Using data mining over open data for a longitudinal assessment of municipal public education in Brazil

Arthur Scanoni[1], Rogério Silva Filho[1], Paulo Adeodato[1] and Kellyton Brito[1]

[1]*Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, s/n - Dois Irmãos, Recife - PE, 52171-900, Brazil*

Abstract

The popularization of large-scale assessments in education, such as PISA in Europe and ENEM in Brazil, and the availability of their related data on contemporaneous government open data repositories, have fostered the creation of public value. Much of the current research aims to analyze education at a student level and focuses on high school education, and few studies on earlier educational stages at the municipal level can be found. This paper presents a study focused on analyzing Brazilian educational open data for assessing public education on a municipal level for elementary and middle school education to understand the correlations between contextual indexes and expenditures and educational achievement. For this, we have applied a data mining-based approach and statistical methods to correlate the features and student performance from 2013 to 2019. The main educational results indicate that the highest positive correlations with students' performance are related to teacher training, followed by financial investments. With regard to the Brazilian open data scenario, certain improvements were identified, such as a large amount of available, updated data. However, historical challenges such as a lack of standards and available machine-readable data are still present.

Keywords

Data Mining, Open Data, Education, Brazil

1. Introduction

Since the 1950s [1], governments and society have agreed that transparency, ”the right to know,” and Open Government Data (OGD), which is verified and used by the general population, may bring about many benefits, such as increased accountability and citizen participation. At the beginning of the 2010s, the movement resurfaced with the possibility of using Web 2.0 to publish and consume data, and many initiatives for publishing open data portals were launched. Thus, new benefits such as delivering better public services and increasing government efficiency and effectiveness have been indicated, especially due to the possibility of society analyzing the publicized data for value generation [2]. These benefits are strongly related to the educational context. Since the 1960s [3], research using educational data aimed at the effects of policies has become standard in educational assessment, and the popularization of OGD has provided a boost to this kind of research, thereby promoting the potential of the findings.

EGOV-CeDEM-ePart 2022, September 06–08, 2022, Linköping University, Sweden (Hybrid)

arthur.scanoni@ufrpe.br (A. Scanoni); rlcsf@cin.ufpe.br (R. S. Filho); pjla@cin.ufpe.br (P. Adeodato); kellyton.brito@ufrpe.br (K. Brito)

- (A. Scanoni); 0000-0002-0297-7507 (R. S. Filho); 0000-0002-0406-2474 (P. Adeodato); 0000-0002-6883-8657 (K. Brito)

Within this context, one important data source is related to large-scale educational assessment (LSA). The databases of these exams are a relevant data source for scientific studies, and enable governments to plan, define and validate educational policies. Although exams at higher educational stages are mostly used to analyze educational performance in LSA [4], elementary and secondary stages, when children learn to read and write, and develop their critical and logical thinking, play an essential role in this process. Moreover, most studies either analyze data on the student level, aiming to analyze and predict student performance [4]. Despite being of value, this analysis nonetheless has limits regarding the use of results for understanding and creating policies and actions at higher levels, such as a municipal or state level. Additionally, being able to characterize the relationship between student performances at municipal/state levels and government policies and expenditures at these levels could identify any gaps and thereby help to support the creation of new policies and validations of those that already exist [5].

Within this context, this paper aims to use the Brazilian open educational data from elementary and secondary education on a municipal level in order to understand the relationship between contextual indexes and expenditures, and Brazilian educational achievements from a longitudinal perspective. We have specifically considered the period from 2013 to 2019, and have applied a data mining process and classical statistical methods to seek correlations that could provide a national perspective on how the characteristics and investments of municipalities are correlated with the performance of early-stage students within the complex Brazilian educational system. The analysis includes educational results and a discussion regarding the challenges of the current Brazilian open data scenario.

The remainder of this paper is organized as follows. Section 2 presents the background to the subject. Section 3 presents the methodology, including the research questions and hypotheses, and an overview of the data mining process carried out to answer the questions. Section 4 presents the experiment, followed by Section 5, which contains the results and evaluation. Lastly, Section 6 presents the concluding remarks.

2. Background and Related Works

Understanding the power of educational policies and the performance of educational systems is a permanent research topic [6]. The research arose as a reaction against the pioneering Coleman study [1], which was the first to identify the predominance of student backgrounds in their outcomes and how factors related to schools and policies have a lesser influence. The essential characteristic that differentiates the current from earlier research is the vast amount of available open data on the educational process [7]. One important branch of research is studying the role played by budgets in the educational system. Although there is a sense that more investment implies better outcomes, some studies, such as [5], have shown that, at least for the higher economies, the manner in which resources are used is more important than the amount spent. In fact, some educational systems, as in the case of Brazil, have increased the budget without, however, changing the outcomes [8].

By considering the Brazilian scenario, some studies have attempted to understand the relationship between contextual variables and educational achievement from Brazilian educational open data, such as [9, 10]. They have often relied upon statistical predictive models through the use of regression coefficients and feature importance in order to unravel how variables are correlated with school achievement. They have also focused on predicting the academic performance of individual students in a specific exam. However, these studies often fail to discuss their model assumptions and how this could verify them in practical scenarios. Also, there is no consensus on an appropriate set of variables to be included in their models. Choices are often arbitrary and may suffer certain influences, mainly due to the lack of standards and the decentralized Brazilian open data scenario, as discussed in Subsection 5.3. Moreover, very few studies regarding the Brazilian scenario have explored the expenditure variables, mainly because this information is not easily gathered. In this direction, [11] and [12] found that expenditure has no or little contribution to increasing test scores, but [13] found that ”financial resources are paramount in producing performance.”

Analyzing the studies, it is clear that there are two types of research. One group used the student and school characteristics to analyze student and school performances, and the other group used financial and expenditure data to analyze the municipal performance. However, some school characteristics may be aggregated to a municipal level, such as teacher data. Hence, these characteristics may be studied jointly with financial data to provide a new viewpoint, capable of supporting policies at different levels.

3. Methodology

The main objective of this research is to use open and publically available data to analyze the characteristics of municipalities and school management, seeking to find correlations between these characteristics and the performance of public-school students. For this, we defined three research questions: RQ1: *Is it possible to define and apply a data mining methodology over open data to find correlations between the main characteristics and investments of municipalities and student performance in large-scale assessments in Brazil?; RQ2: What are the main correlations between the characteristics and investments of municipalities and student performance in large-scale assessments?; and RQ3: How have the correlations between the characteristics and investments of municipalities and student performance varied over time?*

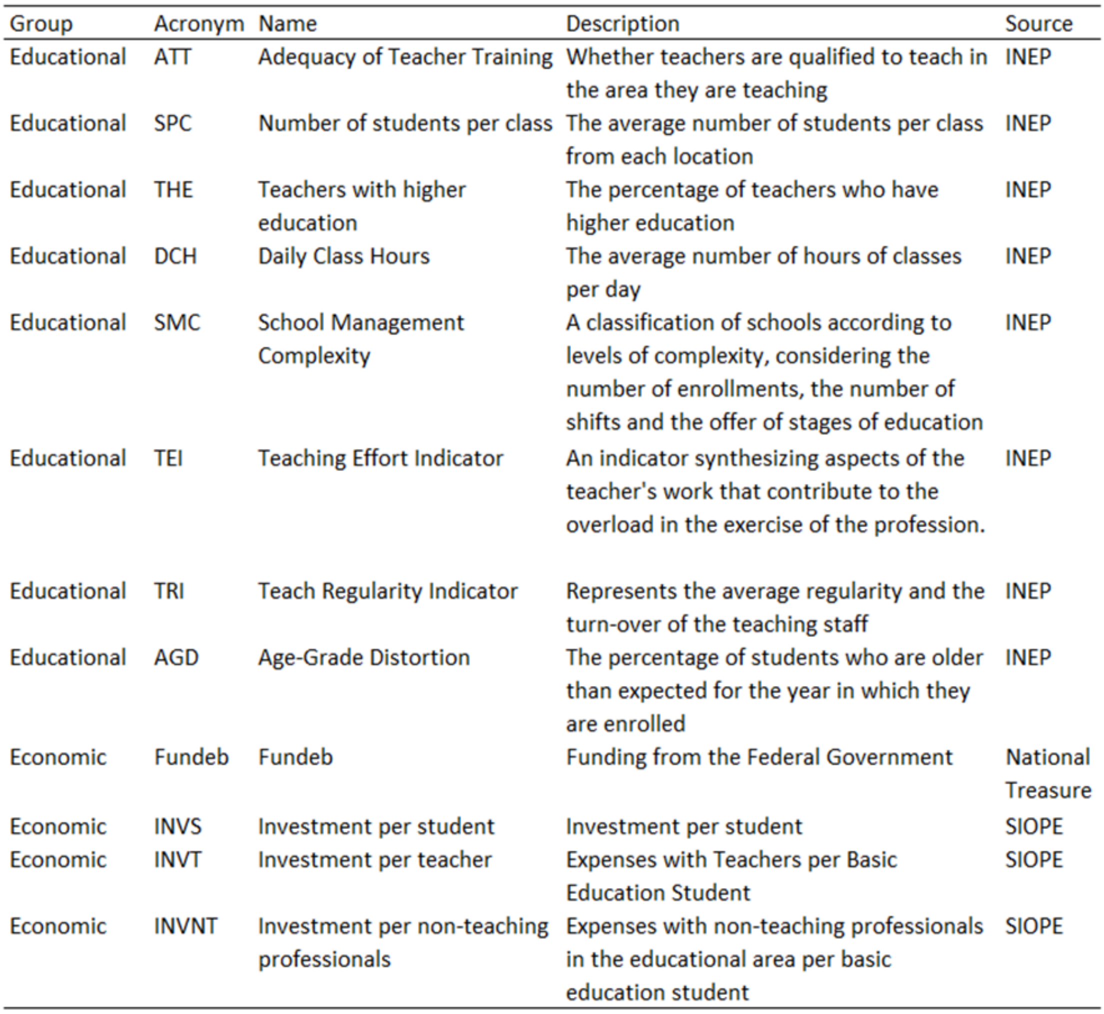
To answer RQ2 and RQ3, we defined a DM process inspired by the well-known CRISP-DM method [14]. Thus, if these questions can be answered, it implies that RQ1 is also answered. The defined process contains five phases: (i) domain understanding; (ii) data collection and understanding; (iii) data preparation; (iv) modeling and analysis; and (v) results reporting and evaluation. Section 4 presents the experiments, with details and implementation of the methodology, from domain understanding to modeling and analysis. The last step is presented in Section 5.

4. Experiments

Domain Understanding: In Brazil, when considering the teaching stage, they may be elementary school (1st to 9th grades) and high school (10th to 12th grades). The elementary school is divided in two stages: initial years (1st to 5th grades) and final years (6th to 9th grades). For a clearer comprehension, in this paper we have named the stages as elementary school, middle school, and high school. With a few minor exceptions, it is mainly the municipalities that are responsible for elementary and middle schools, while state governments are responsible for high schools, and the federal government is responsible for universities. The private sector may be present at any level, but it is not the subject of this study. The most studied LSA in Brazil is the ENEM, which is used for entry to most universities. However, the assessment of early education is still underexplored. For this, there is the IDEB, an education development index. It is based on two main metrics: each school's pass and abandonment rates and the performance of students in a specific exam, called "Prova Brasil". The IDEB has three distinct indexes, one for elementary school, one for middle school, and one for high school. Thus, in order to evaluate municipal education, which is responsible for elementary and middle schools, we have used the respective IDEB index for these two stages.

Data Collection and Understanding: We have focused data collection on the granularity of municipalities, rather than students, and relevant data were divided into two groups, educational data and economic data. The educational data, including school data and IDEB indicators, were collected from the INEP portal [15]. Economic data were collected from the National Treasury website and the SIOPE website [16]. Figure 1 presents a list of the collected data used as features. In addition, the performance indicator, IDEB, was also collected from the INEP portal. Most features presented in Figure 1 have two values, one for the elementary stage and another for the middle stage. The corresponding value will be used in the analysis of each stage. Also, some of the features are continuous and may be directly understood, such as the number of students per class (SPC). However, some of the features are divided into subcategories and present discrete values for each of the subcategories. These categories are defined according to metrics defined by the Brazilian government. The ATT classifies schools into 5 groups, ranging from G1 (percentual of teachers who have a teaching degree in the same subject they teach) to G5 (teachers without a college degree). The SMC ranges from G1 (low complexity) to G6 (high complexity)., TEI ranges from G1 (low effort) to G6 (high effort), and TRI ranges from G1 (low regularity) to G6 (high regularity). As the IDEB is calculated every two years, data were collected for the years 2013, 2015, 2017 and 2019 for each of the 5,568 Brazilian municipalities. Each feature presented in Figure 1 is published as a different dataset for each year. The Brazilian government provides all of the indicators over its open data portals already listed.

Data Preparation: INEP data is in spreadsheet format (XLSX and ODS) and, despite being complete, contains unnecessary data, such as a header with the INEP logo and merged lines and columns. Hence, we manually cleaned spreadsheets before they could be processed. INEP data is also usually presented in different granularities, such as aggregations of the country/regions/states, aggregations of municipalities, and individual schools. It also contains data for each separate stage, such as elementary, middle, and high school. For the objectives of this study, we filtered and used the aggregations by municipalities, and data regarding the elementary and middle stages. The National Treasure data are presented with two values, the value initially transferred plus a correction that may be available later. In the study, we considered the sum of both as the absolute value of spending. For SIOPE data, it was not necessary to edit the data since each one has only the value of each municipality. Fortunately, municipalities are identified by a unique and equal ID across the different datasets. Thus, after cleaning and filtering the data from the three different sources, it was possible to create a unique database joining all of them together.



Group	Acronym	Name	Description	Source
Educational	ATT	Adequacy of Teacher Training	Whether teachers are qualified to teach in the area they are teaching	INEP
Educational	SPC	Number of students per class	The average number of students per class from each location	INEP
Educational	THE	Teachers with higher education	The percentage of teachers who have higher education	INEP
Educational	DCH	Daily Class Hours	The average number of hours of classes per day	INEP
Educational	SMC	School Management Complexity	A classification of schools according to levels of complexity, considering the number of enrollments, the number of shifts and the offer of stages of education	INEP
Educational	TEI	Teaching Effort Indicator	An indicator synthesizing aspects of the teacher's work that contribute to the overload in the exercise of the profession.	INEP
Educational	TRI	Teach Regularity Indicator	Represents the average regularity and the turn-over of the teaching staff	INEP
Educational	AGD	Age-Grade Distortion	The percentage of students who are older than expected for the year in which they are enrolled	INEP
Economic	Fundeb	Fundeb	Funding from the Federal Government	National Treasure
Economic	INVS	Investment per student	Investment per student	SIOPE
Economic	INVT	Investment per teacher	Expenses with Teachers per Basic Education Student	SIOPE
Economic	INVNT	Investment per non-teaching professionals	Expenses with non-teaching professionals in the educational area per basic education student	SIOPE

Figure 1: Features Description

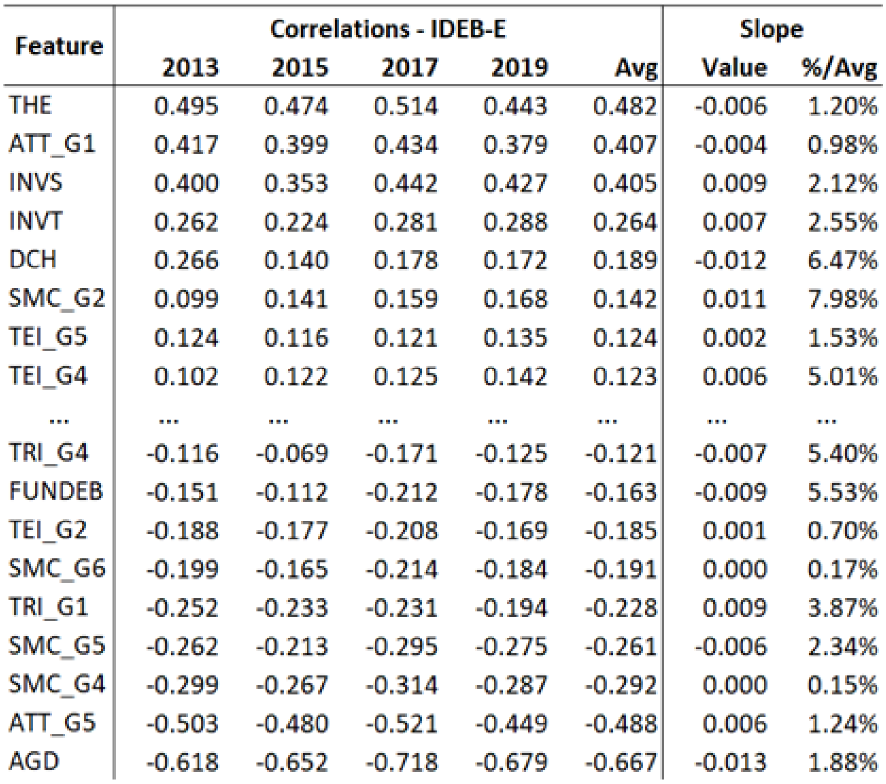
Modeling and Analysis: All features were analyzed individually and correlated with the two performance features: IDEB elementary school (IDEB-E) and IDEB middle school (IDEB-M). We used the Spearman's correlation, which evaluates the monotonic relationships between the variables. It is most suitable in this context since it is not affected by global changes, for example, an increase in IDEB results due to changes in methodology or an overall improvement of education in the country [17]. It is well known that results in exams are the result of several factors, including a diversity of topics such as financial status, parents' level of scholarity and personal effort, as presented in Section 2. Thus, it not expected to find that municipality indexes, characteristics or investments in education present high correlation values. There are also several types of interpretations for the correlation values, and we decided to follow the same interpretations as the Mukaka study [18], i.e., to be a considerable correlation, the absolute value must be greater than 0.3. Moreover, due to the expected low values, we also present and analyze the variable with at least the 5 highest positive and negative correlations.

In order to answer RQ2, we calculated the correlations of every feature for every year to analyze which features were relevant in each year. We then averaged the results for each feature to have a consistent value among the years. Next, to answer RQ3, we calculated a linear regression on the values over the four years, and identified the percentage of the slope compared to the mean of the values. The result is the assessment of the strength, or ”the velocity”, in which the correlation increases or decreases during the period.

5. Results and Evaluation

5.1. Correlation results and discussion

As planned, we calculated the Spearman correlations coefficient for each feature in relation to the performance indicators, IDEB-E and IDEB-M, and averaged the results obtained over the years. Unsurprisingly, most of the features with high correlations, both positive and negative, are the same for both elementary and middle schools. Figure 2 presents the data.

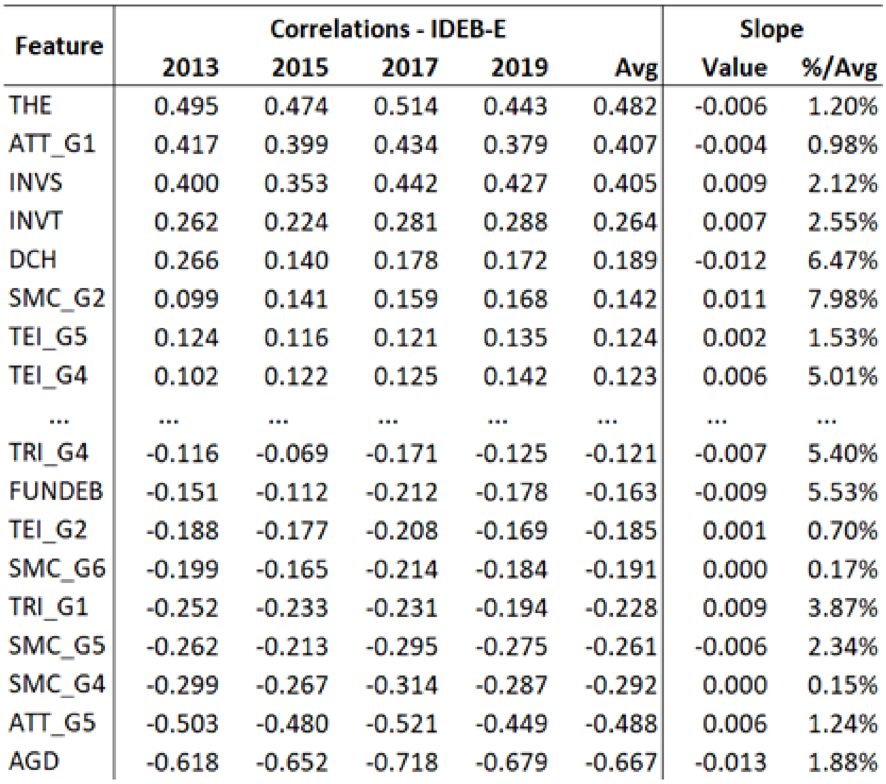


Feature	Correlations - IDEB-E					Slope	
	2013	2015	2017	2019	Avg	Value	%/Avg
THE	0.495	0.474	0.514	0.443	0.482	-0.006	1.20%
ATT_G1	0.417	0.399	0.434	0.379	0.407	-0.004	0.98%
INVS	0.400	0.353	0.442	0.427	0.405	0.009	2.12%
INVT	0.262	0.224	0.281	0.288	0.264	0.007	2.55%
DCH	0.266	0.140	0.178	0.172	0.189	-0.012	6.47%
SMC_G2	0.099	0.141	0.159	0.168	0.142	0.011	7.98%
TEI_G5	0.124	0.116	0.121	0.135	0.124	0.002	1.53%
TEI_G4	0.102	0.122	0.125	0.142	0.123	0.006	5.01%
...	...	...	...	...	...	...	...
TRI_G4	-0.116	-0.069	-0.171	-0.125	-0.121	-0.007	5.40%
FUNDEB	-0.151	-0.112	-0.212	-0.178	-0.163	-0.009	5.53%
TEI_G2	-0.188	-0.177	-0.208	-0.169	-0.185	0.001	0.70%
SMC_G6	-0.199	-0.165	-0.214	-0.184	-0.191	0.000	0.17%
TRI_G1	-0.252	-0.233	-0.231	-0.194	-0.228	0.009	3.87%
SMC_G5	-0.262	-0.213	-0.295	-0.275	-0.261	-0.006	2.34%
SMC_G4	-0.299	-0.267	-0.314	-0.287	-0.292	0.000	0.15%
ATT_G5	-0.503	-0.480	-0.521	-0.449	-0.488	0.006	1.24%
AGD	-0.618	-0.652	-0.718	-0.679	-0.667	-0.013	1.88%

Figure 2: Average Correlations

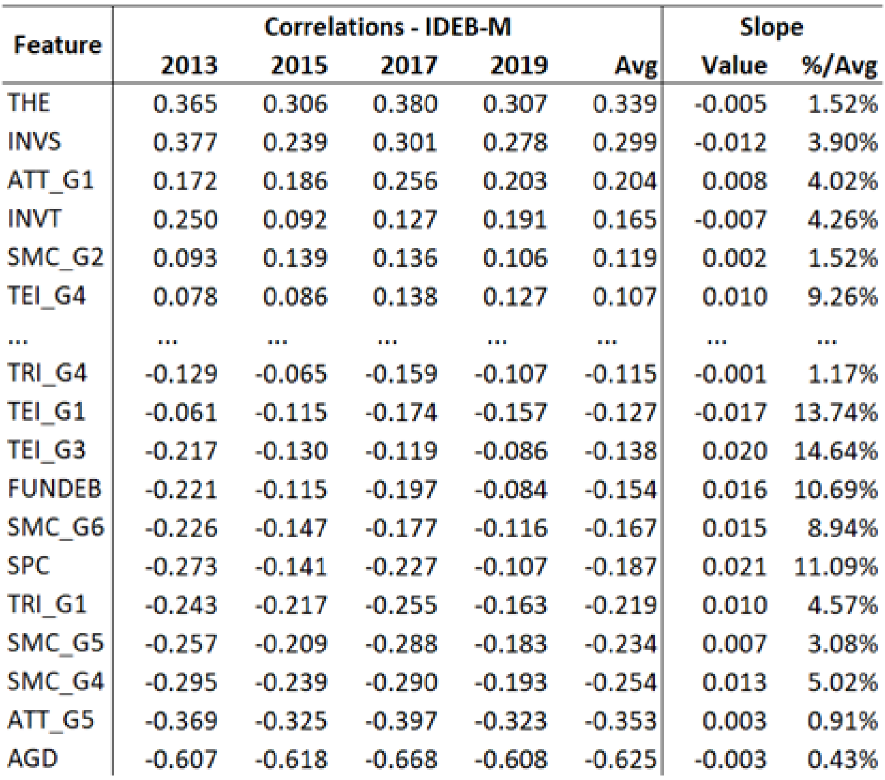
When considering the positive correlations, the data indicated that student performances in a municipality were better when the teachers were more highly qualified: the percentage of teachers with higher education (THE) and the adequacy of teacher training (ATT) at a higher level (G1), presented the higher correlations. It is followed by investment: first the investment per student (INVS), then the investment per teacher (INVT). On the other hand, the feature presenting a high negative correlation in both scenarios was the age-grade distortion (AGD), i.e., in a municipality where a high number of students was out of the expected class, the educational performance was worse. It is important to note that, as correlation is not causality, it may not be assured whether the age-grade distortion is the cause or, most probably, the consequence of a bad student performance. The second higher negative correlation is the opposite of the positive correlation, the number of teachers with no college degree (ATT_G5), followed by the number of schools with a high management complexity (SMC_G4 and SMC_G5). Lastly, also associated with the teacher, (TRI) at the lower level, representing municipalities where the turnover is high, also presented a negative correlation with student performance.

As planned, we also analyzed the **behaviour of features over time**. Taking the correlation values for each year, considering both the IDEB-E and IDEB-M, we calculated the slope of their linear regression and analyzed the trend of each feature over time. With this, we calculated the percentage of variation in relation to the mean, thereby obtaining the tendency if the correlation is consistently increasing or decreasing. To be able to get more insights, we broadened the analysis all variables with p > 0.10. Figures 3 and 4 present the data.



Feature	Correlations - IDEB-E					Slope	
	2013	2015	2017	2019	Avg	Value	%/Avg
THE	0.495	0.474	0.514	0.443	0.482	-0.006	1.20%
ATT_G1	0.417	0.399	0.434	0.379	0.407	-0.004	0.98%
INVS	0.400	0.353	0.442	0.427	0.405	0.009	2.12%
INVT	0.262	0.224	0.281	0.288	0.264	0.007	2.55%
DCH	0.266	0.140	0.178	0.172	0.189	-0.012	6.47%
SMC_G2	0.099	0.141	0.159	0.168	0.142	0.011	7.98%
TEI_G5	0.124	0.116	0.121	0.135	0.124	0.002	1.53%
TEI_G4	0.102	0.122	0.125	0.142	0.123	0.006	5.01%
...	...	...	...	...	...	...	...
TRI_G4	-0.116	-0.069	-0.171	-0.125	-0.121	-0.007	5.40%
FUNDEB	-0.151	-0.112	-0.212	-0.178	-0.163	-0.009	5.53%
TEI_G2	-0.188	-0.177	-0.208	-0.169	-0.185	0.001	0.70%
SMC_G6	-0.199	-0.165	-0.214	-0.184	-0.191	0.000	0.17%
TRI_G1	-0.252	-0.233	-0.231	-0.194	-0.228	0.009	3.87%
SMC_G5	-0.262	-0.213	-0.295	-0.275	-0.261	-0.006	2.34%
SMC_G4	-0.299	-0.267	-0.314	-0.287	-0.292	0.000	0.15%
ATT_G5	-0.503	-0.480	-0.521	-0.449	-0.488	0.006	1.24%
AGD	-0.618	-0.652	-0.718	-0.679	-0.667	-0.013	1.88%

Figure 3: Evolution of correlations over time, IDEB-E



Feature	Correlations - IDEB-M					Slope	
	2013	2015	2017	2019	Avg	Value	%/Avg
THE	0.365	0.306	0.380	0.307	0.339	-0.005	1.52%
INVS	0.377	0.239	0.301	0.278	0.299	-0.012	3.90%
ATT_G1	0.172	0.186	0.256	0.203	0.204	0.008	4.02%
INVT	0.250	0.092	0.127	0.191	0.165	-0.007	4.26%
SMC_G2	0.093	0.139	0.136	0.106	0.119	0.002	1.52%
TEI_G4	0.078	0.086	0.138	0.127	0.107	0.010	9.26%
...	...	...	...	...	...	...	...
TRI_G4	-0.129	-0.065	-0.159	-0.107	-0.115	-0.001	1.17%
TEI_G1	-0.061	-0.115	-0.174	-0.157	-0.127	-0.017	13.74%
TEI_G3	-0.217	-0.130	-0.119	-0.086	-0.138	0.020	14.64%
FUNDEB	-0.221	-0.115	-0.197	-0.084	-0.154	0.016	10.69%
SMC_G6	-0.226	-0.147	-0.177	-0.116	-0.167	0.015	8.94%
SPC	-0.273	-0.141	-0.227	-0.107	-0.187	0.021	11.09%
TRI_G1	-0.243	-0.217	-0.255	-0.163	-0.219	0.010	4.57%
SMC_G5	-0.257	-0.209	-0.288	-0.183	-0.234	0.007	3.08%
SMC_G4	-0.295	-0.239	-0.290	-0.193	-0.254	0.013	5.02%
ATT_G5	-0.369	-0.325	-0.397	-0.323	-0.353	0.003	0.91%
AGD	-0.607	-0.618	-0.668	-0.608	-0.625	-0.003	0.43%

Figure 4: Evolution of correlations over time, IDEB-M

By analyzing data in Figures 3 and 4, most of the features present a non-monotonic variation, i.e., they neither strictly increase nor decrease, and do not strictly present a trend. Also, none of the five features with the highest correlations (positive or negative) presented a percentage slope variation higher than 10%, demonstrating that the most correlated features remained stable. However, some trends may be found, allowing additional analysis. The percentage of schools in a city (SMC) with a low management complexity (level 2 of 6) presented a monotonic increase in relation to IDEB-E, suggesting that having more schools with less students (50-300) is increasing its importance. Also, the TRI_G1 presented an absolute monotonic decrease. This signifies that, if the trend continues, the negative correlation of the teacher regularity indicator (TRI) will soon no longer exist. Additionally, the result of a positive correlation of TEI_G4, a high effort group, consistently increasing, in opposition to the negative correlations of the TEI at lower levels, was unexpected. This result must be carefully assessed in further studies, although an initial reason may be put forward. The lower level groups include teachers working only one daily period, which may indicate that teaching may not be their primary activity.

Presented data answers RQ2 and RQ3 by using the defined data mining methodology over open data, thereby answering RQ1 affirmatively.

5.2. Open data discussion

Four years after being one of the founders of the Open Government Partnership (OGP), [19] indicated some of the challenges of the Brazilian scenario. We may summarize these challenges at that time as: (i) a lack of available data, because of the small quantity of datasets published in the repositories; (ii) multiple and decentralized data sources, because although a national open-data portal exists, it is incomplete and there are many other state and municipal repositories not integrated within it; (iii) zombie data, without an update policy; (iv) a lack of standards for publishing, because each variety of publisher chooses what and how to publish, as well as the data format; and (v) one-way data, not allowing citizens to return data to the government.

In this study we have verified that, considering the current state of educational open data, some of these challenges have been solved, and others still remain. As improvements, we highlight that there is no lack of available educational data, since we were able to collect the data easily with regard to all municipalities in Brazil. Also, we cannot consider the data as zombie data, because despite a certain amount of delay in publishing, since they are released annually, all the expected and updated data were found. As challenges, we indicate that there are still multiple and decentralized data sources and a lack of standards for publishing. Even being published by the Federal government, the gathered data was not found in the national open data repository, but on three different, independent sources, each with a different format and standards. In particular, the data gathered from SIOP is not even found on an open or transparency portal: it was gathered from a system that must be operated by an individual who has to access the ”management reports” option, filter the desired data, and then performs the download. Moreover, most of the collected data is ”almost” machine readable, one of the premises of open data. INEP data is in spreadsheet format (XLSX and ODS) and, despite being complete, it contains unnecessary data, such as a header with INEP logo and a variety of merged lines and columns. Hence, it is necessary to manually clean the spreadsheets before they may be processed. Lastly, almost all the collected data can not be found or downloaded without human intervention, and was not available to be accessed from an API (Application Programming Interface).

Considering these challenges, we argue that many improvements still need to be implemented in order to enhance the Brazilian open data scenario and foster the potential of using government data.

6. Concluding Remarks

This paper has presented a data mining approach over educational open data for a longitudinal assessment of municipal public education in Brazil, by finding correlations among the characteristics and investments of municipalities and student performances in large-scale assessments. For this, we defined a methodology which consisted of five steps: (i) domain understanding; (ii) data collection and understanding; (iii) data preparation; (iv) modeling and analysis; and (v) results reporting and evaluation. Data was collected from Brazilian open data portals, including educational and financial data, and data from the performance of students at a municipal level, including the 5,568 municipalities for the period from 2013 to 2019. Spearman correlations were calculated in order to find correlations, and the regression slope over the period was calculated and analyzed in order to verify the trends.

The main results indicate that the highest positive correlations are linked to teacher training, followed by financial investments in students and teachers. On the other hand, the age-grade distortion presents a high negative correlation. All of the highest correlations remained stable over time, but correlations regarding the lower level of school management complexity and the number of full-time teachers present a positive trend.

Some improvements were also identified in the Brazilian open data scenario, such as a large amount of available, updated data. However, historical challenges, such as multiple, decentralized data sources, the lack of standards for publishing, and the existence of datasets that are not machine readable, and thus require human intervention, are still challenges.

We argue that correlation is not causality. Also, we recognize that many other features may be correlated and influence student performance, not only those studied. Thus, the results of this study must be analyzed with caution. Future work may focus on identifying why some features present a higher correlation with student performance, and how public policy may be modeled to boost this performance. Furthermore, Brazil is a large country, well known for having regional differences, but data from all municipalities were analyzed together. Thus, future studies focused on identifying the regional differences may be suitable, as well as studies comparing Brazilian results with results of other countries, especially in Latin America. Lastly, future studies using multivariate correlations and machine learning models for simulating the prediction of student performance, based on the municipalities features, may be promising.

References

[1] W. Parks, Open government principle: Applying the right to know under the constitution, Georg. Wawhingt. Law Rev 26 (1957).

[2] M. Janssen, Y. Charalabidis, A. Zuiderwijk, Benefits, adoption barriers and myths of open data and open government, Information Systems Management 29 (2012) 258–268. doi:10.1080/10580530.2012.716740.

[3] J. Coleman, Equality of educational opportunity, Equity Excell. Educ 6 (1968) 19–28. doi:10.1080/0020486680060504.

[4] S. Salloum, M. Alshurideh, A. Elnagar, K. Shaalan, Mining in educational data: Review and future directions, in: Proceedings of the International Conference on Artificial Intelligence and Computer Vision, volume AICV2020, 2020, p. 92–102. doi:10.1007/978-3-030-44289-7_9.

[5] O.E.C.D., Does money buy strong performance in pisa?, PISA Focus 13 (2012) 4,.

[6] D. Hernández-Torrano, M. Courtney, Modern international large-scale assessment in education: an integrative review and mapping of the literature, Large-scale Assessments Educ 9 (2021-12) 17,. doi:10.1186/s40536-021-00109-1.

[7] C. Fischer, Mining big data in education: Affordances and challenges, Rev. Res. Educ 44 (2020) 130–160. doi:10.3102/0091732X20903304.

[8] P. Adeodato, R. Filho, Where to aim? factors that influence the performance of brazilian secondary schools, in: Proceedings of The 13th International Conference on Educational Data Mining, 2020, p. 5.

[9] C. Gomes, E. Jelihovschi, Presenting the regression tree method and its application in a large-scale educational dataset, Int. J. Res. Method Educ 43 (2020) 201–221. doi:10.1080/1743727X.2019.1654992.

[10] R. Filho, P. Adeodato, K. Santos Brito, Interpreting classification models using feature importance based on marginal local effects, 2021.

[11] M. Haddad, R. Freguglia, C. Gomes, Public spending and quality of education in brazil, J. Dev. Stud 53 (2017-10) 1679–1696. doi:10.1080/00220388.2016.1241387.

[12] A. Santos, F. Medeiros, Relationship of federal funding to ideb results in a state in brazil: an approach based on educational data mining, in: 2020 15th Iberian Conference on Information Systems and Technologies (CISTI, 2020, p. 1–4. doi:10.23919/CISTI49556.2020.9140924.

[13] A. R. M. Valle, R. Gomes, Analyzing the importance of financial resources for educational effectiveness, Int. J. Product. Perform. Manag 63 (2014-01) 4–21. doi:10.1108/IJPPM-08-2012-0085.

[14] C. Shearer, The crisp-dm model: the new blueprint for data mining, J. data Warehous 5 (2000) 13–22.

[15] Instituto nacional de estudos e pesquisas educacionais anísio teixeira, "dados abertos - inep, 2022. URL: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos.

[16] Fundo Nacional Educação, Sistema de informações sobre orçamentos públicos em educação, 2022. URL: https://www.fnde.gov.br/siope/relatorio-gerencial/dist/indicador.

[17] P. Schober, C. Boer, L. Schwarte, Correlation coefficients, Anesth. Analg 126 (2018-05) 1763–1768. doi:10.1213/ANE.0000000000002864.

[18] M. Mukaka, A guide to appropriate use of correlation coefficient in medical research, Malawi Med. J 24 (2012).

[19] K. S. Brito, M. S. Costa, V. Garcia, S. L. Meira, Is brazilian open government data actually open data? an analysis of the current scenario, Int. J. E-Planning Res 4 (2015-04) 57–73. doi:10.4018/ijepr.2015040104.