RIO PREFEITURA

COLUMBIA GLOBAL CENTERS | LATIN AMERICA
RIO DE JANEIRO

COLUMBIA UNIVERSITY

BIG DATA & EDUCATION

An Exploratory Study of Rio de Janeiro City Public Schools System

Alberto Zeraik, Bruno Bondarovsky, Eduardo Padua,
Fernando Ivo Cavalcante, Luiz Eduardo Ricon,
Victor Zajdhaft

RIO PREFEITURA

COLUMBIA GLOBAL CENTERS | LATIN AMERICA
RIO DE JANEIRO

COLUMBIA UNIVERSITY

January, 2015

1. Introduction

Rio de Janeiro municipal public school network is one of the largest in the world, comprising over 1,500 schools (from Pre-K to Elementary), 42,000 teachers and 650,000 students. In recent years, the city's public educational system has been engaged in a comprehensive agenda of reforms, and although schools and students performances have been continually improving over the years, they are still far from ideal.

All over the world, from the local, community level up to the national and international stages, the challenges in Education are becoming increasingly critical for economic and social development, as students are continually demanded to develop new skills and competencies more fit to the 21st century globalized world economy. That puts crescent pressure over teachers, principals and educational managers, especially in the public sector. And as governments raise their investments in Education[1], it is becoming proportionally hard to identify the most effective strategies, those that could deliver the greatest results in the shorter term. In many ways, it's not about having the money, but how to spend it.

There are many possible approaches to those issues, but the increasing social-economic relevance of the Internet and the plunging costs in data collection, storage and processing are leading to the generation and use of huge volumes of data, putting the use of Big Data Solutions among the most promising ones for the Education sector.

1.1 Big Data & Education

The term Big Data is commonly used to describe a wide range of new techniques and technologies for processing and analyzing large volumes of data. As other forms of Data Driven Innovation (DDI), Big Data Analytics (BDA) is becoming a key trend in 21st

[1] *Recently, the Brazilian Federal Government has approved the destination of 10% of the country's GDP towards Education and 75% of the federal revenues obtained from the "Pre-sal" deep sea oil exploration towards Public Health and Public Education investments.*

century efforts towards value creation, fostering new processes, products and creating significant competitive advantages.

In the Public sector, the use of DDI and BDA is an opportunity “to do more with less”, optimizing processes and public investments, applying new technologies to cross and correlate a wide range of available data from several different sources, be them structured or not, intersecting unrelated agencies in order to address key issues, and also tangent through secondary subjects that can affect primary targets.

Up to this day, most efforts involving Big Data in Education focused on individual student performance, applying IT tools and Big Data solutions to individualize and personalize teaching strategies in order to maximize the student’s outcomes. There is a clear trend in Education for applying technology to custom-tailor tests, assessments, evaluations, exercises and other school procedures, in order to provide teachers with tools to treat each student as a unique individual. In this study, we mainly focused on external issues of educational environment that may influence students performance.

Gathering data from several different sources, including social-economic statistics and a wide array of data from several agencies in the city’s administration, the goal of this study was to confront these data with the main educational performance indicators, trying to identify interesting or hidden correlations between them, looking not only at the general trends, but also (and maybe more importantly!) observing the so-called “points outside the curve”, or in other words: treating outliers not as discardable or faulty data, but as elements that could provide public managers with unusual insights that could lead to innovative solutions, in order to improve students and school performance, optimize investments and, at the same time, highlight best practices that could be analyzed and replicated to the whole system.

2. Methodology

It was our initial assumption that the whole environment surrounding the students might affect their ability to perform. Traditional methodologies have already tracked a wide variety of Key Performance Indicators (KPI’s) in order to measure performance on Education. This proposal does not question these KPIs, but rather understands that now there is a whole set of new perspectives that can be side-measured and that could not be measured before, due to technological limitations imposed by the tools available in the past.

Therefore, this exploratory study addressed the status of Rio de Janeiro's Public Schools, trying to identify key factors that could explain the gap between schools with satisfactory results and those that, although sharing the same level of public investment, still have not been able to perform. Our proposal was to gather and compare data from several different sources and correlate them to schools performance indicators from local and national standardized tests.

To do so, this study applied Big Data to monitor school performance for significant deviations from its historical profile, or from a profile of similar peers. Our assumption was that we might have so much data already collected about the schools without knowing exactly what is going on. If a particular school score deviates significantly from the general trend, then we considered it as an outlier. With this method, we expected to reveal best practices or hints to search locally for what might be interfering with the school's performance.

After looking at available information about similar projects using Big Data in Public Education, this study tried to go beyond the obvious, proposing a methodology based on the use of Big Data technology to identify behaviors or conditions that might impact on school and student's performance. Our goal was to offer policy makers and public education managers a new tool to support their decision making process in order to achieve higher educational outcomes.

Several studies try to identify the determinants of the schools adequate or poor performances in standardized tests. However, with the technological evolution it has become possible to analyze an infinite number of variables in order to identify in complete and specific way each factor causing variations in student's performance in a given school. This is only possible through the increasing processing capacities and the use of Big Data solutions.

In this study, we tried to confront schools performance indicators with several indicators from other sources, related to external elements, trying to assess how these external factors might affect the school performance.

Due to the limits and constraints to this study, be it time, budget or institutional limitations, we decided to use only widely available data, selecting indicators that we could combine and compare for the same array of public schools. These indicators were grouped into three different categories: Students, School and Environment. We decided not to use data about the families because our focus was in the environment, or in better terms, in the external indicators, to evaluate if they could impact significantly on schools performance or not.

For each school, only the following[2] indicators were analyzed:

- Cost - Average annual cost per student
- ideb_ai - The Basic Education Development Index (IDEB) is an indicator created by the federal government to measure the quality of education in public schools . The last IDEB, held in 2013, states the medium note of Brazil and 5.2 in the early years , 4.2 and 3.7 in the final years. IDEB is calculated based on student learning in mathematics and Portuguese (Brazil Exam) and school flow (pass rate). In this case, “ai” refers to the early years of elementary school (1st to 5th grades).
- ideb_af – the same for the final years of elementary school (6th to 9th grades).
- iderio_ai - As the IDEB is held every two years (odd years) by the federal government through the “Brazil Exam”, the city created the IDE Rio for the pair years, with the application of “Rio Exam”.
- iderio_af - – the same for the final years of elementary school (6th to 9th grades).
- NSE - The Socioeconomic status (SES) summarizes the characteristics of individuals in relation to their income, occupation and education, allowing classes to analysis of similar individuals in relation to these characteristics.
- Isp - is an indicator of police reports which aggregates records of theft, robbery, death threats and more serious cases such as murders
- Ids - SUSTAINABLE DEVELOPMENT INDICATORS - The construction of sustainable development indicators in Brazil is part of the international

[2] *These indicators were chosen in part because of their potential relevance to the objectives of this study and also due to the limitations and difficulties we faced to obtain, treat and validate the data. If more time, budget or institutional leverage and support were available, the list would probably be different. But these limitations are within the scope of an exploratory study such as this one.*

set of efforts for implementation of the formulated ideas and principles of the United Nations Conference on Environment and Development in Rio de Janeiro in 1992, with regard to the relationship between the environment, society, development and information for taking decisions.

- Hdi - The Human Development Index (HDI) is a summary measure of long-term progress in three basic dimensions of human development: income, education and health (longevity).
- Hdi-l – Longevity. The item longevity is evaluated according to the life expectancy at birth. This indicator shows how many years a person, born in a specific year and specific locality is supposed to live in average. It reflects health and sanitation conditions of the locality, since life expectancy is highly influenced by early childhood deaths.
- Hdi-e – Education. To assess the extent of education the HDI calculation considers two indicators. The first, weighing twice, is the literacy rate of people with fifteen or more years old. The second indicator is the schooling rate: the sum of people, regardless of age, enrolled in any course, whether primary, secondary or higher, divided by the total number of people between 7 and 22 years of the locality.
- Hdi_r - income: Income is calculated based on the GDP per capita (per person) in the country. As there are differences between the cost of living from one country to another, the income measured by the HDI is in PPP dollars (Purchasing Power Parity), which eliminates these differences.
- casos_dengue - This indicator includes the cases recognized and registered as cases of dengue fever in the city.
- 1746_auxilio_interno - The 1746 is the direct relationship channel between the citizen and the City to request services. This indicator includes calls to action made to the Municipal Guard within public schools.
- 1746_dengue - This indicator refers to calls related to complaints of transmitters of dengue mosquito outbreaks.
- 1746_estacionamento_irregular - This indicator refers to calls related to irregular parking within the city.
- 1746_iluminação_publica - This indicator refers to calls related to problems relating to public lighting, whose responsibility is the RIOLUZ

- 1746_pavimentacao - This indicator refers to calls related to potholes in the streets and sidewalks of the city.
- Dis - age-grade gap

3. Operationalizing the data

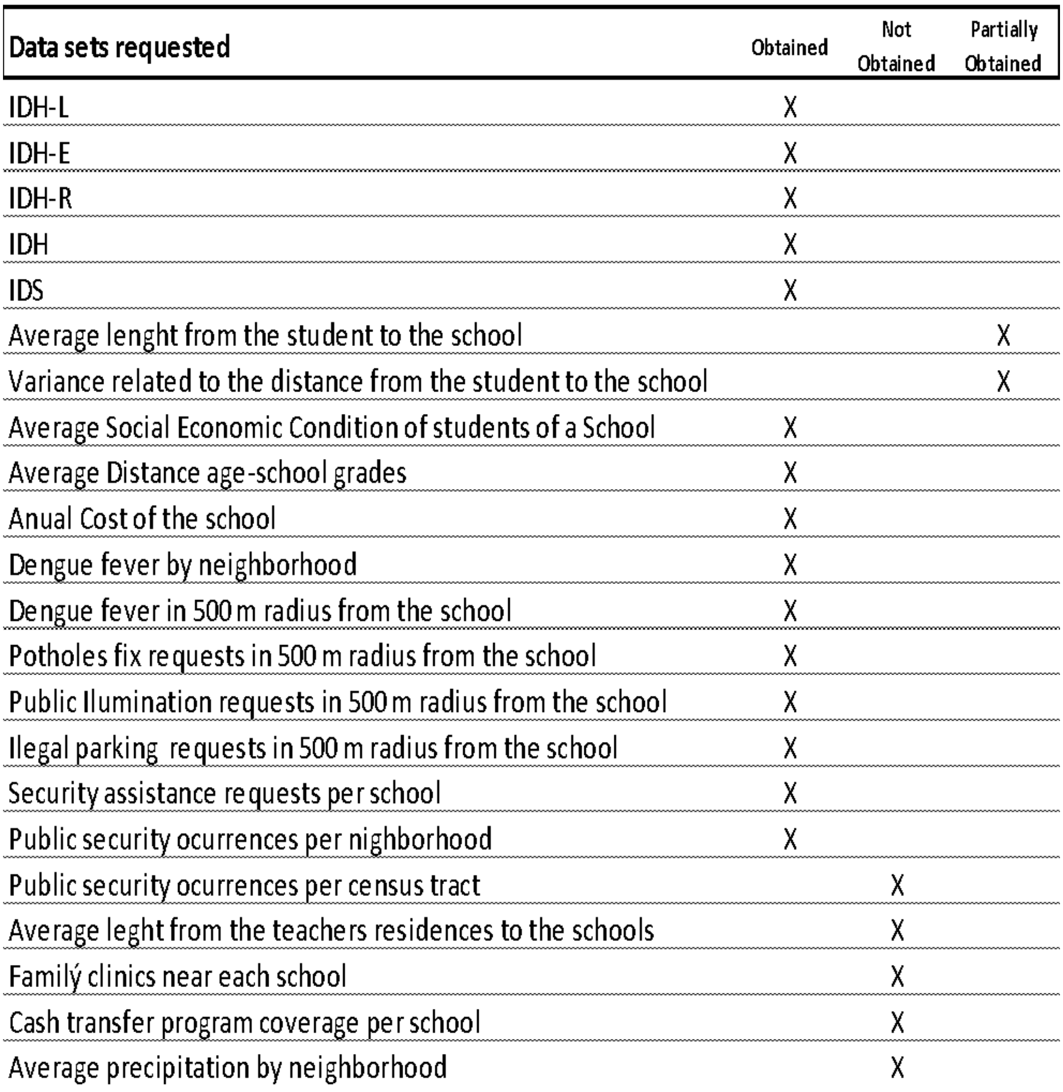
It is known that a Big Data project is a complex initiative, usually expensive, involving very large volumes of data, both from structured and unstructured sources, but it still involves a heavy and long process of transforming data so that they can be correlated.

Thus, it is not feasible to execute an effective Big Data project within the scope defined for this study, which led us to propose a model that could meet this project purpose, test some intersections with a suitable set of indicators and then suggest a more thorough initiative based upon the analyzed model and its expansion.

The first step in this context was to understand how school performance may be affected by the components that make up the system, in order to pursue actions that can promote gains in performance.

We selected sets of indicators which could contribute to the proposed analysis. The table below shows the indicators that were required, those that were achieved, the ones that were partially achieved and the ones that were not achieved.

The techniques used to assess the weight of the components in school performance were Multiple Linear Regressions (MLR) and several intersections.



Data sets requested	Obtained	Not Obtained	Partially Obtained
IDH-L	X		
IDH-E	X		
IDH-R	X		
IDH	X		
IDS	X		
Average lenght from the student to the school			X
Variance related to the distance from the student to the school			X
Average Social Economic Condition of students of a School	X		
Average Distance age-school grades	X		
Anual Cost of the school	X		
Dengue fever by neighborhood	X		
Dengue fever in 500 m radius from the school	X		
Potholes fix requests in 500 m radius from the school	X		
Public Ilumination requests in 500 m radius from the school	X		
Ilegal parking requests in 500 m radius from the school	X		
Security assistance requests per school	X		
Public security ocurrences per nighborhood	X		
Public security ocurrences per census tract		X	
Average leght from the teachers residences to the schools		X	
Familý clinics near each school		X	
Cash transfer program coverage per school		X	
Average precipitation by neighborhood		X	

3.1 Data Sources[3] and indicators

Domain: Student

Datasets:

- Students: 600 registries (academic number, address, assigned school, school address, home latitude, home longitude) Source: Academic Management System Database from Rio de Janeiro's Department of Education
- Social Economic condition: 885 registries, grouped by schools (school identification code, school name, Average SEL – Social Economic Level). Source: www.qedu.org.br/

[3] All the following data refer to 2013 scenarios.

Domain: School

Datasets:

- School: 1450 registries. (school identification number, school name, school address, school IDEB for for 5th grade, school IDEB for 9th grade, school IDERIO for 5th grade, school IDERIO for 9th grade, average annual grade for 5th grade). Source: Academic Management System Database from Rio de Janeiro's Department of Education

- Social Economic Conditions: IDH: 130 registries (IDH-L, IDH-E, IDH-R, IDH)Source: Wikipedia; IDS: 226 registries, Source: Wikipedia

Domain: Surroundings

Datasets:

- 1746 Requests. Numbers by neighborhood and in 500 meters radius outside the school, considering Dengue fever spots, requests for security support at the school, illegal parking, public illumination, potholes and sidewalk repair. Source: 1746 Database

- Police Reports by neighborhood: 152 registries Source: Public Security Institute

- Resources: Annual Cost per School: 1500 registries Source: Strategic Information System from Rio de Janeiro

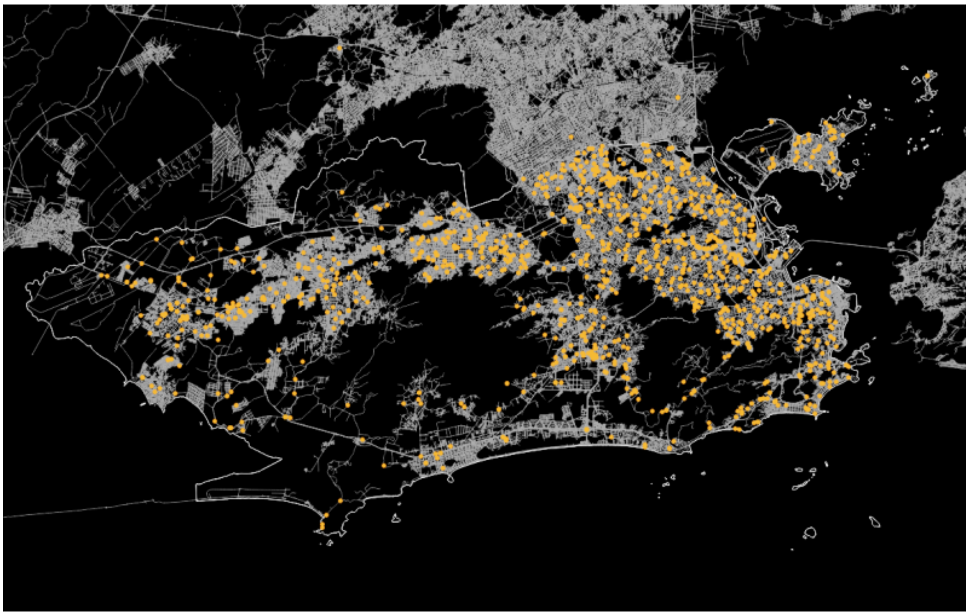
3.2 Crossing Data

Two data groups were crossed separately. First one with the entire sample at disposal. Second one based on a 40 records set, divided on top and bottom 20 IDEB records on 2013. These record sets were processed again, using the average grade on regular tests during 2013 as the performance result.

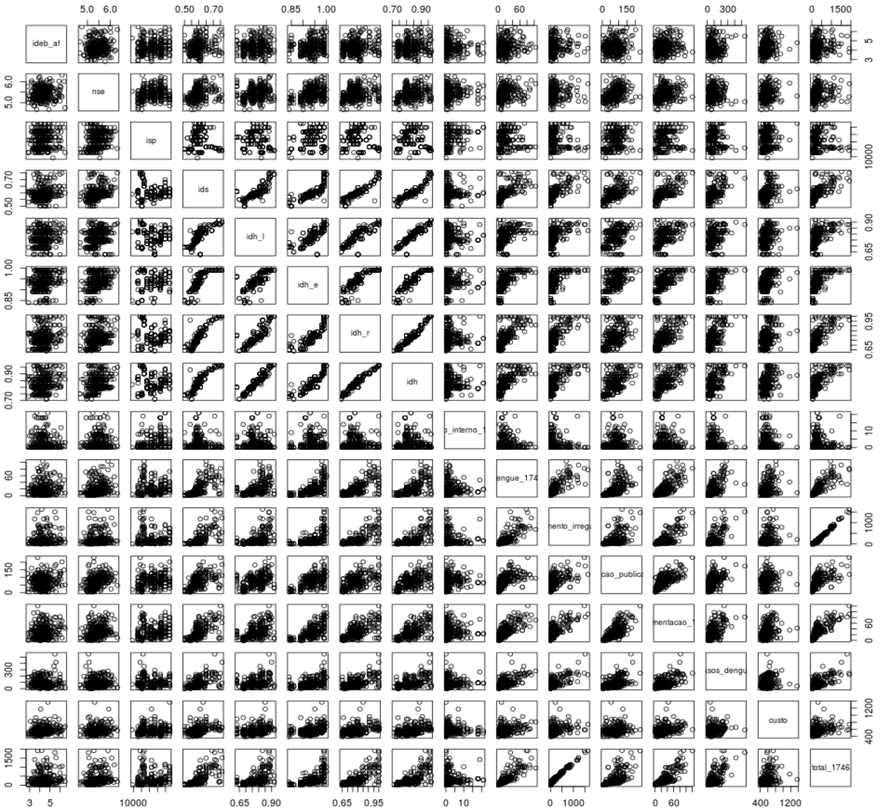
There are fundamental differences between IDEB and the average grade on tests. The first one considers non regular tests applied specifically to build the indicator, drop out records and reprobation as well. The second one consists exclusively on Portuguese language, essay and mathematics exams.

3.3 Full set

There can be no blank spaces on probabilistic analysis. During the process to extract and transform data to make it uniform, some schools didn't have the related data from several sources, narrowing down the potential 1450 registries to 471 registries fully completed.

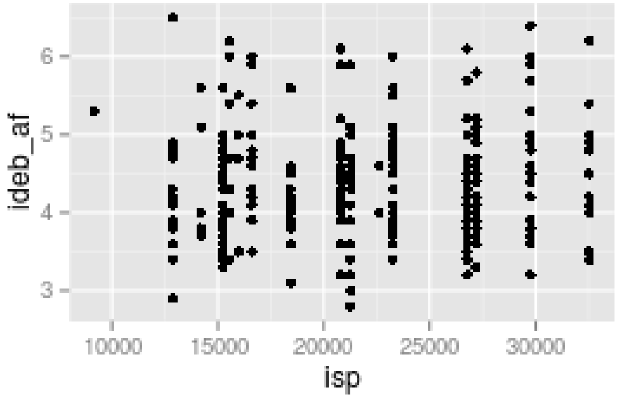


Using Multiple Linear Regressions (MLR) for the indicators referred to the schools, we obtained the results shown below

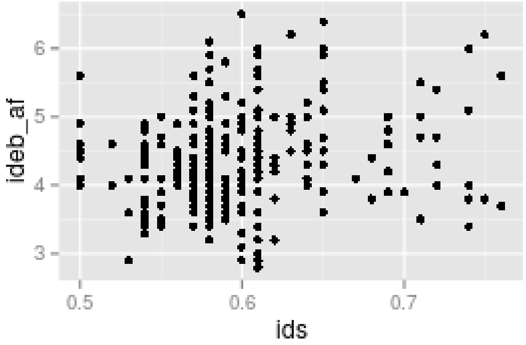


This matrix shows in the diagonal line all the intersected indicators, and we can only notice any correlations between some indicators (especially between IDH and 1746 indicators), but there is no correlation between these indicators and Ideb performance.

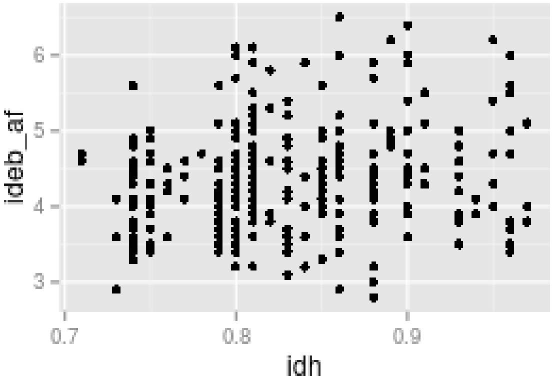
It can be explained in the charts on the next page, in which we cross the variables in pairs with IDEB results:



The graph on the left shows that the police incidences are not distributed linearly, having no influence on display. We believe that this factor is potentially valuable in the analysis, but the data were grouped by areas of 4 or 5 neighborhoods, i.e. with low sectorization, which could point to an influence differentiation in a school and consequently in its performance.

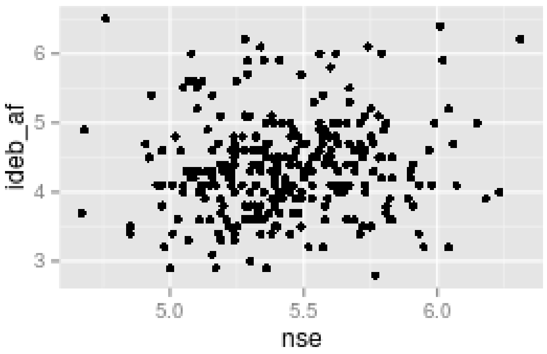


Similar behaviors can be seen in the chart to the right, where the IDEB is compared to the indicator of social development. We understand that in this case, as in the next (Human Development Index), we witness the same phenomenon seen in the police reports, where the grouping of indicators by neighborhood hinders the analysis as it does not influence the school.



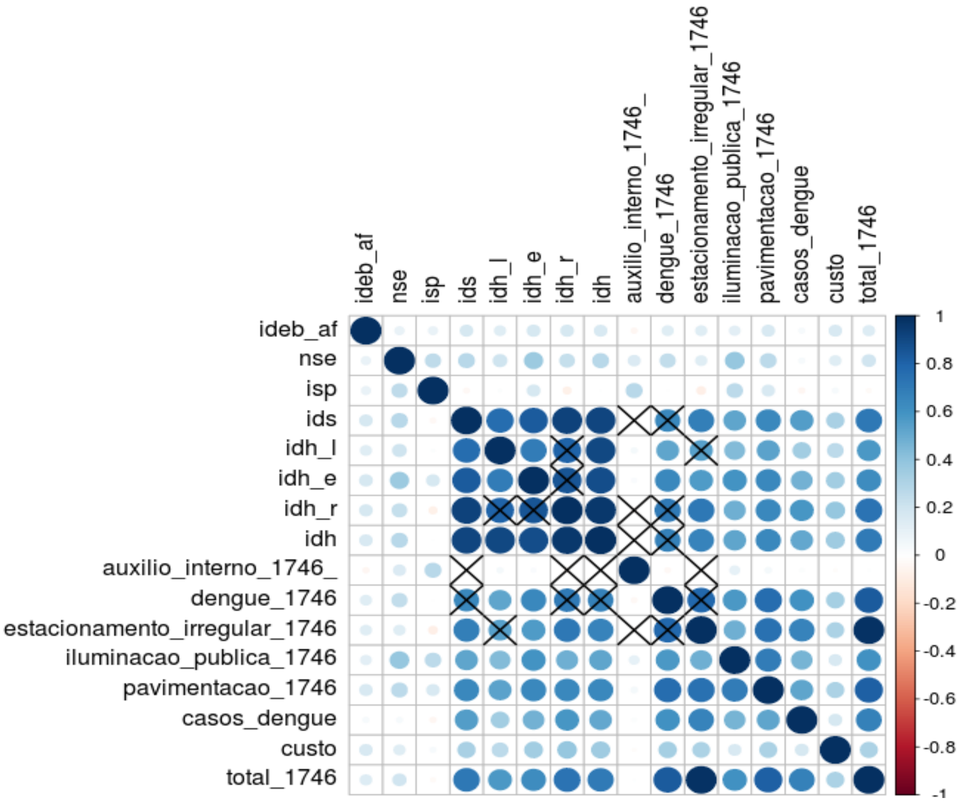
Moreover, unlike instances indicator, both the IDS as the HDI are composed of sub-indicators, more precise to the analysis of specific cases, such as sanitary condition, employability, public resources around, water supply etc. As a suggestion for further developments, one should address the sub-indicators in future crosses, as well as the location of these census tracts in both the school and the student, since it is expected that the influence of the environment will be more active in the house than at school, especially considering that many households are located in the favelas and schools are not.

The distribution in the Social Economic Level (below) begins to behave more similarly to an influenced behavior, even though it's not enough to get to any solid conclusion.



The information of school expenditures does not show any correlation with performance.

Another view of this cross-influence and its relation to the IDEB is as follows, where no significant interrelationships are represented by "X" and the intensity of the significant correlation is demonstrated by color gradation / circle size. Although the correlation is apparent between several variables to each other, again we see that none is linked to the IDEB.



3.4 New Regressions without the non-significant variables

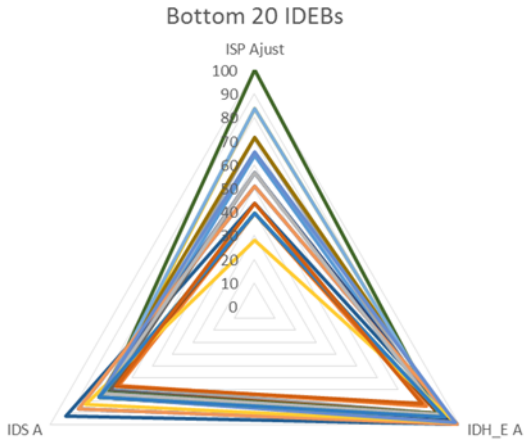
The next step consists of discarding the non-significant variables, so as the variables with inner correlation as well, and process another regression considering only NSE, ISP, IDS, IDH, Dengue fever, Annual Cost and total 1746, but the results remain the same. There aren´t any correlations between those variables.

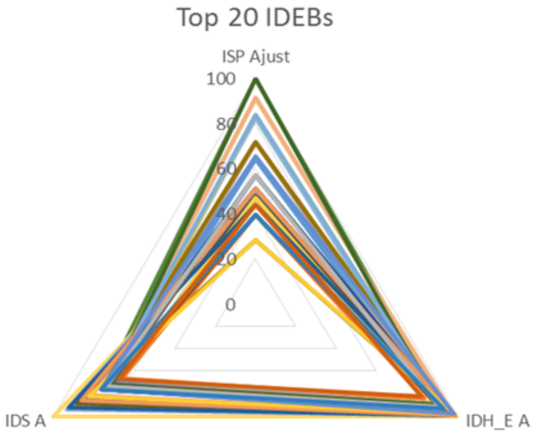
3.5 Another approach with a reduced set of schools

Since we didn’t find any correlation with previous sample, another approach was required to aid the analysis of given variables against the performance one. We added to that dataset the average distance between students and the schools, from now on called “mobility”.

The sample of 20-20 schools was operationalized in separate with the objective of verifying if there was any behavior in the variables that could stand out when we consider two radically opposed groups of schools in terms of performance.

The distance between the students households and the schools didn’t result on any correlation whatsoever.

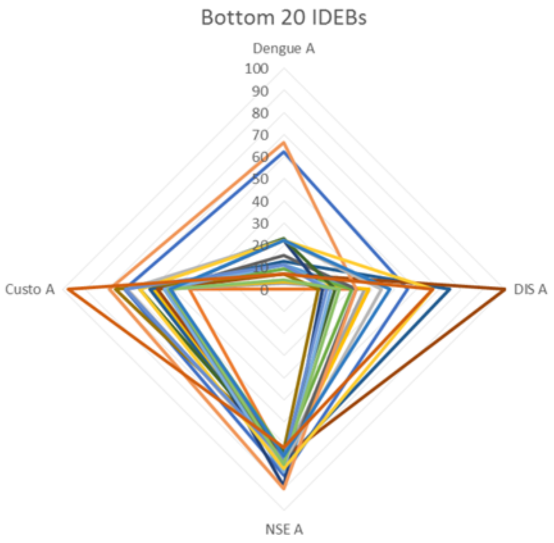


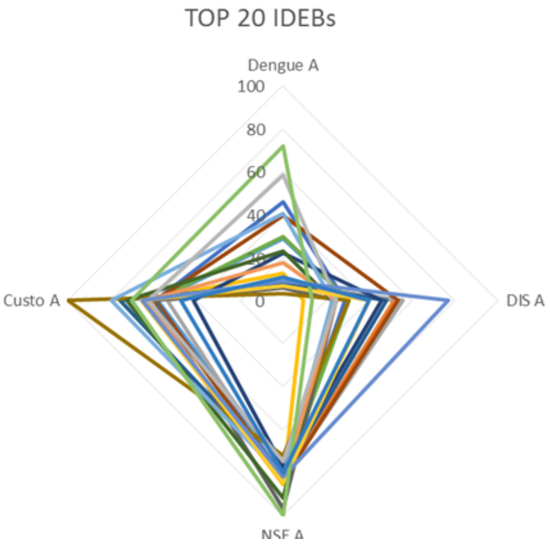


The following images compare the variables behavior of the better and worst performance schools. First we plot those variables that the regression proved to be insignificant in order to reinforce the conclusion.

It´s possible to notice the same behavior either on the top 20 IDEBs and the bottom 20 IDEBs, reassuring the conclusions of the first analysis.

The remaining variables, except the mobility, already evaluated, we have the charts below:





The same conclusions can be extracted from both charts, since there is little difference between the top and bottom 20 schools. The only highlight can be attributed to the age-series gap (DIS) in the case of the top performing schools, but this is not conclusive, since it seems to be an outlier and the small sized sample can lead to wrong conclusions, since there is no concentration. The same can be said about the dengue influence, which surprisingly has more representativity in the top performing schools.

4. Conclusions

In our initial proposal, our goal was to measure indicators on four important domains: Student, Family, School and School Environment (or “surroundings”). However, throughout the development of the project, we were able to access only a limited amount of data, most of them related to the environment and “the world outside the school”.

In the first applied method we find out there were too many indicators that were not really being able to correlate with the performance indicators and that was due to granularity characteristics of those indicator that were grouped by neighborhood or by area (groups of neighborhoods).

In the second applied method, although we have eliminated the weak variables identified in the first method, the results didn’t show any strong correlation with the performance indicators. That result leads us to two hypotheses:

- The environment may not strongly interfere in student performance;

or

- The aggregated indicators used are not good hints of influence in the performance.

In the last method, we compared only few indicators for the top 20 and worst 20 schools ranked in the performance IDEB indicator and the only relevant indicator that has a significant behavior was the “age-grade relation” which pointed to the need for more inner schools and inner classes indicators.

After all these conclusions, we assume that our greatest achievement was to put together for the first time a method and a structure to analyze student performance in a multi area environment. The project involved people from 6 different areas, including the PENSA Group, devoted to BIG DATA analyses that had never initiated a study on education.

5. Recommendations and further developments

The first and more important recommendation to follow with the study is to search for data within the school: its assets, the way they are used, the school climate, school principal performance and parents and community engagement in the school life and development. We also recommend that any environment indicators used should be related just to school surrounding, so schools will not share big area’s indicators that don't reveal their differences and finally we recommend to refrain from using aggregated indicators since they mix too many variables, but to try to analyze specific pure indicators that could lead to a more concrete conclusion.