Anomalies in Students Enrollment Using Visualization

Nishith M Thakkar, Lenin Mookiah, Douglas A. Talbert, and **William Eberle**
{nmthakkar42,lmookiah42}@students.tntech.edu and {dtalbert,weberle}@tntech.edu
Tennessee Technological University
Cookeville, TN 38505

Abstract

In recent years, data visualization has gained popularity as a method for discovering anomalies. In this paper, we study the application of visualization to anomaly detection in student enrollments by race, particularly in elementary and secondary education. We use data from the *U.S. Department of Education* that measures each state's progress towards implementation of the No Child Left Behind Act (NCLB) and also contains details for state level grant programs authorized by the Elementary and Secondary Education Act (ESEA) as amended by the NCLB. We demonstrate that a data visualization approach is able to effectively discover interesting anomalies in student enrollment among different states in the United States. We also use visualization to study anomalous patterns in various federal funding initiatives at the state level. In summary, we show that visualizations aide in the discovery of interesting patterns and anomalies.

Introduction

The Department of Education's elementary and secondary programs annually serve approximately 50 million students from 14,000 school districts, including around 32,000 private schools and 98,000 public schools. The programs from the department provide loan, grant, and work-study assistance to more than 12 million post-secondary students [1]. The data regarding academic standing, racial buildup and other related metrics are provided for usage in the public domain.

Because of the large volume of data, it is possible that a *visualization* might help education specialists discover potentially interesting patterns not readily available from statistical reports. First, interesting patterns could help the department in effectively planning their objectives and vision. For example, where is most funding allocated? Where can we better leverage funding? What areas have shortcomings or limitations?, etc. (Cohn 1987). Second, visualizations could aide the department in the discovery of trends and patterns in dropouts across demographics, i.e., race, gender, age, etc. (Rumberger 1983). Third, the department may be aided by knowing how one state performs in comparison to another with respect to, for example, race (Hemphill and Vanneman 2011). While traditional statistical reports exist, the results can be daunting to read and comprehend. The hypothesis of this paper is that visualization is a better approach, as it more intuitively summarizes all of the data, enabling a relevant decision maker to discover anomalies and other interesting patterns.

In this work, we present an approach for *visualizing school performance data* along with results for two specific problems: discovering anomalies in student enrollment by race, and how federal funding is applied to schools at a state level in the U.S.

[1] http://www2.ed.gov/about/overview/fed/role.html

Related Work

Cohn analyzed the effects of financial factors on academic progress. He presents the effects of the investment by federal and state grants on student success and its effect on the academic performances of the states (Cohn 1987). Rumberger examines the extent of the high school dropout problem in 1979 and investigates both the stated reasons students leave school and some of the underlying factors influencing their decision (Rumberger 1983). Particular attention is focused on differences by sex, race, and family background. McGuinn reported about the Obama Administration's program of *Race to the Top* and how it influences state's decisions over education policies and the evolution of the No Child Left Behind Act (McGuinn 2011). Additionally, they explore the role of the federal government in education. The different behaviors which are exhibited by high school students in the decision making process for their career are studied and analyzed in the work by M. Tang et al. (Tang, Pan, and Newmeyer 2008). They examine a collection of patterns observed by the students and what influences the final decisions made by the students.

In short, much of the research are case studies that examine the impact of government grants for innovation/teaching-evaluation methods, factors affecting high school students' career, and the impact of technology on teaching. What is presented in this paper takes a different look at the data through *visualization* techniques to *discover anomalies* in student enrollment by race as well as federal funding.

Data Source and Data Preparation

For the visualization, yearly data of primary and secondary schools for each state is collected from the Education Department's website [2]. In addition, NCLB reports provided by the states to the Department of Education for each academic year are available on the website. The data is broken down into sections and aggregated. Specifically, the aggregated data is partitioned into the following sections: State Facts and Figures, Achievement Data, Accountability Data, Teachers, Options for Parents, and Annual Measurable Objectives. Each of these aggregated data can be easily exported into a Microsoft Excel spreadsheet. In this paper, we are particularly interested in the data from the *State Facts and Figures* report for the latest available year, 2014.

The Excel data is then converted into a comma-separated format (CSV) for input into our data visualization tool. We also use population distribution information of each race/ethnicity as reported from the U.S. Census Bureau's 2015 Current Population Survey[3]. The race-wise population by state is only available since the year 2015, and thus we are limited. However, we conjecture that a comparison of the census population from 2015, with student enrollments from the academic year 2013-2014, should not affect our analysis and conclusions, as demographics should be similar.

Experimental Setup and Visualization

First, we discuss our definitions of *growing populations* and *anomaly*. Second, we present our experimental setup.

Definition of growing/receding population. The particular racial population that has the highest increase in the percentage of population vs. the enrollment figures compared to the other ethnic groups in the same state is defined as the growing population race. Similarly, the racial population with the largest decrease is defined as the receding population race.

Definition of anomaly. We define an anomaly as the largest differences in the percentage of a race's student enrollment with respect to its population.

We will use the D3 JavaScript library (D3.js[4]) for the data visualizations presented in this paper. The D3.js library is designed to visualize large datasets using standard web technologies, which are HTML, Scalable Vector Graphics (SVG) and CSS. SVG, which is an XML-based vector image format, is adept at two-dimensional graphics with support for interactivity and animation (Lane 2007).

In this work, we use two types of visualizations. First, we want to be able to study the relationship between race and population rate, as well as between race and student enrollment rate. We call this a *bipartite visualization*, as shown in the three examples represented in Figure 1, Figure 2, and Figure 3, corresponding to the three states Georgia, Hawaii, and Alaska, respectively. Specifically, the visualization is useful for discovering anomalies and interesting patterns, and the associated tools allow for further drilling down into the data, as will be further discussed in the next section.

[2]http://eddataexpress.ed.gov/state-tables-main.cfm

[3]http://kff.org/other/state-indicator/distribution-by-raceethnicity/

[4]https://d3js.org/d3.v3.min.js

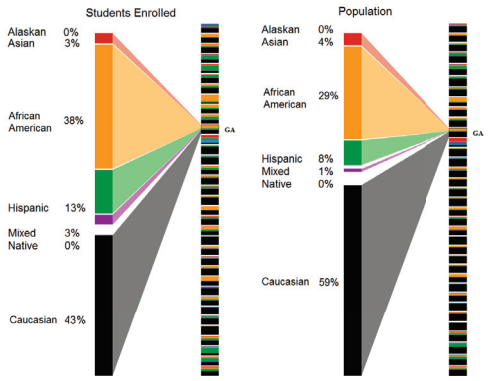

Figure 1: Bipartite visualization corresponding to Georgia (GA) in Table 1.

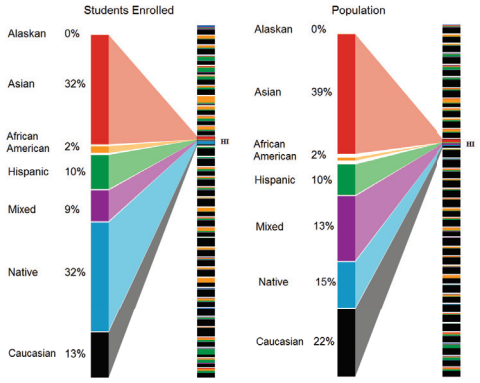

Figure 2: Bipartite visualization corresponding to Hawaii (HI) in Table 1.

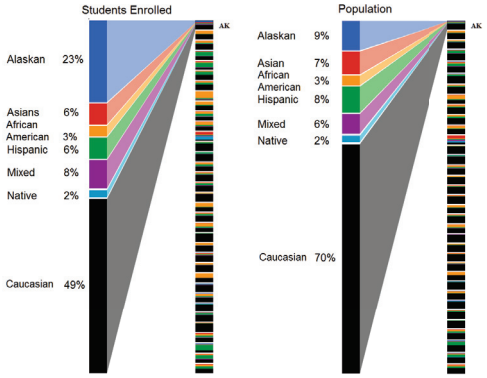

Figure 3: Bipartite visualization corresponding to Alaska (AK) in Table 1.

We study these visualizations of each state in order to discover anomalies related to the rise or fall of a population rate, and/or student enrollment rate, among different races such as Hispanics, African-Americans, Alaskans, Asians, Native Americans, Mixed, etc., where "mixed" refers to in-

State	Growing Population Race (%)	Growing Population (%)	Growing Enrollment (%)	Difference (%)	Receding Population Race (%)	Receding Population (%)	Receding Enrollment (%)	Difference (%)
Alaska (AK)	Alaskan	9%	24%	15%	Caucasion	67%	51%	-16%
California (CA)	Hispanic	28%	54%	26%	Caucasion	55%	25%	-30%
Georgia (GA)	Black	29%	38%	9%	Caucasion	59%	43%	-16%
Mississippi (MS)	Black	37%	50%	13%	Caucasion	60%	46%	-14%
New Mexico (NM)	Hispanic	31%	62%	31%	Caucasion	58%	25%	-33%
Illinois (IL)	Hispanic	14%	25%	11%	Caucasion	68%	51%	-17%
Nevada (NV)	Hispanic	22%	41%	19 %	Caucasion	60%	37%	-23%
New Jersey (NJ)	Hispanic	17%	24%	7 %	Caucasion	63%	49%	-14%
Hawaii (HI)	Native	14%	33%	19%	Mixed	19%	9%	-10%

Table 1: Anomalies found using *bipartite visualization* from each state as examples shown in Figure 1, Figure 2, and Figure 3 for the states of Georgia (GA), Hawaii (HI), and Alaska (AK), respectively. It shows a state-wise difference for found anomalies between percentage of population growth and enrollment growth among different races. Also, it shows state-wise differences between percentage of receding population and receding enrollment among different races.

dividuals reported as belonging to two or more races. A total of 9 anomalies (i.e., unexpected deviations in population rates) are shown in Table 1. It should be noted that we chose the states of Georgia, Hawaii, and Alaska because their populations have a varied population in terms of race (Hispanic, African-American, Native-American, and Alaskan), as shown in Table 1. One will notice that Hispanic is the most dominate race in the table. Therefore, since we are trying to discover anomalies (i.e., something that is rare) in the data, we will exclude showing any visualization examples for Hispanics and show examples for only the following races: African-American (Georgia), Native (Hawaii), and Alaskan (Alaska).

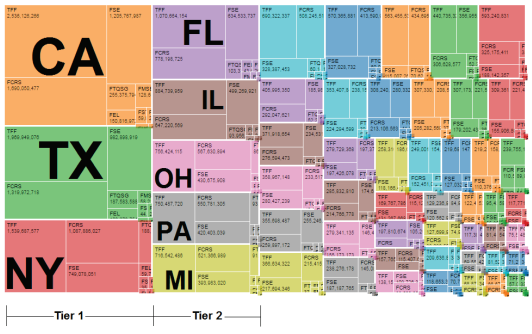

Figure 4: Funding Visualization using TreeMap

Second, we want to visualize anomalous funding at the state level. We use a treemap visualization as shown in Figure 4. We call this visualization *funding visualization using treemap*. For each state, there are 10 different types of funding - Total Federal Funding for Elementary and Secondary Education Programs, Funding for Assessing Achievement (State Assessments), Funding for College-and Career-Ready Students (Title I, Grants to LEAs), Funding for English Learner Education (English Language Acquisition), Funding for Homeless Children and Youth Education, Funding for Improving Teacher Quality State Grants, Funding for Migrant Student Education (State Agency Program), Funding for Neglected and Delinquent Children and Youth Education (State Agency Program), Funding for School Turnaround Grants (School Improvement State Grants), and Funding for Special Education–Grants to States. Treemap helps find the states that get significant funding and also identify prominent types of funding states get from the federal government. Demos of the visualizations are available for the *bipartite visualization* [5] and for the *funding visualization using treemap* [6].

Results and Discussion

In this section, we discuss the results of our visualizations of the data. We first discuss results from the *bipartite visualization*, which is created for the number of children enrolled in schools per state based on race, versus the total population in that particular state with data from *State Facts and Figures* from the NCLB report database. The anomalies found using the visualization are shown in Table 1. We discuss the results of the data visualizations shown in the Figure 1, Figure 2, and Figure 3. First, as shown for Georgia (GA) in Figure 1, the African-American population has a growing

[5] http://run.plnkr.co/plunks/HtqtWV33u3Mq7A0NIjEU/
[6] http://run.plnkr.co/plunks/F6lyESVhG6X4pyMU0pOS/

enrollment of 9%, where as the Caucasian population has a dropping enrollment of 16%. Second, looking at Figure 3 that represents Alaska (AK), we see a large difference in the number of Alaskan students (24%) versus the total Alaskan population (9%). Also, the total percentage of the Caucasian population is (67%) versus just Caucasian students (51%). This shows there are significantly more Alaskan children as compared to the general Alaskan population.

There is one common trend that can be seen among some of the states in Table 1. There is a considerable gap in the enrollment of Caucasian students relative to the overall Caucasian population. This might be because either there is a relatively decreasing population of Caucasian children with respect to other races or there is an increasing population of the other races. The only state where there is a different trend is Hawaii, as shown in Figure 2, where a difference of 10% can be seen in the mixed race population versus student enrollments of a mixed race. Whereas, the Native or Other Pacific Islander population/enrollment increased from 14% to 33%.

There are two key advantages for using such visualizations. First, there is a comprehensive comparison of all states in single view. Second, the visualizations provide two views: (1) a race-wise distribution of student enrollments/population over each state, and (2) how populations of a specific race are distributed across different states - and potentially other questions could be answered. For example, notice that 1% of the Alaskan population is distributed over different states, and by hovering over Älaskan, one can see that the Alaskan population distribution is primarily over the states New Mexico (22%), North Dakota (18%), Arizona (8%), Michigan (10%), Northern Carolina (6%), Nevada (6%), and Nebraska (4%). Use of this information by an entity like the Department of Education, could enable them to a better understanding of the movement of races throughout the U.S. There are potentially other such interesting information that could be inferred from the visualization.

Funding visualization using TreeMap as shown in Figure 4 is used to visualize various funding received for the states partitioned by the type of funding it was getting. As can be seen in the visualization shown in Figure 4, there are three states that receive the most funding, California (CA), Texas (TX), and New York (NY), and five other states that receive considerable funding, Florida (FL), Illinois (IL), Ohio (OH), Pennsylvania (PA), and Michigan (MI). The key to representing information this way is that someone who might be interested in understanding the data can get a *quick* answer to their question (i.e., "Which states receive the most federal education funding?"). While this information can be found in a table representation of the data, when you are dealing with lots of data (i.e., big data), answers like this may not be easy to derive quickly.

Additionally, a few other trends and/or anomalies can be seen using this type of visualization. In general, the type of funding named *Total Federal Funding for Elementary and Secondary Education Programs* is the highest source of funding for all the states. The second most valuable grants received by the states were for *College/Career ready students*. Also, we discover there has been no funding for migrant students in three states: Rhode Island, Washington D.C. and Connecticut. Again, using a *visualization* can have an advantage over viewing data table (like a Microsoft Excel spreadsheet) for the discovery of trends and anomalies because: (1) visualizations provide an overall comprehensive bird's eye view of all states in a single view, and (2) fine-grained details of how funds are distributed at different layers can be captured by *drilling-down*. The figures in the paper are a set perspective of a dynamic query-able visualization that we have focused upon. The policy makers can query the visualizations as per their requirements for the needed perspectives.

Conclusion

In this work, we presented a data visualization approach that allows one to more easily discover trends and anomalies in educational demographics and federal funding at the state level. We collected data from multiple data sources and implemented two different visualizations, namely *Bipartite Visualization* and *Funding Visualization using TreeMap*. Using *Bipartite Visualization*, we show that our visualization approach can effectively find anomalies in student enrollments. Using *Funding Visualization using TreeMap*, we show that funding-related anomalies can be found.

There are a few directions for future work. First, we could visualize using temporal data of student enrollments. Such a visualization would help users to intuitively see the change in enrollment. Second, a more sophisticated visualization could be studied to let users initially choose to cluster a group of states into geographical groups such as south-eastern states, north-western states, etc., and then subsequently look for patterns in the visualizations. Finally, more visualization techniques like Heat Maps and Star Charts should also be studied.

References

Cohn, E. 1987. Federal and state grants to education: Are they stimulative or substitutive? *Economics of Education Review* 6(4):339–344.

Hemphill, F. C., and Vanneman, A. 2011. Achievement gaps: How hispanic and white students in public schools perform in mathematics and reading on the national assessment of educational progress. statistical analysis report. nces 2011-459. *National Center for Education Statistics*.

Lane, D. 2007. Scalable vector graphics. *AMC* 10:12.

McGuinn, P. 2011. Stimulating reform: Race to the top, competitive grants and the obama education agenda. *Educational Policy* 0895904811425911.

Rumberger, R. W. 1983. Dropping out of high school: The influence of race, sex, and family background. *American Educational Research Journal* 20(2):199–220.

Tang, M.; Pan, W.; and Newmeyer, M. 2008. Factors influencing high school students' career aspirations. *Professional School Counseling* 11(5):285–295.