Correlational Analysis Between School Performance and Municipal Indicators in Brazil Supported by Linked Open data

Bruno Elias Penteado
University of São Paulo
ICMC - Institute of Mathematics and
Computer Science
brunopenteado@usp.br

ABSTRACT

The advance in quality of public education is a challenge to public managers in contemporary society. In this sense, many studies point to the strong influence of socioeconomical factors in school performance but it is a challenge to select proper data to perform analyses on this matter. In tandem, it has happening a growth in provision of big quantities of educational indicators data, but in isolate cases, and by different agencies of Brazilian government. For this work, we use both education and economic indicators for analysis. The following socioeconomical indicators were selected: municipal human development index (MHDI), social vulnerability index (SVI), Gini coefficient and variables extracted from DBpedia, as part of the connection of this data to the Web of data: GDP per capita and municipal population. These data were used as independent variables to look into their correlations with Brazilian Basic Education Development Index (IDEB) performances at municipal level, supported by the application of linked open data principles. OpenRefine was used to extract the data from different sources, convert to RDF triples and then the mapping of the variables to existing ontologies and vocabularies in this domain, aiming at the reuse of existing semantics. The correlational analysis of the variables showed coherence with the literature about the theme, with significative magnitude between IDEB performances and the indicators related to income and parent education (SVI and HDI), besides moderate relations with the other varibles, except for the municipal population. Finally, the consolidated dataset, enriched by information extracted DBpedia was made available by a SPARQL endpoint for queries of humans and software agents, allowing other applications and researchers to explore the data from other platforms.

Keywords

LOD, Linked open data, educational indicators, school performance.

1. INTRODUCTION

Educational quality is a matter of great importance to public managers and society at large. Although the term 'educational quality' is quite polemic, the scope adopted in this work is restricted to school performance, i.e., the academic performance, quantitatively measured by statistical indicators composed by scores in large scale assessments, leaving other wider educational aspects out.

These indicators were created at federal level in the end of the 90's and beginning of 2000's as a form of quantifying the status and advances of teaching in the country, in different administrative spheres. Since then, a great concern for public educational managers has been how to measure and improve this academic attainment in their educational system, giving rise to some school performance indicators, such as the IDEB (Brazilian Basic Education Development Index) at federal level. The data is collected from predefined criteria of scope and periodicity and are used as primary source for establishing public policies, educational research in different subareas and for the society. On the other hand, there is the collection and provisioning of many other data, in different spheres other than education, such as demographic, economic, health, and environmental data, among others. However, such rich datasets are mostly available in isolation, with each dataset in its delimited scope, some with files dependent on proprietary standards and not suited for automated queries [1][2].

With this scenario, this work aims to gather data available in the Web about school performance in public system, which guide public policies, such as the IDEB, and make them available using linked open data principles, having the municipalities as the level of granularity. In addition, also gathers several municipal indicators that can be used to better contextualize the respective performances in function of the socioeconomical reality faced by the schools. Based on that, the dataset is extended to use other information available in DBpedia. As a case study, the variables of academic performance are used as dependent variables and the socioeconomic variables are treated as independent variables in order to analyze their correlations. Thus, this tool can be useful for both public managers at federal and state level, for data analysis and decision-making and researchers in public management or educational policies fields. This work brings two contributions: it exposes the aggregated data from these different and important data sources in linked open data format and the correlational analysis of these socioeconomic indicators of every municipality in Brazil and their academic performances, fundamental to the creation of public policies.

In section we 2 present a contextualization of school performance and the main municipal socioeconomic indicators. Section 3 reviews related literature. Section 4 presents the details of the material used, their data sources, formats, problems found and the methodological approach. Section 5 shows the results of na application over the data as well as the modeling and publishing of the data as linked open data. Finally, section 6 discusses the results and possible extensions for this work.

2. CONTEXTUALIZATION

An *indicator* can be thought of as a measure, usually quantitative, with a social sense, used to replace, quantify or operationalize an abstract social concept, of theoretical (for academic research) or programatic interest (for policy formulation) [3]. The indicators chosen for this work are detailed below.

IDEB

The brazilian federal government assesses the quality of basic education (similiar to US K-12) primarily by the Basic Education Development Index, or IDEB, in Portuguese acronym. IDEB was created in 2007, as part of the federal Education Development Plan. The main goal was to create a synthetic indicator of educational quality, based on two important components: *academic approval rate*, obtained from brazilian School Census[1] and the *Prova Brasil scores* – large scale assessment applied in the end of each academic cycle (5th, 9th and 12th grades) in mathematics and reading, every two years (odd years). The schools selected for the assessment are those with at least 20 students enrolled in those respective grades. Finally, IDEB considers both components and transforms them into a scale ranging from 0 to 10 – similar to an academic grade, for an easier interpretation – trying to balance one component controllable by the school (approval rate) with another external evaluation (large scale assessment). Thus, if a school chooses to pass all their students without actual learning, this will become clear by analyzing its performance in the IDEB [4].

Education indicators such as the IDEB are attractive because they allow monitoring the country's education system. Its importance, in terms of diagnostics and guidance of political actions focused in education system improvement, is in: a) detect schools and/or education systems whose students show lower performance in terms of attainment and proficiency; b) monitor temporal evolution of these system's students [5], on a scale that allows longitudinal comparison.

The Bases and Guidelines for Education Act (LDB) [6], in its fifth paragraph, states the responsibilities of municipalities, in collaboration with the federal and state governments, for basic education by determining that it must provide and watch over the access and maintenance of population in school age for preschool, kindergarten and primary education. States are responsible for secondary education and Federal government has the role of organizing the education system as a whole and regulate higher education.

Socioeconomic indicators

There is a certain degree of consensus, in the filed of education policies, on the factor that most influences academic performance of a student is his socioeconomic status, in terms of cultural and economic capital. In the study known as *Coleman Report* [7], multiple US public schools were researched by applying tests and contextual questionnaires, in order to understand educational opportunities to different ethnic groups in the country. One of the results was the strong relation of socioeconomic level and the school performance. Other studies in this sense were carried out in Brazil, with similar results [8]. It should be noted, however, that this relation is not a deterministic one; that is, many schools present lower or higher performances of what is expected to their socioeconomic level, this mediated by other factors.

In this work, three socioeconomic indicators are used, available in the granularity adopted (ie, by municipalities), all made by the last time in 2010:

- *Municipal Human Development Index* (MHDI): the HDI is a concept elaborated by UNO in the 90's towards more information about social development not only by economic advances but also improvements in human well-being, considering quality of life of the population, in a broader sense than just economic, in existence at that time. The MHDI is a methodological adaptation of HDI at the municipal level, though measuring the same dimensions, is a more adequate to assess conditions in a smaller social nucleus. Currently, the MHDI considers the geometric mean of three factors: income, health and education. It ranges from 0 to 1, and the higher its value, the better its development index.

- *Gini coefficient* (*Gini index*): measure used to estimate the inequality of income and was created by the italian statistician Corrado Gini in 1912. It analyzes the correlation between poorest and wealthiest populations, rating them according to income levels. Numerically, it ranges from 0 to 1. The closer to zero the better the income distribution; the closer to one, more unequal.

- *Social vulnerability index* (SVI): this measure points to the absence or insufficiency of some assests, resources or structure (as income stream, adequate conditions of living, access to education services, among others) that should be available to every citizen, fostering favorable life conditions and social inclusion, created by the brazilian Institute of Applied Economic Research[2] (IPEA) from demographic census data, ranging from 0 to 1, with the higher the value the higher the vulnerability.

In addition, other variables were selected, by enriching the dataset with DBpedia data:

- *GDP por capita*: represents the total GDP of a region divided by the number of its inhabitants, with the idea that citizens would benefit from an increase in the aggregated production in their municipality. Similarly, GDP per capita is not a personal income measure and does not consider differences in wealth distribution.

- *Population*: this variable was selected to investigate if there is any relation of a municipality's population size and school performance, supposing that bigger municipalities have more students and schools and, thus, a more complex organizational structure, which, in turn, influences its performance.

- *Municipality geographic location*: latitude and longitude coordinates, to support georeferencing applications.

Semantic Web and linked data

The term *semantic web* refers to W3C's vision of the Web of linked data, allowing people – through semantic technologies like RDF, SPARQL, OWL, SKOS, among others – to create data repositories on the Web, build vocabularies and write rules to handle these data,

---

[1] http://portal.inep.gov.br/basica-censo

[2] http://www.ipea.gov.br/

providing an environment where application can query, inference and interpret these data. The main goal of the Web of data is "*to*

enable computers to do more useful work and to develop systems that can support trusted interactions over the network" [3].

The concept of *linked data* was created by Tim Berners Lee and refers to best practices for publishing and linking data on the Web, by establishing principles, techniques and standards for publishing, retrieving and interlinking structured data in a standard way. His vision is the creation of a big global graph, formed by RDF triples that share understanding about the data, besides the data itself [9]. The following principles should be considered:

- Use of URI to identify things/resources;
- Use of HTTP protocol to access resources;
- Provide information in open standards (RDF or SPARQL);
- Link resources to other URI to discover more information;

Lee also suggests a 5-star rating system to show the level of maturity in the provision of data [10]. In *Level 1*, the data are available on the Web, with open licence, but in unstructured format; such as a PDF file; in *Level 2*, the data are available in a structured manner, but in a proprietary format, such as an Excel file; in *Level 3*, the data are open, structured and in open format, such as CSV; in *Level 4*, in addition to previous requirements, the data uses W3C standards, such as RDF, SPARQL and URI to identify resources; in *Level 5*, the data are connected to other data sources, as linked open data, so to expand its context, as shown in Figure 1.

Figure 1. An illlustration of the 5-star rating for the definition of the maturity level in opoen data[4].

Provision of data

The brazilian Law on Access to Information [11], an initiative of federal brazilian government to open and disseminate public data, states: "*everyone has the right to receive from public agencies information of his private interest, or of collective and general interest, provided in period established in law, on liability, with exception for those which secrecy is critical to society and federal union safety*" (author's translation). Consequently, many datasets are being provisioned, including educational data.

However, a major part of educational data currently available, collected by federal government in every public school across the country, in elementary, middle and high schools are available in isolation, having each dataset limited in its narrow scope and with some files dependent on proprietary formats and difficult access for automated queries [2].

3. RELATED WORK

Soares and Alves analyzed the school and municipal effects through hierarchical linear regression, considering the performances of students all over the country in IDEB 2005 - 2011 editions [8]. The study shows strong influence both from school socioeconomic variables and from municipalities where the schools are located. The authors point to empirical evidences in several locations and periods, and suggest that school performance reflects very straightforwardly social, demographic and cultural characteristics of the students, synthesizing their previous academic experiences [7][12]. They also observe that the "quality of education cannot be analyzed disregarding school and student context". In [13] it was studied the IDEB in function of the percentage of inhabitants which live with income lower than half minimum wage and observed a strong influence in its result, particularly at municipal level. A similar research analyzed the IDEB scores in function of: percentage of families served by the *Bolsa Família*[5] program, cost per student, population and region, both at school and at municipal level and observed a significant effect for all independent variables, showing that the presence of students in poverty conditions has substantial negative effect in the school IDEB [14]. Many other studies also looked into these relations showing great influence of socioecomic variables over academic performance.

With regard to linked open data for education, Mota and Rossi [15] created a strategy for publishing information embedded in School Census and followed linked open data principles, as part of a project called Web-PIDE. The data was imported to a relational database and a tool used both to generate an OWL ontology and to load the data triples into Virtuoso[6] triplestore. In [16] was proposed an integration of multiple web applications of education indicators, taken from School Census and *Prova Brasil*, also through Web-PIDE portal and the creation of a data warehouse from the datasets. The integration was made through web services for visual inquiries and access of the indicators of large scale assessments by human users. Oliveira and Turine [17] worked with ENEM[7] microdata to create a data warehouse to support analytic queries for decision-making. They proposed a process model of linked data publishing, consisting in the triplification of ENEM data and its release for queries via SPARQL in Virtuoso triplestore. Cabral et. *al* [18] used this same data to create a tool to help the visualization of the data collected in this exam. The data was imported in a relational database and the mean performance of all students in the respective municipalities was computed. Two ontologies were created and the D2QR[8] was used to map the data to the ontologies, converting database tuples in RDF triples and using Virtuoso triplestore.

---

[3] http://www.w3.org/standards/semanticweb/

[4] http://5stardata.info/en/

[5] *Bolsa Família* is a monetary (income) transference program made to poor individuals or families in Brazil.

[6] http://virtuoso.openlinksw.com

[7] ENEM is a large-scale assessment for high school students, used as an entrance exam to public universities. Microdata available at: http://dados.gov.br/dataset/microdados-do-exame-nacional-do-ensino-medio-enem

[8] http://d2rq.org/

Non-academic portals, like *Qedu*[9], *Educação Inteligente*[10] and *Data Escola Brasil*[11] provide user-friendly applications on the data about the schools and their academic performances. *Educação Inteligente* also provides an API for access to some data in their repository. However, these portals do not support linked data, semantic information or open queries via SPARQL.

This work uses other data sources than the previous academic works, enabling the link to indicators not so explored in the linked data community. As an example of application on this semantic infrastructure, we performed a correlational study between socioeconomic indicators and IDEB of each municipality in Brazil.

4. METHOD AND MATERIAL

The socioeconomic indicators (MHDI, SVI, Gini index) were extracted from IPEA (Institute of Applied Economic Research), in CSV format, by manually selecting the parameters and indicators to the requested granularity. IDEB data with information on school performance, at municipal level of public schools, were collected in the form of zipped Excel spreadsheet[12], making it difficult to read them automatically. There are no mentions of APIs to access both those datasets automatically, to the author's knowledge. Another problem is the use of many explanative headers and trailers in the spreadsheets. Due to these complexities, and as both sources did not provide programatical access to the data, they were downloaded and enriched off-line. In order to enrich this data aggregation, we opted to add contextual data of municipalities, by linking them to their URIs in DBpedia and extending the available set of variables with: population, GDP per capita, latitude and longitude.

Many issues in the collection, preprocessing and in quality of data were faced during the data consolidation. For instance: incomplete data (since only few datasets are published as structured open resouces), decentralized data sources (some information are present in multiple agencies, withou a single official source; there is a separation of datasets per year) and lack of standard for data publishing (both in format and data available). In turn, some issues in portuguese DBpedia were also found: the data on MHDI is innacurate and outdated (thus, prevailing IPEA's data); data about longitude have incorrect sign; some municipalities had profound alterations not present in DBpedia (like changing its name or emancipation); some municipalities not correctly annotated in the ontology; the difficulty in automatically reconciling the names of municipalities and resources in DBpedia, since a single municipality may have small variations for its name. The reconciliation procedure was made partially manual, checking for accuracy, since OpenRefine could not reconcile a great part of the data.

This case study uses a non-experimental methodology, of correlational descriptive design, relational (not predictive), that aims to explore relations that might exist among the variables other than the cause-effect relation, given its natural occurence [19]. The methodological design is transversal, because the variables are identified just in a point in time, with no regard in analyzing the passing of time. Finally, is also censitary, because all municipalities (the unit of analysis chosen) in Brazil are included in this work. The dependent variables were retrieved from official databases from the government, available to general public.

For consolidation, conversion and publishing of these data, the following tools were used: OpenRefine to extract and convert the data from the data sources, reconcile with DBpedia, and generate RDF triples. Protégé was used to create a ontology and a SPARQL endpoint was set up. The full procedure is illustrated in Figure 2:
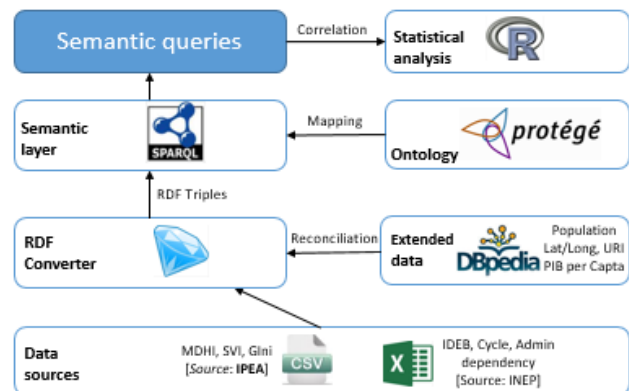
Figure 2. Provisioning of data adopted in this work.

For the statistical analysis, the R software, version 3.2.2 was used, along with *SPARQL R*[13] and *Hmisc* packages, which can, respectively, fetch and import the data from the endpoint and calculate the correlation matrix and its statistical significance. Figure 3 depicts the R script and the query used to generate the analysis data.

5. RESULTS

The results are shown in two parts, each aligned to its respective goal.

5.1 Linked open data modeling

Both school performance and socioeconomic indicators data of Brazilian municipalities were provisioned in Level 5, via endpoint SPARQL, hosted by a free hosting triplestore (Dydra)[14]. Figure 4 shows the ontology developed for this task to represent these data based on the analysis of the data collected[15]. This work also reused other ontologies for classes and properties. From it, can be observed the following entities: *Municipality* and relations: *latitude*, *longitude, code, name* datatype properties and *state*, *indicator* object properties. The concept of *Indicator* is linked to concepts like *year of reference, source, value* and *title* and its specializations – *education* (IDEB), *demographic* (Population) or *socioeconomic* (MHDI, Gini, SVI and GDP per capita). *Education indicator* is related to a *Cycle*, concept commonly used to denote if the indicator refers to different stages: primary, secondary or high school; it is also related to an *Administrative Level* like: municipal, state, federal or private. Brazilian government has also developed a controlled vocabulary for subjects, named *VCGE*[16], to classify data unregarding of the responsible agency, used in *EducationIndicator*. The URI of the municipalities and states, reconciled from Dbpedia were also mapped, enabling the extension via *owl:sameAs* property. Though they were not used in the query, the resulting dataset could be used to query for other data from those DBpedia instances.

---

[9] http://www.qedu.org.br

[10] http://educacao.dadosabertosbr.com

[11] http://www.dataescolabrasil.inep.gov.br/dataEscolaBrasil/

[12] http://portal.inep.gov.br/web/portal-ideb/planilhas-para-download

[13] http://linkedscience.org/tools/sparql-package-for-r/

[14] http://dydra.com/brunopenteado/br_municipalities_v2/sparql

[15] https://github.com/brunopenteado/OpenData/blob/master/Indicadores_v3.owl

[16] http://vocab.e.gov.br/id/governo

```
library(SPARQL)
library(Hmisc)

endpoint <- "http://dydra.com/brunopenteado/br_municipalities_v2/sparql"
options <- "output=csv"

sparql_prefix <- "PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpprop-pt: <http://pt.dbpedia.org/property/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <https://github.com/brunopenteado/OpenData/blob/master/Indicadores_v3.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>"

q <- " SELECT
?Name ?State ?MDHI ?SVI ?Gini ?Population
?PibPerCapita ?latitude ?longitude ?IDEB_Municipal_5 ?IDEB_Municipal_9 ?IDEB_State_5 ?IDEB_State_9
WHERE {
  ?mun a dbpedia-owl:Municipality .
  ?mun rdfs:label ?Name .
  ?mun dbpedia-owl:state ?state .
  ?state dbpprop-pt:sigla ?State .
  ?mun geo:lat ?latitude.
  ?mun geo:long ?longitude .
  optional { ?i vocab:indicatorOf ?mun . ?i dbpedia-owl:humanDevelopmentIndex ?MDHI . }
  optional { ?j vocab:indicatorOf ?mun . ?j dbpedia-owl:giniCoefficient ?Gini . }
  optional { ?k vocab:indicatorOf ?mun . ?k vocab:SVI ?SVI . }
  optional { ?l vocab:indicatorOf ?mun . ?l dbpedia-owl:populationTotal ?Population . }
  optional { ?m vocab:indicatorOf ?mun . ?m dbpprop-pt:pibPerCapita ?PibPerCapita . }
  optional { ?e vocab:indicatorOf ?mun . ?e vocab:refersToLevel 'Municipal' .
     ?e vocab:refersCycle '5' . ?e dbpedia-owl:value ?IDEB_Municipal_5 . }
  optional { ?f vocab:indicatorOf ?mun . ?f vocab:refersToLevel 'Municipal' .
     ?f vocab:refersCycle '9' . ?f dbpedia-owl:value ?IDEB_Municipal_9 . }
  optional { ?g vocab:indicatorOf ?mun . ?g vocab:refersToLevel 'State' .
     ?g vocab:refersCycle '5' . ?g dbpedia-owl:value ?IDEB_State_5 . }
  optional { ?h vocab:indicatorOf ?mun . ?h vocab:refersToLevel 'State' .
     ?h vocab:refersCycle '9' . ?h dbpedia-owl:value ?IDEB_State_9 . }
} "
res <- SPARQL(endpoint,paste(sparql_prefix, q),extra=options)$results

cols <- c(3,4,5,6,7,10,11,12,13)
m <- res[,cols]
matrix <- rcorr(as.matrix(m))

View(matrix$r[c(6:9),-c(6:9)])
View(matrix$P[c(6:9),-c(6:9)])
```

Figure 3. SPARQL query and R script used to fetch the municipalities' indicators and calculate the correlations.

Using the ontology model given in Figure 4 the query of Figure 3 can be made in order to fetch the municipality school performance and their socioeconomic indicators. The *OPTIONAL* statements are used here since not every municipality has all the cycles or administrative levels contemplated (for instance, a particular municipality may not have state schools for 5th grade).

Figure 4. Graphical representation of the ontology used to represent the compiled data in this work.

5.2 Correlational analysis

From the modeling of the data, it is possible to perform some sorts of analyses on the aggregated data. In this work, it was calculated the correlation matrix relating the indicators and IDEB performances. The results are shown in Table 1, using Pearson correlation among the variables, all of them continuous. It is noted a strong correlation (magnitude ranging from 0.4 and 0.7) with SVI and MHDI with IDEB for 5th and 9th grades, both for municipal and state schools, having greater influence in primary level (5th grade, both municipal and state), ie, decreasing along grades and greater for municipal than state schools, as expected – since it is municipalities responsibility this education stage. Gini coefficient showed moderate correlation with IDEB. GDP per capita also showed moderate correlation for municipal IDEB and a slight lower for state schools. The population variable didn´t show any correlation in the performance levels analyzed.

Table 1. Correlation matrix among the variables of interest*.

	SVI (2010)	Gini (2010)	MDHI (2010)	Income per capita (2009)	Population (2010)
IDEB Municipal schools 5th grade (2013)	-0,737*	-0,463*	0,709*	0,311*	0,010
IDEB State schools 5th grade (2013)	-0,584*	-0,349*	0,538*	0,273*	-0,005
IDEB Municipal schools 9th grade (2013)	-0,615*	-0,335*	0,608*	0,306*	0,034
IDEB State schools 9th grade (2013)	-0,499*	-0,337*	0,458*	0,144*	-0,039**

*: p-value < 0.001, **: p-value <0,05

6. DISCUSSION

This work aimed at exercising a case study of relevance for educational context and appropriate for the concepts of linked open data, by using public government data that serve as basis for public policies in Brazil. It was found a very technologically complicated scenario regarding the data sources and its enrichment. The variables analyzed represent a very used set of socioeconomic factors also used as basis for public policies in multiple sectors of economy. Even being from different domains, data are complementary to each other, as described in literature, since academic performance is closely related to the socioeconomic status of students and schools.

Although these data are open and licence-free, some of them still depend on proprietary software and of much effort of a person to collect and compile in a useful manner. As example of issues for automatic processing, we illustrate: some of the data are available as zip files contaning other zip files with multiple large files in it. Spreadsheets have a lot of formatting and headers, with many levels, making it difficult to read; unique codes are actually composite keys, which may appear disaggregated in different sources; the data are available at multiple location, with no guarantee of uniqueness and making it difficult to choose the 'correct' source. The yearly datasets - with different formatting - also makes the process harder to execute (yet it makes it easier for humans to read); some other indicators were also intended to be included in this analysis, but they were not available in open format, in the desired granularity or for all the municipalities. So, for more complex analysis with live, different and complimentary indicators from diverse areas, it is still an open issue.

The data were connected and the correlation of the municipal indicators with the academic performances of public schools was estimated, considering every Brazilian municipality, bringing light to the discussion of these results and pointing to further work. However, the results should be analyzed with caution, because an indicator in isolation, no matter how good it is, models just a tiny fraction of reality. Moreover, it is about correlation and not causation.

Among the selected variables, *SVI* and *MHDI* were the ones that presented greater magnitude in correlation (r > 0.5 for 5th grade). *SVI* has as components both *education*, by measuring lower school level for adults, child and young which are not enrolled in school, and *income*, such as adult's unemployment rate, informal occupation rate, child labor, among others. *MHDI* has as components: population longevity, income and education. As for education, it is composed, in turn, by educational level of the adult population and educational flow of young people. Thus, as both encompasses both income and education information, they presented a higher correlation than the others did.

The i*ncome per capita* showed more modest results (r > 0.3 for municipal schools). *Gini coefficient*, which measures income inequalities, showed moderate correlation to IDEB (r > 0.33) and with some regularity along grades and administrative levels. As IDEB is computed only for public schools, which usually have poorer students, the relation with this indicator is weaker than expected. *Population* did not show correlation with performance, contradicting the notion that a more complex school system, due to a greater number of agentes involved, would affect school performance. This may also suggest a higher impact of economic factors, since this variable is not related to income at all.

The indicators that more strongly incorporate income and education levels were the ones with greater correlation, consistent with the literature in education policies reasearch (such as [7][8][13][14]). As stated in [8]: "IDEB does not capture educational quality, but a better economic situation". In relation to administrative dependency (municipal or state school), the effects where grater for municipal schools than state schools, what is expected, since these schools are primarily the municipality's responsibility. Another point to consider is the dropout rate along the grades, to be included in future work, since it may have attenuated the magnitude of the variables along the grades. Another direction for future work is the regression analysis of the variables, so that to detect colinearity in the effects, the effects in different regions and how they behave in conjunction. Another direction is refine the analysis by school indicators, with more granularity. Technically, the use of Data Cube Vocabulary is also good choice to represent these statistical data and is needed a more automatic way to access the datasource in an on-line setting.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Isotani, S., and Bittencourt, I. 2015. Dados Abertos Conectados. Novatec: São Paulo. 176 p. http://ceweb.br/livros/dados-abertos-conectados/. Access: 2015-10-20.

[2] Bandeira, J., Ávila, T., Alcantara, W., Sobrinho, A., Bittencourt, I. I., Isotani, S. 2015. Dados abertos conectados para a Educação. Jornada de Atualização em Informática na Educação, p 1-24. http://br-ie.org/pub/index.php/pie/article/view/3551. Access: 2015-11-20.

[3] Jannuzzi, P. M. 2001. Indicadores sociais no Brasil: conceitos, fontes e aplicações. Campinas: alínea/PUC-Campinas, 2001.

[4] Neri, M. C., and Buchmann, G. 2008. The Brazilian Education Quality Index (Ideb): measurement and incentive upgrades. In: LACEA/LAMES, 2008, Rio de Janeiro.

[5] INEP. 2015. Technical note - Ideb. http://download.inep.gov.br/educacao_basica/portal_ideb/o_que_e_o_ideb/Nota_Tecnica_n1_concepcaoIDEB.pdf. Access: 2015-10-20.

[6] Brasil, Bases and Guidelines for Education Act. 1996. Lei nº 9.394/96, 12-20-1996.

[7] Coleman, J. S., Campbell, E. Q., and Hobson, C. J. 1966. Equality of educational opportunity. Washington, DC: U.S. Government Printing Office, 1966.

[8] Soares, J. F., and Alves, M. T. G. 2013. Effects of schools and municipalities in the quality of basic education. Cad. Pesqui. 2013, vol.43, n.149, pp. 492-517. ISSN 0100-1574. http://dx.doi.org/10.1590/S0100-15742013000200007

[9] Lee, T. B., Hendler, J., and Lassila, O. 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. v.1, p. 34 - 43, 2001

[10] Lee, T. B., 2006. Linked data: design issues. http://www.w3.org/DesignIssues/LinkedData.html. Access: 09-17-2015.

[11] BRASIL. 2011. Lei de Acesso à Informação. Lei nº 12.527/11, de 18 de novembro de 2011.

[12] Bourdieu, P., Passeron, J. C. 2008. A reprodução: elementos para uma teoria do sistema de ensino. Petrópolis: Vozes.

[13] Andrews, C. W., and Vries, M. S. De. 2012. Pobreza e municipalização da educação: análise dos resultados do IDEB (2005-2009). Cad. Pesqui. 2012, vol.42, n.147, pp. 826-847. ISSN 0100-1574.

[14] Duarte, N. S. 2013. O impacto da pobreza no Ideb: um estudo multinível. Rev. Bras. Estud. Pedagog. 2013, vol.94, n.237, pp. 343-363. ISSN 2176-6681. http://dx.doi.org/10.1590/S2176-66812013000200002.

[15] Mota, F. M. and Rossi, L. L. 2010. Estudo e avaliação das técnicas e das ferramentas para o projeto de data warehouse. Coxim - MS, UFMSCPCX.

[16] Savitraz, J. D. 2010. Uma abordagem de integração e exploração visual de dados educacionais na plataforma Web-PIDE. FACOM-UFMS.

[17] Oliveira, D. A. Z., and Turine, M. A. S. 2012. Uma estratégia para publicação de um Open Linked Data baseado em Data Warehouse. 2012. 121 p. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Computação, Fundação Universidade Federal de Mato Grosso do Sul.

[18] Cabral, S. P., Beduschi, N. B., Zancanaro, A., Todesco, J. L. and Gauthier, F. A. O. 2012. Aplicando Linked Data na publicação de dados do ENEM. ONTOBRAS-MOST, vol. 938 of CEUR Workshop Proceedings, page 176-181.

[19] Cohen, L., Manion, L. and Morrison, K. 2001. Research methods in education. NewYork: Routledge Falmer. 5th Edition