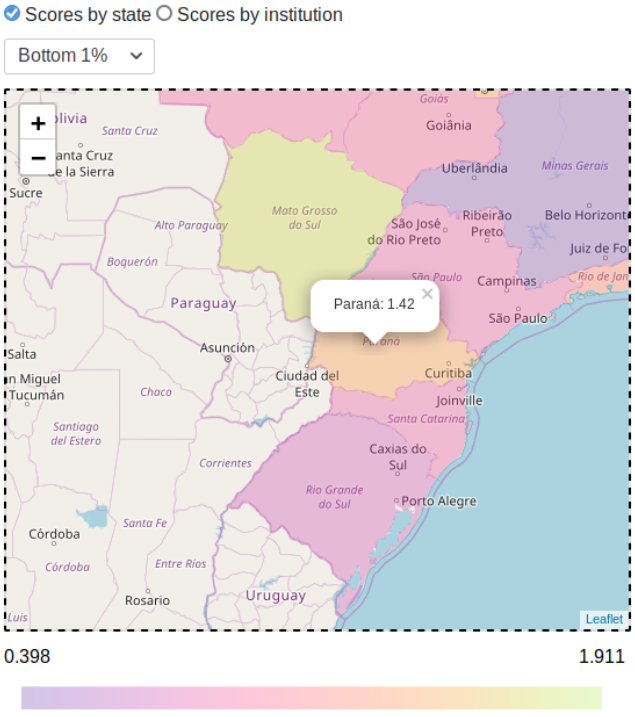

Figure 4: Percentages of bottom scores for the states, with Paraná highlighted.

wealthy neighboring states. Further analysis into the cases of states like Espírito Santo, Ceará and Paraná may reveal what factors can positively or negatively influence the higher education system in a state.

5.1 Future Work

Given the nature of the data set investigated, there are a multitude of possible future avenues of investigation. These may include simply teasing out other interesting patterns, involving the grades, the socioeconomic factors or a combination of both. Another possibility is the development of more sophisticated visualizations, which allow for the discerning of finer details, such as details at the institution level, or visualizations which attempt to tease out patterns and groupings at the individual student level. Finally, the data also allows for the application of statistical and machine learning methods, in order to investigate correlations and causality relations between the different variables, in particular between socioeconomic variables and the performance of students.

REFERENCES

[1] M. R. F. d. Brito et al. O sinaes e o enade: da concepção à implantação. *Avaliação: Revista da Avaliação da Educação Superior*, 13(3):841–850, November 2008.

[2] I. N. de Estudos e Pesquisas Educacionais Anísio Teixeira. Microdados - inep, 2018. [Online; accessed 3-July-2018].

[3] I. B. de Geografia e Estatística. Ibge — agência de notìcias — contas regionais 2015: queda no pib atinge todas as unidades da federação pela primeira vez na série, 2017. [Online; accessed 3-July-2018].

[4] R. E. Verhine, L. M. V. Dantas, and J. F. Soares. Do provão ao enade: uma análise comparativa dos exames nacionais utilizados no ensino superior brasileiro. *Ensaio: Avaliação e Polìticas Pùblicas na Educação*, 14(52):291–310, 2006.

[5] E. Viggiano and C. Mattos. O desempenho de estudantes no enem 2010 em diferentes regiões brasileiras. *Revista Brasileira de Estudos Pedagógicos*, 94(237), 2013.

[6] N. P. B. Vista, M. F. Figueiró, and P. M. M. Chicon. Técnicas de mineração de dados aplicadas aos microdados do enade para avaliar o desempenho dos acadêmicos do curso de ciencia da computação no rio grande do sul utilizando o software r. In *Anais do I Seminário de Pesquisa Científica e Tecnológica*, 2017.