SUBSIDY TO PUBLIC INSPECTIONS: Identification of Municipalities with discrepant spendings in Basic Education

Renata Guanaes Machado
Auditor Federal de Finanças e Controle - Controladoria-Geral da União

SUMMARY

Public spending must be constantly monitored by government agencies. In this context, the application of technology to produce strategic information that supports actions to combat corruption and mismanagement of public resources becomes essential. With the availability of the Information System on Public Budgets in Education (SIOPE), the present work employs data mining techniques for the detection of atypical expenses in Elementary Education, performed by municipalities in 2018 - which may constitute occasional events (such as works in schools) or represent evidence of irregularities. Clustering of municipalities and anomaly detection algorithms were applied to a group of similar municipalities. The results (if the municipality is anomalous and its degree of anomaly) can be added to the planning of control actions and, also, support the adoption of appropriate measures by other instances, such as the Ministry of Education and social control councils.

Keywords: Clustering of municipalities. Anomaly detection. Public spending. Basic education.

SUMMARY

SUMMARY .................................................................................................................................... 2

1 INTRODUCTION .................................................................................................................. 5
  1.1 MOTIVATION .................................................................................................................................... 5
  1.2 PROBLEM AND JUSTIFICATION .................................................................................................... 5
  1.3 GENERAL AND SPECIFIC OBJECTIVES ......................................................................................... 6

2 METHODOLOGY USED ....................................................................................................... 6

3 BUSINESS UNDERSTANDING PHASE ................................................................................. 7
  3.1 THE SIOPE FOR MONITORING EDUCATION SPENDING ........................................................... 7
  3.2 FUNDEB EXPENSES ......................................................................................................................... 9
  3.3 THE ROLE OF THE OFFICE OF THE COMPTROLLER GENERAL ............................................... 10
  3.4 IDENTIFICATION OF PROBLEMS OR CHALLENGES .................................................................. 11
  3.5 BUSINESS AND DATA MINING OBJECTIVES ............................................................................. 12

4 DATA UNDERSTANDING AND PREPARATION PHASE ..................................................... 13
  4.1 UNDERSTANDING THE SIOPE MUNICIPAL DATA ..................................................................... 13
    4.1.1. Linked Programs ................................................................................................................ 13
    4.1.2. Expense Groups ................................................................................................................. 13
    4.1.3. Education’s Subfunctions structured in Folders and Subfolders ..................................... 14
    4.1.4. Accounting Accounts (Nature and Element of Expense) ........................................... 16
  4.2 DATA PREPARATION - SPENDING DATAFRAME ....................................................................... 16
    4.2.1. Data Collection and Cleaning .......................................................................................... 16
    4.2.2. Inclusion of columns: “Classification” and “Type of Expense” .................................... 17
    4.2.3. Inclusion of additional data: economic and demographic data ...................................... 18
    4.2.4. Filter data for context to Elementary School ............................................................... 19
    4.2.5. Spending Dataframe Summary ...................................................................................... 20
  4.3 DATA PREPARATION - DATAFRAME OF MUNICIPALITIES ....................................................... 23
    4.3.1. Data Pivot ............................................................................................................................ 23
    4.3.2. Consolidation of G/L Accounts ........................................................................................ 23
    4.3.3. Summary of the municipalities dataframe ................................................................... 24

5 MODELING PHASE ............................................................................................................ 25
  5.1 EXPLORATORY DATA ANALYSIS .................................................................................................. 25
  5.2 CLUSTERING OF SIMILAR MUNICIPALITIES ............................................................................. 26
    5.2.1. Objectives of the clustering of municipalities ................................................................... 26
    5.2.2. Deciding on the scaling of the data ................................................................................. 26
5.2.3. k-Means Clustering ........................................................................................................... 29
5.2.4. DBSCAN Clustering ........................................................................................................... 32
5.2.5. Agglomerative Clustering ................................................................................................. 33
5.2.6. Validation of the clustering algorithms ........................................................................... 35

5.3 ANOMALY DETECTION ........................................................................................................... 37
5.3.1. Delineation of outlier detection strategies .................................................................. 38
5.3.2. The Python Outlier Detection (PyOD) library ............................................................... 38
5.3.3. Choice of Cluster to be submitted to the algorithms .................................................. 40
5.3.4. Results of anomaly detection .......................................................................................... 40
5.3.5. Validation of anomaly detection models ......................................................................... 42

6 EVALUATION AND IMPLEMENTATION PHASE ................................................................47

7 CONCLUSION .....................................................................................................................49

REFERENCES ............................................................................................................................... 51
1. INTRODUCTION

1.1. MOTIVATION

Public corruption costs governments around the world many millions every year, and combating it is a challenge for the public sector and society (THEISING, 2019). For some authors, the mismanagement of public money can lead to even worse losses than acts of corruption (ANGELICO, 2012; AGUIAR, 2019). In this context, the joint use of cutting-edge technologies, intelligence methods and data analysis tools, in recent years, has proven to be a powerful and effective ally not only to tackle corruption, but also to minimize or avoid the waste of public resources.

Several control agencies have, among their main objectives, the monitoring of public spending and the production of strategic information to support decision-making in internal control. Such monitoring can occur in various contexts, such as: procurement and public bidding; government programs; social benefits; agreements and transfers; among others. Among the various tools and methods that support internal control, one can mention: analytical data processing, statistical techniques, big data, business intelligence, artificial intelligence, data mining and machine learning. All this tool framework allows the discovery of high value-added knowledge from large government databases, the enhancement of the analysis capacity and the modernization of control, whether internal or external.

At this point it becomes fundamental to define some terms. Data Science is a set of fundamental principles that guide the extraction of knowledge from data; Data Mining is the extraction of knowledge itself, through technologies that incorporate such principles (PROVOST and FAWCETT, 2016). Thus, machine learning is one such technology, in the form of induction algorithms (stating a generalized truth from the observation of some elements).

1.2. PROBLEM AND JUSTIFICATION

This work was motivated by the need to conduct a systematic analysis on the data of the Information System on Public Budgets in Education (SIOPE), under the management of the National Fund for Education Development (FNDE). In this system, states and municipalities register their revenues and spendings on public education - which are later transmitted to other control instances for the purpose of validating the information. The FNDE then generates the management indicators and the indicators of compliance with legal provisions - such as, for example, the compliance with the application of 25% of revenues in the Maintenance and Development of Education (MDE).

Nevertheless, some reports produced by control agencies, such as the Office of the Comptroller General (CGU) and the Federal Audit Court (TCU), pointed out flaws in the reports produced by the SIOPE, which did not prevent the detour of public resources dedicated to education in several municipalities. These problems are detailed in item 3.4 of this work, after the contextualization of the SIOPE, the Fund for Maintenance and Development of Basic Education and the Appreciation of Professionals in Education (FUNDEB) and the specific attributions of the CGU as a control agency.

In view of these facts - the availability of the SIOPE data and the concrete existence of deviations in the resources allocated to education - the following research problem arose: from the records in the SIOPE, how can we identify possible evidence of irregularities in the municipalities’ public spending on Basic Education?
Finally, the justification for this work is the possibility of generating knowledge of strategic value in the theme of monitoring public spending on Education, which can give a higher quality of data to the SIOPE, as well as provide complementary subsidies to the audit and inspection work performed by the control bodies.

1.3.  GENERAL AND SPECIFIC OBJECTIVES

Given the opportunity to explore the SIOPE data, the following general objective was defined: “To propose mechanisms for identifying discrepancies in public spending made by municipalities, specifically in Elementary Education in 2018, to produce a report with value-added information for the control instances”, which provides the specific objectives below:

- contextualization about the legislation on policies and investments in education;
- understanding of the SIOPE system;
- understanding about the performance of the control instances;
- reading audit and inspection reports from the control agencies, to understand the flaws in the SIOPE and the deviations in the application of federal funds in Basic Education;
- perform exploratory analyses on the SIOPE data to raise initial hypotheses about the data;
- choose the data mining techniques to be used for the identification of discrepancies in public spending.

2.  METHODOLOGY USED

The present work was performed following the reference methodology Cross-Industry Standard Process for Data Mining (CRISP-DM), commonly used for data mining projects (CHAPMAN, 2000). In this methodology, the life cycle of a data mining project, as presented below, is composed of the following phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

**Figure 1** – The CRISP-DM life cycle

The business understanding phase consists of identifying the institutional context, understanding the business problem that the organization hopes to solve, and identifying the highest priority
needs. Next, one must determine the business objectives and their success criteria; verify the available resources (personnel, data, hardware, and software); determine the data mining objectives from the business objective; and, finally, create the project plan.

The data understanding phase involves collecting, checking the quality, and exploring the data with descriptive statistics to make discoveries and detect possible correlations.

The data preparation phase performs treatments on the data to make it suitable for the application of the mining algorithms in the next phase. Such treatments include: data cleaning, replacement of missing values, selection of relevant attributes, dimensionality reduction, creation of derived attributes, integration of external data, among others.

The modeling phase comprises the selection and application of techniques to create models and find patterns or knowledge discoveries - examples of these techniques are classification, clustering, and regression. Testing procedures must be performed (such as confusion matrix and accuracy measures) to verify the quality and validity of the results.

The evaluation phase aims to analyze whether the results of the model meet the business objectives and success criteria defined earlier, and whether the model can be implemented or whether it needs to be improved through new iterations of the previous phases.

The implementation phase consists of putting the model obtained into effective production, including, when applicable, the preparation of a final report with the results achieved.

3. BUSINESS UNDERSTAND PHASE

This phase consists of surveying existing problems to then establish the business objectives, their success criteria, and the data mining objectives. Thus, a previous contextualization about the SIOPE system is necessary; about FUNDEB as the main source of federal funding for Basic Education; and about the supervisory role of the CGU over federal funds passed on to states and municipalities.

3.1. THE SIOPE FOR MONITORING EDUCATION SPENDING

The SIOPE, instituted by MEC Ministerial Ordinance no. 06, of June 20, 2006, and made operational by the FNDE, came about as a response to a demand by the then Minister of Education, Cristovam Buarque (2003), who wanted to identify how much was invested in Brazilian public education. As published on the FNDE’s website:

“The SIOPE is an electronic tool instituted for the collection, processing, dissemination, and public access to information regarding the education budgets of the Union, the states, the Federal District, and the municipalities, without prejudice to the proper attributions of the Legislative Powers and the Audit Courts.” (BRAZIL, FNDE, 2019).

This system allows federal entities to register, on a bimonthly basis, the revenues and expenses incurred with public education, including the remuneration of teaching professionals and expenses funded with resources from related federal programs, such as the National School Meals Program (PNAE) and the National School Transportation Support Program (PNATE). Although the data are of a declaratory nature, there are several integrity rules, built into the system, that check the data posted before transmission to the FUNDEB’s Council for Monitoring and Social Control (CACS)
and to the FNDE. For example, the system informs you if there is an omission of any tax revenue that is entered in the STN systems; or when there are students in Early Childhood Education, declared in the School Census (INEP), without the registration of expenses in this same type of education.

After data validation, the system indicates if each federative entity reaches the legally established percentages - bearing in mind that article 212 of the Federal Constitution (BRASIL, 1988) decrees that states and municipalities must invest a minimum of 25% of tax and transfer revenues in MDE. Furthermore, Law #9,394/96 (BRASIL, 1996) - the Law of Directives and Bases for National Education (LDB) - specifies, in its article 70, the eligible expenses with MDE resources:

Art. 70 - Expenses [...] will be considered as education maintenance and development:
I - remuneration and improvement of the teaching staff and other education professionals;
II - acquisition, maintenance, construction and conservation of facilities and equipment necessary for teaching;
III - use and maintenance of goods and services related to education;
IV - statistical surveys, studies, and research aimed primarily at improving the quality and expansion of education;
V - performing the necessary intermediate activities for the operation of the educational systems;
VI - granting of scholarships to public and private school students; [...] 
VIII - acquisition of didactic and school material and maintenance of school transportation programs. (BRAZIL, 1996)

There are penalties imposed on units that do not use the SIOPE. The insertion, updating, and transmission of data to the CACS and FNDE are mandatory and have defined deadlines, under penalty of blocking the transfer of financial resources (voluntary transfers and agreements with the Federal Government) and the irregular situation in the Auxiliary Information Service for Voluntary Transfers (CAUC).

Finally, the system allows the periodic generation of educational management indicators, supporting the definition, implementation, and monitoring of educational public policies by several actors, as shown in the figure below.

**Figure 2 – SIOPE’s Partner Network**

In any case, the SIOPE is an important tool for monitoring and evaluating public spending on education - both by the Ministry of Education/FNDE and control agencies, and by educational managers and CACS. Moreover, the information recorded, together with indicators and consolidated reports, are made available by the FNDE on a website for citizens to access, ensuring transparency and social control of public resources for education.

3.2. FUNDEB EXPENSES

In the SIOPE there is a record of an important source of resources: it is the FUNDEB. This fund was created in 2006 by Constitutional Amendment No. 53/2006, replacing the Fund for Maintenance and Development of Basic Education and the Appreciation of Teaching (FUNDEF), and is regulated by Law #11,494/2007 (BRAZIL, 2007). All FUNDEB resources must be invested exclusively in Basic Education (particularly, in the valuing of teachers) - this amount reached, according to the figure below, the value of R$ 150 billion in 2018. It ends in 2020, but a bill has been created to make it permanent.

Figure 3 – FUNDEB Investments Evolution

The FUNDEB is a special redistributive fund, accounting in nature and statewide. As shown in the figure below, it is formed by 20% of the resources coming from taxes and transfers from the states and municipalities, and by a financial portion of federal resources (as a complement). Subsequently, the amounts are recalculated (based on the cost per student, calculated by the FNDE, and the number of students enrolled, surveyed by INEP) and redistributed to the entities, proportionally and equally.

Figure 4 – FUNDEB composition and redistribution

Source: JEDUCA (2019).

The FUNDEB is a special redistributive fund, accounting in nature and statewide. As shown in the figure below, it is formed by 20% of the resources coming from taxes and transfers from the states and municipalities, and by a financial portion of federal resources (as a complement). Subsequently, the amounts are recalculated (based on the cost per student, calculated by the FNDE, and the number of students enrolled, surveyed by INEP) and redistributed to the entities, proportionally and equally.
This redistribution seeks to guarantee the minimum national value per student/year to each federative entity, contributing to reduce resource inequalities among the education networks and universalizing the supply of education to all citizens.

### 3.3. THE ROLE OF THE OFFICE OF THE COMPTROLLER GENERAL

At the federal level, the CGU is the central body of the Internal Control System (SCI), the Correction System, and the Ombudsman System of the Executive Branch (BRASIL, CGU, 2020), and it is responsible for the defense of the public patrimony, internal control, public auditing, correction, and ombudsman functions, in addition to actions to promote transparency; to encourage public integrity; and to prevent and combat corruption.

The Internal Regulations of the CGU (BRASIL, CGU, 2020b) provide for its organizational structure, and the Federal Secretariat for Internal Control (SFC) is responsible for exercising the competencies of the central body of the Federal Executive Branch’s ICS:

- **XIV** - to inspect and evaluate the execution of government programs [...];
- **XVI** - conduct audits on the management of federal public resources [...];
- **XVIII** - investigate illegal or irregular acts or facts practiced by public or private agents in the use of federal public resources [...].

Among the activities performed in internal control is the Evaluation of Government Programs (AEPG) and the Inspection of Federative Entities (FEF), which verify the regularity of the application of federal public resources transferred to states and municipalities. These instruments undergo constant improvements, such as the creation of the Vulnerability Matrix in 2015, a risk analysis tool composed of twelve indicators - this generates a ranking of municipalities with greater weaknesses in the application of resources, and from which those to be inspected are selected.

With respect to the federal funds from FUNDEB for the municipalities, the CGU conducts field inspection activities in the municipalities chosen by the Vulnerability Matrix or by public drawings. These actions result in Evaluation Reports, with detailed analyses of the operational and financial
execution of FUNDEB - such as, for example, the adequacy of the payroll of the teaching staff or the correct execution of contracting processes for the supply of goods and services to schools.

Some reports - in particular, those resulting from the FEF in two municipalities in the state of Piauí, produced in 2016 (BRASIL, CGU, 2019a and 2019b); and in one municipality in the state of Maranhão, produced in 2019 (BRASIL, CGU, 2020a) - pointed to irregularities in the application of FUNDEB resources, passed on to municipal governments, among them:

- Realization of ineligible expenses with FUNDEB resources, or expenses incompatible with the FUNDEB’s objective;
- Payment of professionals without proven performance in teaching;
- Spending without supporting documents;
- Payments made for services not received;
- Overbilling by quantity in the acquisition of goods, furniture, equipment, and permanent materials with FUNDEB resources;
- Overbilling by quantity in the execution of construction, renovation and expansion works in schools with FUNDEB funds.

It can be stated that methodologies and instruments for internal control assist in the process of selecting municipalities and enable, when in the field, the detection of existing irregularities. Nevertheless, with the availability of revenue and spending information in the SIOPE (in the period from 2005 to 2019) and with the concomitant use of database crossing, statistical techniques and data mining algorithms, it is possible to add value to existing methods - such as the Vulnerability Matrix - in the sense of indicating perspectives not yet explored.

3.4. IDENTIFICATION OF PROBLEMS OR CHALLENGES

Article 212 of the Federal Constitution (BRASIL, 1988) and the LDB (BRASIL, 1996) foster the decentralization of Basic Education and the municipalization of basic education, granting autonomy to states and municipalities to organize and maintain the official institutions of their education systems. However, this administrative decentralization, in a country of great territorial extension, generates an educational system of colossal proportions - more than 180,000 basic education schools and 48 million enrollments, according to the 2019 School Census (BRASIL, INEP, 2020). Moreover, the substantial volume of financial resources (in billions of reais) invested by the Federal Government results in a funding model that is difficult to monitor and, consequently, in a greater complexity of internal control mechanisms.

In addition to the CGU as an internal control body, there are also the activities of the FNDE, which monitors the applications of FUNDEB resources in states and municipalities through the SIOPE (BRASIL, MEC, 2018); and the activities of the CACS, which are charged with performing the monitoring and social control over the distribution, planning, and application of FUNDEB resources (BRASIL, CGU, 2019d), as well as analyzing the accounts reported in the SIOPE by the federative entities. It should be clarified that “monitor” does not have the same meaning as “supervise”, according to the note issued by the MEC:

“The inspection and control regarding the application of Fundeb resources [...] are the responsibility of
the local audit courts and the Public Prosecution Service of the states, safeguarding the competence of the Federal Public Prosecution Service, for states that receive the federal contribution of resources. The MEC, through the FNDE, is responsible for monitoring the application, which is done through the SIOPE [...]. However, Siope is a monitoring system whose basis is declaratory. It means that the inspection and control are only exercised directly for the purposes of auditing, inspection and eventual punishment, by the local audit courts and the Public Prosecutor’s Office. (ARCOVERDE and TOLEDO, 2019)

However, the Annual Audit Report (AAC), on the accountability of the FNDE as an audited unit (BRASIL, CGU, 2019c), pointed to the following findings on the SIOPE as a control system for the application of FUNDEB resources:

a. the actions of the various additional control instances (MEC, FNDE, CACS) are not sufficient to prevent losses, deviations and/or frauds in the applications of FUNDEB resources, detected and documented in several audit reports prepared by CGU (BRASIL, CGU, 2019a, 2019b, 2020a);

b. the presentation of the reports generated by the SIOPE on its website: “do not allow to understand the situation of education in the entities, requiring specialized work for data treatment, understanding of the information and creation of parameters that give meaning to the numbers. It is the comparison of spending and results among similar federative entities that allows to qualify the performance of the control” (BRASIL, CGU, 2019c).

c. The CACS do not have the appropriate management information to act as a control body.

In a survey conducted by CGU/CGEBC on the Court of Accounts’ Judgments, it was found that the TCU, in Judgment No. 618/2014 - Plenary (BRASIL, TCU, 2020), evaluated the SIOPE as to the reliability of the information on MDE spending and listed the following findings, among others:

a. impossibility to certify that the information provided by the federal entities in the SIOPE reflects the spendings in MDE, and that the personnel spending is reliable;

b. validation of information before data transmission only takes place on revenue items - expense items do not have the same level of control.

Given the facts mentioned above, the main problems identified are: difficult monitoring of investments in education; high complexity of internal control mechanisms; inadequate SIOPE management reports for use by the CACS; and lack of prior validation of information in the SIOPE for expense items.

3.5. BUSINESS AND DATA MINING OBJECTIVES

Given the context so far, priority was given to the problem: “SIOPE management reports not suitable for use by CACS” and the following business objective was formulated: conduct exploratory analysis and data mining techniques on expenses linked to Basic Education, to produce value-added information for control bodies. Therefore, this information may provide additional subsidies to the Vulnerability Matrix for the work of the CGU, and may also improve the monitoring of the application of resources by the FNDE and the CACS.

In this work, the initial scope was established to indicate the municipalities with discrepancies in their spending on Primary Education in 2018 - that is, with inconsistent expenses compared to a certain standard, which may be evidence of failures or irregularities in public spending. As a success criterion, it is expected to obtain at least 1% of federal entities with discrepancies in their declared
educational expenses, and that some entities are identified as anomalous in all detection algorithms to be used.

Regarding data mining, the following objectives were set:

a. Perform exploratory analysis on the data to generate new knowledge;
b. Use at least three clustering techniques to create groups of similar municipalities;
c. In a cluster of similar municipalities, use at least five algorithms to detect anomalies in the spendings of these municipalities.

4. DATA UNDERSTANDING AND PREPARATION PHASE

These phases involve data collection, exploration, and treatment; and they occurred simultaneously throughout the execution of this work. At the end of the chapter, we have the following dataframes: spending dataframe (tabular data) for use by exploratory analyses, in which each record represents data on one spending; and municipalities dataframe for use by clustering techniques, in which each record represents a municipality with its vector of characteristics. Both dataframes contain only municipal spendings.

4.1. UNDERSTANDING THE SIOPE MUNICIPAL DATA

The existing components of the SIOPE are described for familiarization with the application domain: spending groups, programs, education sub-functions, and accounting accounts.

4.1.1. Linked Programs

These are FNDE resources passed on to the federative entities and, of course, this field is only filled in for the records of the Bound Spending group. In the Municipal SIOPE of 208 for Primary Education, the following programs are registered:

- PNAE, PNATE, PDDE;
- Linked to the Social Contribution on Education Allowance;
- Other Transfers of FNDE Resources;
- Agreement Transfers - Education;
- Other Education Resources;
- FUNDEF Lawsuit – Judiciary Bonds.

4.1.2. Expense Groups

Expenses are classified, described in the table below.

<table>
<thead>
<tr>
<th>Table 1 - Description of each Expense Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Own expenses defrayed with taxes and transfers</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
4.1.3. Education’s Subfunctions structured in Folders and Subfolders

The inclusion of expenses in the SIOPE meets the Brazilian budget model, using the functional and programmatic classification. According to the Ordinance # 42, April 14, 1999, of the Ministry of Planning, the function represents the highest level of aggregation of the various areas of spending that fall under the public sector (BRASIL, MPDG, 1999). Therefore, it refers to the government’s area of action (education, health, social security, etc.). Since it deals with the education function, the table below lists the education sub-functions registered in the system.

**Table 2 - Education Function Subfunctions in the Municipal SIOPE.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative Support Subfunction infrastructure</td>
<td>121 - Planning and Budget</td>
<td>122 - General Administration</td>
<td>123 - Financial Management</td>
<td>125 - Standardization and Inspection</td>
<td>126 - Information Technology</td>
<td>128 - Human Resources Training</td>
<td>131 - Social Communication</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020).
These are also considered administrative support subfunctions and are located as subfolders under some teaching modality (Parent folder numbered 361 to 365).


In allocating a given expense, after choosing the expense group, one or more sub-functions that represent it are selected, as shown in Figure 5. Since this is a hierarchy of sub-functions, it was agreed to use the fields Pasta_Pai (Parent Folder) and Pasta - the Pasta_Pai is a typical sub-function (the education modality) or the item “Own Expenses Financed with Taxes and Transfers” - referring to sub-functions 242, 243, etc. which are school expenses that are independent of an education modality.

The Portfolio is just another sub-function, subordinated to the Pasta_Pai.

In the following figure, below the typical sub-functions numbered 361 to 365 (educational modalities) are unfolded the administrative support sub-functions. These are medium activities that favor the development of school activities and influence, indirectly, the execution of typical sub-functions - naturally, they are prorated expenses by the number of enrollments in each education modality, subsidizing the calculations of the cost per student and facilitating the calculation, at the national level, of how much is spent in each of these areas (BRASIL, MEC, 2018).

---

**Figure 5 - Hierarchy of Subfunctions used to classify an expense**

<table>
<thead>
<tr>
<th>Subfunction Others</th>
<th>others considered in the MDE calculation</th>
<th>Apenas Pasta</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>331 - Worker Protection and Benefits (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>722 - Telecommunications (Distance Education)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>782 - School Transportation (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>841 - Domestic Debt Refinancing (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>842 - External Debt Refinancing (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>843 - Internal Debt Service (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>844 - External Debt Service (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>846 - Other Special Charges (*)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subfunction Others</th>
<th>others NOT considered in the MDE calculation</th>
<th>Apenas Pasta</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>242 - Assistance to the Handicapped</td>
<td></td>
</tr>
<tr>
<td></td>
<td>243 - Child and Adolescent Care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>271 - Basic Welfare</td>
<td></td>
</tr>
<tr>
<td></td>
<td>272 - Provident Fund Scheme</td>
<td></td>
</tr>
<tr>
<td></td>
<td>273 - Pension Plan</td>
<td></td>
</tr>
<tr>
<td></td>
<td>274 - Special Welfare</td>
<td></td>
</tr>
<tr>
<td></td>
<td>306 - Food and Nutrition - School Feeding (*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>392 - Cultural Dissemination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>695 - Tourism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>812 - Community Sports</td>
<td></td>
</tr>
<tr>
<td></td>
<td>813 - Leisure</td>
<td></td>
</tr>
</tbody>
</table>

(*) These are also considered administrative support subfunctions and are located as subfolders under some teaching modality (Parent folder numbered 361 to 365).
4.1.4. Accounting Accounts (Nature and Element of Expense)

The Chart of Accounts in the SIOPE follows the determinations of the Budget Technical Manual (MTO), the basis for the preparation of the Union's Fiscal and Social Security Budgets (BRASIL, MPDG, 1999). According to Figure 6, after selecting the spending group and sub-functions (on the left), the right side shows the hierarchical accounting accounts in analytical (in white) and synthetic (in blue, disabled for completion). You finally choose the account most related to the expense being recorded. The yellow color is the account that is currently selected by the user in the application.

Figure 6 - Using G/L Account to Classify an Expense

4.2. DATA PREPARATION - SPEND DATAFRAME

4.2.1. Data Collection and Cleaning

The scope of the selection of data from the Municipal SIOPE was restricted to the spendings of municipalities in the year 2018, with the most relevant attributes as listed in the table below. Cleaning procedures are summarized in Table 4.
Table 3 - SIOPE data model fields relevant to the work

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CodUF</td>
<td>State Code</td>
</tr>
<tr>
<td>NameUF</td>
<td>State Name</td>
</tr>
<tr>
<td>SigUF</td>
<td>State Acronym</td>
</tr>
<tr>
<td>CodMunicipio</td>
<td>Municipality code</td>
</tr>
<tr>
<td>NameMunicipality</td>
<td>Name of municipality</td>
</tr>
<tr>
<td>Classif_Pasta</td>
<td>Equivalent to the Expense Group (Own, FUNDEB, Linked).</td>
</tr>
<tr>
<td>NameProgram</td>
<td>FNDE’s Program Name (only for linked spendings).</td>
</tr>
<tr>
<td>CodPasta_Pai</td>
<td>Parent Folder Code (represents the teaching modality).</td>
</tr>
<tr>
<td>NameFolder_Father</td>
<td>Parent Folder Name (represents the teaching modality).</td>
</tr>
<tr>
<td>FolderCode</td>
<td>Folder Code (education subfunction).</td>
</tr>
<tr>
<td>FolderName</td>
<td>Name of the Portfolio (education sub-function).</td>
</tr>
<tr>
<td>CodCC</td>
<td>G/L account number, without periods.</td>
</tr>
<tr>
<td>Cod_CC_f</td>
<td>G/L account number with dots separating two-digit groups.</td>
</tr>
<tr>
<td>NameCC</td>
<td>G/L account name.</td>
</tr>
<tr>
<td>Cod_NameCC_f</td>
<td>Field that joins the Name to the G/L Account Code - required because there are different G/L accounts that have the same name.</td>
</tr>
<tr>
<td>Updated Allocation (DA)</td>
<td>Budgeted appropriation (plus supplements, minus recorded cancellations).</td>
</tr>
<tr>
<td>Commitment Expenses (DE)</td>
<td>Expense originating from an act issued by a competent authority that creates an obligation of payment for the State.</td>
</tr>
<tr>
<td>Liquidated Expenses (DL)</td>
<td>Verification of the right acquired by the creditor, based on documents proving the delivery of the material or the provision of the service.</td>
</tr>
<tr>
<td>Expenses Paid (DP)</td>
<td>It consists in the discharge of the acquired good or the contracted service.</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020).

These data were accessed in the Jupyter Notebook environment (PROJECT JUPYTER, 2019) and consolidated into different dataframes by employing Python language (PYTHON, 2001) and a diversity of libraries, including: Pandas for data analysis (THE PANDAS PROJECT, 2019); PYODBC for database connection; and SEABORN (SEABORN, 2019) for graph generation. To use the data mining (clustering) algorithms, the modules of the open-source machine learning library SCIKIT-LEARN (PEDREGOSA, 2011) were used.

Table 4 - Data cleaning procedures

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA, DE, DL, DP</td>
<td>Null fields have been filled in with ZERO value (this is due to some spending types not filled in by the federal entities - a consequence of the pivoting in the “Spending Type” column).</td>
</tr>
<tr>
<td>Program Name</td>
<td>This field applies only to Bound Expenses. Other records (own expenses and FUNDEB) with a null value were updated to “Not applicable”.</td>
</tr>
<tr>
<td>CodPasta Pai</td>
<td>Null fields were updated to “Not applicable”, because they refer to sub-functions that do not fit any education modality. Although null, the NomePasta_pai field is filled in with the value “Own Expenses defrayed with Taxes and Transfers”.</td>
</tr>
</tbody>
</table>

Fonte: Elaborada pelo autor (2020).

4.2.2. Inclusion of columns: “Classification” and “Type of Expense

The data collection lists all synthetic (grouped) and analytical (into which values are entered) G/L accounts. Thus, it became necessary to indicate the classification of accounts into synthetic and analytical, so that the expense dataframe would contain only the analytical ones that would be subject...
to data analysis.

**Figure 7** – Quantitative records of analytical and synthetic accounts

<table>
<thead>
<tr>
<th>Lista de Grupos de Despesa:</th>
<th>3 valores</th>
</tr>
</thead>
<tbody>
<tr>
<td>['Desp próprias - não EF', 'Desp FUNDEF', 'Desp com Recursos Vinculados']</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contagem de contas analíticas (unique):</th>
<th>476</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contagem de contas sintéticas (unique):</td>
<td>40</td>
</tr>
</tbody>
</table>

Considering all groups of Despesas dos Municípios:
- Total of registros de despesas dos Municípios: 2026497 despesas
- Total of registros de despesas Analíticas dos Municípios: 923778 despesas
- Total of registros de despesas Sintéticas dos Municípios: 1102919 despesas

Source: Prepared by the author (2020).

Moreover, due to many accounting accounts (more than 300 accounts), each analytical accounting account was manually classified into a more generic spending category, to consolidate the accounts into 9 major groups of type of spending: Remuneration, Training, Didactic, Feeding, Transport, Maintenance, Investments, Partnership and Other. This classification was performed by the CGU team, and validated by a SIOPE manager at FNDE. The figure below shows the analytical spending count in each Spending Type group.

**Figure 8** – Analytical Spend Count in each Type of Spending

<table>
<thead>
<tr>
<th>Index</th>
<th>Total registros</th>
<th>Total de registros em %</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Manutenção</td>
<td>409755</td>
<td>44.36</td>
</tr>
<tr>
<td>1. Remuneração</td>
<td>284333</td>
<td>28.51</td>
</tr>
<tr>
<td>7. Investimentos</td>
<td>99349</td>
<td>10.75</td>
</tr>
<tr>
<td>4. Alimentação</td>
<td>47067</td>
<td>5.10</td>
</tr>
<tr>
<td>5. Transporte</td>
<td>30084</td>
<td>4.23</td>
</tr>
<tr>
<td>3. Didático</td>
<td>33816</td>
<td>3.66</td>
</tr>
<tr>
<td>9. Outros</td>
<td>18912</td>
<td>2.05</td>
</tr>
<tr>
<td>8. Conveniadas</td>
<td>6763</td>
<td>0.73</td>
</tr>
<tr>
<td>2. Formação</td>
<td>4699</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020).

**4.2.3. Inclusion of additional data: economic and demographic data**

The initial attempts at modeling expenses did not produce good results, as it is not recommended to compare all expenses without due attention to the particularities of each municipality (larger municipalities, with more teachers and students, spend much more, and vice versa). In view of this, there was a return to the data preparation phase for the inclusion of external data to characterize the municipalities for clustering purposes. The table below summarizes the external data added to the spending dataframe.

**Table 5** – Data from external sources added to the study of municipal spendings

| IBGE Data | IBGE municipality code; region, mesoregion and microregion; population in the 5,570 Brazilian municipalities (BRASIL, IBGE, 2019). |
The enrollment figures, by education modality and for each municipality, are data provided by INEP and registered in the SIOPE. Regarding the numbers of schools and teachers, the selection conditions were obtained from INEP’s “Basic Education Filters” document, which deals with instructions for the use of the Microdata of the Basic Education Census - 2018 (BRASIL, INEP, 2019e), and executed in INEP’s database.

The Basic Education Development Index (IDEB) (BRASIL, INEP, 2019a), prepared by the MEC, is a quality indicator for primary and secondary education, covering the public and private networks, and is the result of crossing performance (Brazil Test and National Basic Education Assessment - ANEB) with school performance (approval). The following indicators have been added to the spending dataframe:

- IDEB Early Years (IDEB_AI) and IDEB Final Years (IDEB_AF): refer to each municipality’s IDEB - Ensino Fundamental (EF) - Early and Final Years (2017 reference) score (only the score of urban schools in the municipal network); and
- IDEB High School (IDEB_EM): refers to the IDEB - High School grades (2017 reference) of each municipality (only the grades of state schools).

The dropout rate represents the proportion of students who, in 2014 were enrolled in grade k (serialized stage of primary or secondary education), and in 2015 were not enrolled. At the time this data was sought, the most recent rate available was for the year 2014 for the year 2015.

Regarding the Municipal Human Development Index (MDI), generated by the United Nations Development Program (UNDP)-it is a composite measure of indicators from three dimensions of human development: longevity, education, and income, to assess the development of Brazilian municipalities (BRASIL UNDP, 2019).

- Education HDI (HDI_E): geometric mean of the subindex of children and young people's school attendance, with a weight of 2/3, and the subindex of adults' schooling, with a weight of 1/3.
- HDI Longevity (HDI_L): obtained from the Life Expectancy at Birth indicator (minimum and maximum values are 25 and 85 years, respectively).
- HDI Income (HDI_R): obtained from the Income per Capita indicator (minimum and maximum values are 8.00 and 4,033.00).

### 4.2.4. Filter data for context to Elementary School

According to the scope defined for the present work, only the Elementary School records were kept in the expense dataframe.
Figure 9 – Number of analytical spending records in each education modality

<table>
<thead>
<tr>
<th>GrupoDespesa</th>
<th>CodPasta_Pai</th>
<th>NomePasta_Pai</th>
<th>Total Registros</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desp FUNDEF</td>
<td>361</td>
<td>Ensino Fundamental</td>
<td>131131</td>
</tr>
<tr>
<td>Desp com Recursos Vinculados</td>
<td>301</td>
<td>Ensino Fundamental</td>
<td>171223</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>361</td>
<td>Ensino Fundamental - Exceto FUNDEF</td>
<td>273225</td>
</tr>
<tr>
<td>Desp com Recursos Vinculados</td>
<td>362</td>
<td>Ensino Médio</td>
<td>7214</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>362</td>
<td>Ensino Médio</td>
<td>5155</td>
</tr>
<tr>
<td>Desp com Recursos Vinculados</td>
<td>363</td>
<td>Ensino Profissional</td>
<td>1668</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>363</td>
<td>Ensino Profissional (Qualificação para o Trabalho)</td>
<td>3276</td>
</tr>
<tr>
<td>Desp com Recursos Vinculados</td>
<td>364</td>
<td>Ensino Superior</td>
<td>3863</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>364</td>
<td>Ensino Superior</td>
<td>5705</td>
</tr>
<tr>
<td>Desp FUNDEF</td>
<td>365</td>
<td>Educação Infantil (Creche)</td>
<td>39964</td>
</tr>
<tr>
<td>Desp FUNDEF</td>
<td>365</td>
<td>Educação Infantil (Pré-Escola)</td>
<td>45982</td>
</tr>
<tr>
<td>Desp com Recursos Vinculados</td>
<td>365</td>
<td>Educação Infantil (Creche)</td>
<td>39150</td>
</tr>
<tr>
<td>Desp com Recursos Vinculados</td>
<td>365</td>
<td>Educação Infantil (Pré-Escola)</td>
<td>38131</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>365</td>
<td>Educação Infantil (Creche) - Exceto FUNDEF</td>
<td>74684</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>365</td>
<td>Educação Infantil (Pré-Escola) - Exceto FUNDEF</td>
<td>74690</td>
</tr>
<tr>
<td>Desp próprias - não EF</td>
<td>Não se aplica</td>
<td>Despesas Próprias Custeadas com Impostos e Transferências</td>
<td>9918</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020).

It is understood that each education modality uses different sub-functions and accounting accounts. It is likely that High School, for example, will have much more expenses with laboratory infrastructure (Physics, Chemistry and Biology) than Elementary School and Kindergarten (which will not have this type of expense). Consequently, the studies must be performed separately by education modality - and the present work was limited to the studies of the expenses applied to Primary Education (CodPasta_Pai field records equal to the value 361) - remembering that the modalities Youth and Adult Education and Special Education are included in the context of Primary Education.

4.2.5. Spending Dataframe Summary

After performing all the treatments, one has a dataframe of executed municipal spendings for Elementary Education in the year 2018 - its main features are summarized in the figure below. Some comments are necessary:

- There is a total of 4,989 municipalities, because not all municipalities had submitted their statement of accounts at the time the SIOPE system was received by the CGU;
- Some fields initially inserted into the dataframe (“Education Cost per Pupil” and “Education Spending on Teacher, per Pupil”) were discarded as they presented inconsistencies that would hinder the effectiveness of the data mining algorithms;
- The existing zero values for the variables “Municipal HDI”, “Number of enrollments in Primary Education”, “IDEB - Early Years”, “IDEB - Final Years”, and “Education Cost per Student” mean that there are no data (no values were registered for these variables for certain municipalities);
- The existing zero values for the Evasion Rate variable mean that there really was no evasion of students.
Figure 10 – Summary of attributes of the municipal expenses dataframe

Total de registros de despesas dos Municípios: 576279 despesas
Total de Municípios: 4989 Municípios

Total de colunas do dataframe: 50 colunas

Colunas do dataframe de despesas:

Regiões cujos Municípios entregaram a declaração ao SIOPE: 5 regiões

Estados cujos Municípios entregaram a declaração ao SIOPE: 4089 Municípios em 26 estados.

Lista de Programas: 9 valores

Lista de Grupos de Despesa: 3 valores
[‘Desp. próprias – não EF’, ‘Desp FUNDEB’, ‘Desp com Recursos Vinculados’]

Lista de Pastas Pai: 2 valores
[‘Ensino Fundamental’, ‘Ensino Fundamental - Exceto FUNDEB’]

Lista de Pastas: 21 valores

Lista de valores de Tipo de Gasto da despesa: 9 valores

Variação da população dos Municípios: 781 a 12.252023 milhões
Variação do índice IDHM dos Municípios: 0.0 a 0.862
Variação da quantidade de escolas nos municípios: 1 a 1530
Variação da quantidade de professores nos municípios: 4 a 38406
Variação do número de matrículas no EF: 0 a 458634
Variação da taxa de evasão (EF) nos municípios: 0.0 a 21.5
Variação do IDEB - EF Anos Iniciais: 0.0 a 9.1
Variação do IDEB - EF Anos Finais: 0.0 a 7.2
Custo da educação por Aluno: 0.0 a 224507.0
Despesa da educação com Professor, por Aluno: 1401.95 a 186027.96

Source: Prepared by the author (2020).

Table 6 – Description of the fields present in the expense dataframe

<p>| UF-Code, UF-Name, UF-Symbol | UF Information. | CodMun, CodIBGE, CodIBGE_Complete, NameMunicipality | County Information. |</p>
<table>
<thead>
<tr>
<th>Region, MesoRegion, NameMesoRegion, MesoRegion, MicroRegion, NameMicroRegion</th>
<th>Region Information.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NameProgram</td>
<td>FNDE’s Program Name (only for linked spendings).</td>
</tr>
<tr>
<td>ExpenseGroup</td>
<td>Own, FUNDEB or Linked.</td>
</tr>
<tr>
<td>Type of Expense</td>
<td>Classification of the G/L account in a more generic type of expense.</td>
</tr>
<tr>
<td>ParentFolder_Code and ParentFolderName</td>
<td>Represents the teaching modality.</td>
</tr>
<tr>
<td>FolderCode and FolderName</td>
<td>It represents the education sub-function.</td>
</tr>
<tr>
<td>CodCC</td>
<td>G/L account number, without periods. Ex. 34490523400.</td>
</tr>
<tr>
<td>CodCC_f</td>
<td>G/L account number with dots separating two-digit groups. Ex. 3.44.90.52.34.00</td>
</tr>
<tr>
<td>NameCC</td>
<td>G/L account name. Ex. Machines, Utensils and Equipment. Equip.</td>
</tr>
<tr>
<td>Cod_NameCC</td>
<td>Field that joins the Name to the G/L Account Code - required because there are different G/L accounts that have the same name.</td>
</tr>
<tr>
<td>Cod_NameCC_f</td>
<td>Field that appends the Name to the formatted G/L Account Code.</td>
</tr>
<tr>
<td>DA, DE, DL, DP</td>
<td>Expense Type (Updated Appropriation, Expense Paid, Expense Liquidated, and Expense Paid).</td>
</tr>
<tr>
<td>Vlr_FUNDEB_STN</td>
<td>Field added to the dataframe by request of CGEBC (outside the scope of this work).</td>
</tr>
<tr>
<td>Pop_estimated</td>
<td>Estimated population of the municipality (source: IBGE).</td>
</tr>
<tr>
<td>HDI (*)</td>
<td>HDI (Source: UNDP).</td>
</tr>
<tr>
<td>HDMI_E (*)</td>
<td>HDI Education (Source: UNDP).</td>
</tr>
<tr>
<td>HDI_L (*)</td>
<td>HDI Longevity (Source: UNDP).</td>
</tr>
<tr>
<td>HDHM_R (*)</td>
<td>HDI Income (Source: UNDP).</td>
</tr>
<tr>
<td>QtySchools</td>
<td>Number of schools (Source: INEP).</td>
</tr>
<tr>
<td>QtyTeachers</td>
<td>Number of teachers (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_361 (*)</td>
<td>Number of students enrolled in Primary School (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_362</td>
<td>Number of students enrolled in high school (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_363</td>
<td>Number of students enrolled in Professional Education (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_365</td>
<td>Number of students enrolled in Higher Education (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_365_1</td>
<td>Number of students enrolled in Early Childhood Education (daycare) (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_365_2</td>
<td>Number of students enrolled in Early Childhood Education (pre-school) (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_366</td>
<td>Number of students enrolled in Youth and Adult Education (Source: INEP).</td>
</tr>
<tr>
<td>NUM_MATR_367</td>
<td>Number of students enrolled in Special Education (Source: INEP).</td>
</tr>
<tr>
<td>IDEB_AI (*)</td>
<td>IDEB score in Elementary School - Early Years (Source: INEP).</td>
</tr>
<tr>
<td>IDEB_AF (*)</td>
<td>IDEB score in Elementary School - Final Years (Source: INEP).</td>
</tr>
<tr>
<td>IDEB_EM (*)</td>
<td>IDEB score in High School (Source: INEP).</td>
</tr>
<tr>
<td>TxEvasao_EF</td>
<td>Elementary school dropout rate (Source: INEP). Zero values mean that there was no student dropout.</td>
</tr>
<tr>
<td>StudentCost (*)</td>
<td>Value of the Cost per Student (calculation made by FNDE). It will not be used in the present work.</td>
</tr>
<tr>
<td>ExpenseProf</td>
<td>Amount of Expenses per Teacher (calculation made by FNDE). It will not be used in the present work.</td>
</tr>
</tbody>
</table>

(*) Zero values mean missing values, not reported.

Source: Prepared by the author (2020).
4.3. DATA PREPARATION - DATAFRAME OF MUNICIPALITIES

The execution of certain data mining algorithms presupposes an input data format in which each object is represented by an attribute vector (or feature vector). Because of this, further preparations were performed on the data.

4.3.1. Data Pivot

To generate feature vectors, a pivoting was performed on the spending dataframe - the result is a new dataframe of municipalities, in which each row is a municipality and the columns are all its attributes.

Table 7 – Creating new columns after pivoting the data

<table>
<thead>
<tr>
<th>Expense Group</th>
<th>Own Expenses, FUNDEB Expenses, Bound Expenses.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Expense</td>
<td>tgRemun (Remuneration), tgFormacao (Training), tgDidatico (Didactic Material), tgAlim (Food), tgTransp (Transport), tgManut (Maintenance), tgInvest (Investments), tgConv (Covenant), tgOutros (Other)</td>
</tr>
<tr>
<td>Folder Name (subfunction)</td>
<td>Financial Administration, General Administration, Planning and Budget, Food and Nutrition - School Meals, School Transportation, Social Communication, Information Technology, Expenses defrayed with Petroleum Royalties and Indemnity Resources, Special Education, Youth and Adult Education, Elementary Education, Elementary Education - Excluding FUNDEB, Human Resources Training, Standardization and Inspection, Other Special Charges, Worker Protection and Benefits, External Debt Refinancing, Internal Debt Refinancing, External Debt Service, Internal Debt Service</td>
</tr>
<tr>
<td>G/L Account</td>
<td>from 3.31.90.01.00.00 to 3.46.00.00.00 (more than 300 analytical G/L accounts)</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020).

4.3.2. Consolidation of G/L Accounts

The dataframe of municipalities resulting from the pivoting presented many columns (387), mainly accounting accounts. The exploratory data analyses, performed initially, proved a high number of zero values for many of these accounts, which motivated the replacement of some analytical accounts by the equivalent synthetic account, because it was not intended to analyze the accounting accounts at a very detailed level. In other words, it is enough to identify the municipalities that have anomalies in spendings related to, for example, Patronal Obligations (composed of 14 analytical accounts, as listed in Figure 11), and not in each of these 14 accounts.

Figure 11 – Example of a synthetic G/L account to be considered in the dataframe
Source: Prepared by the author (2020).

The list below shows the G/L accounts in which the replacement of the analytical accounts by the synthetic equivalent was made, generating a new dataframe with less than 200 columns.

- Accounts (3.31.90.04.**) under Permanent Contract. (3.31.90.04.00.00)
- Accounts (3.31.90.11.**) in Salaries and Fixed Benefits - Civilian Personnel (3.31.90.11.00.00)
- Accounts (3.31.90.13.**) in Patronal Obligations (3.31.90.13.00.00)
- Accounts (3.33.50.43.**) in Social Grants (3.33.50.43.00.00)
- Accounts (3.33.90.30.**) in Consumable Material (3.33.90.30.00.00)
- Accounts (3.33.90.36.**) in Other Third-Party Services - PF (3.33.90.36.00.00)
- Accounts (3.33.90.39.**) in Third-Party Services - PJ (3.33.90.39.00.00)
- Accounts (3.33.90.47.**) in Tax and Contributory Liabilities (3.33.90.47.00.00)
- Accounts (3.44.90.51.**) in Works and Facilities (3.44.90.51.00.00)
- Accounts (3.44.90.52.**) in Equip. and Permanent Material (3.44.90.52.00.00)

4.3.3. Summary of the municipalities dataframe

After performing the pivoting and the consolidation of some accounting accounts, we have a dataframe of municipalities with their characteristics: IBGE information (UF, region, total population); INEP information (number of enrollments, schools, and teachers; IDEB scores; dropout rate); UNDP information (IDHM); and spending information, types of spending, sub-functions, and accounting accounts.

Figure 12 – Summary of the attributes of the municipalities dataframe
5. MODELING PHASE

This phase comprises the selection and application of techniques to create models and discover knowledge. In the present work, Exploratory Data Analysis (EDA) was performed to discover relevant facts. Then, Clustering (to create clusters of similar municipalities) and Anomaly Detection (to identify, in each cluster, anomalous spendings) techniques were used.

5.1. EXPLORATORY DATA ANALYSIS

AED aims to summarize and visualize data before creating models, allowing to understand their properties, inspect their qualitative characteristics and discover new patterns (LEEK, 2015). In the present work, AED was performed with descriptive statistics and graph plotting to learn about the data and summarize its most relevant characteristics; to detect hidden patterns; and to identify correlations between variables.

In the spending dataframe, the intent was to extensively explore municipalities’ paid spendings on Elementary Education in 2018 to:

- to have a sense of order of magnitude of the amounts of spending;
- detect behaviors of the total expenses paid per expense group, subfunction, type of expense and per accounting accounts; and
- preliminarily, to verify the existence of expenses of abnormal values.

In the dataframe of municipalities, the objective was to study more specifically the behaviors of the other variables. As examples of these studies, one can cite:

- What is the distribution of the population, the number of schools and teachers? What about the HDI and IDEB indicators?
- How does each spending group behave?
- What correlations exist between the variables?
All the AED performed in both dataframes are in jupyter notebooks, which can be consulted if needed.

5.2. CLUSTERING OF SIMILAR MUNICIPALITIES

5.2.1. Objectives of the clustering of municipalities

In the exploratory analyses, it was possible to perceive some correlations between the data - such as population with INEP’s quantitative data (above 0.86) and the HDI indices with IDEB_AI (above 0.50) - motivated the clustering of similar municipalities, without considering the spending data or the accounting accounts. It is assumed that similar municipalities (with similar populations, similar indicators, and similar numbers of schools, teachers, and students enrolled) must also have similar educational expenses, at least in the same order of magnitude.

Four clustering algorithms were tested - this chapter details the two algorithms with the best results: k-Means and Agglomerative Clustering. Some results obtained with the DBSCAN algorithm are also demonstrated.

After creating the groupings of municipalities, some graphs were created to facilitate the visualization of the coherence of these sets - that is, to validate if a certain algorithm was able to separate the data well.

5.2.2. Deciding on the scaling of the data

Variable scaling is a method used to standardize a range of independent variables, and is also referred to as data normalization (SRIVASTAVA, 2019). One should therefore standardize the data of municipalities’ characteristics (IBGE, INEP, and UNDP) on the same scale before using any clustering algorithms - especially when the similarity is based on the calculation of distances between points.

There are several techniques for scaling variables - in this work, the scaling techniques defined in the table below were tested for each clustering algorithm used. Population graphs with number of students enrolled (all in scaled values) were created to support the choice for the most appropriate scaler.

Table 8 – Techniques used for the transformation of variables

<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Scaler</td>
<td>It assumes that the data is normally distributed on each variable and therefore scales so that the distribution has mean value zero with standard deviation value 1. If data is not normalized, it is not a good scaling alternative.</td>
</tr>
<tr>
<td>Robust Scaler</td>
<td>It uses a similar approach to MinMax scaling, but uses the interquartile range instead of the range 0 to 1. It is suitable for data with the presence of outliers.</td>
</tr>
<tr>
<td>MinMax Scaler</td>
<td>Reduces the data to take the range of values between 0 and 1, or -1 to 1 if there are negative values. This technique is suitable for distributions that are not Gaussian or when the standard deviation of the variables is small. It is sensitive to outliers.</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020), adapted from (SRIVASTAVA, 2019).

In the graphs in Figure 13, the left side displays all the points (to visualize the groups with the most distant points), while the right side zooms in, to present the groups that are formed with the most concentrated points.
Figure 13 – Data scaling with K-Means, DBSCAN and Aggl. Clustering

**K-Means**

![K-Means Diagram]

- **Problems:**
  - Joins similar municipalities in the same cluster.
  - Poor separability of distant points.
  - Inadequate alternative.

- **Remarks:**
  - Good separation for points far apart.
  - There is a mixture of close points.

---

**K-Means**

![K-Means Diagram]

- **Problems:**
  - Joins similar municipalities in same cluster.
  - Poor separability of distant points.
  - Inadequate alternative.

- **Remarks:**
  - Good separation for points far apart.
  - There is a mixture of close points.

---

**K-Means**

![K-Means Diagram]

- **Problems:**
  - Joins similar municipalities in the same cluster.
  - Poor separability of distant points.
  - Inadequate alternative.

- **Remarks:**
  - Good separation for points far apart.
  - There is a mixture of close points.

---

**K-Means**

![K-Means Diagram]

- **Problems:**
  - Joins similar municipalities in the same cluster.
  - Poor separability of distant points.
  - Inadequate alternative.

- **Remarks:**
  - Good separation for points far apart.
  - There is a mixture of close points.
**Agglomerative Clustering**

![Image of plots showing agglomerative clustering results](image)

Source: Prepared by the author (2020).

The results presented prove it:

- K-Means: RobustScaler showed the best results - the data is well separated on both sides of the graph, although there is a little mixing of the points closest to zero;

---

**DBSCAN**

![Image of plots showing DBSCAN results](image)

Source: Prepared by the author (2020).
• Agglomerative Clustering: RobustScaler is the best alternative among all graphs presented (as outlined in green);
• DBSCAN: Although there are few groups, StandardScaler proved to be the most suitable for separating the data from these groups.

The figure below shows other graphs created, with other variables, using the Agglomerative Clustering algorithm and Robust Scaler scaling technique.

**Figure 14** – Data scaling with Aggl. Clustering and RobustScaler

5.2.3. **k-Means Clustering**

Given the number k of clusters, K-Means partitions a set of points into K groups (ZAKI and MEIRA JR., 2014), so that each point is allocated to the group that is closest to it. Thus, the clusters are created based on the minimum distances between the points belonging to each cluster and its centroid (center point of the cluster). One of the challenges in using K-Means for clustering is finding the optimal number of clusters - the one that maximizes the differences between clusters (inter-cluster) and minimizes the variations within a cluster (intra-cluster), as shown in the figure below.
When few clusters are formed, you have inter-cluster maximization, but intra-cluster variations suffer, since points that are too far apart can be in the same cluster. On the other hand, by increasing the number of clusters, the inter-cluster differences become small, although the advantage is that intra-cluster variations decrease.

It is therefore necessary to find the optimal point, in which the points of each cluster are as homogeneous as possible and which clusters formed are sufficiently different from each other. A few methods of selecting this optimal point were used with the data from the municipalities, namely: Elbow (inertia calculation), Gap Statistics and Silhouette Coefficient. The figure below displays the results found in each method.

Although the suggestions vary, the value of 7 clusters was chosen because 3 clusters are already allocated to few anomalous counties. Figure 17 shows the Silhouette Analysis plot (ALVES, 2019) for the value of 7 clusters, using population and number of students. One notices that each extreme point...
(right side) was in different clusters (these are SP and RJ municipalities that are, in fact, distinct), while concentrated points (left side) were appropriately separated according to population and number of students ranges.

**Figure 17** – Silhouette analysis for k-Means clustering with 7 clusters.

![Silhouette analysis](image1)

Source: Prepared by the author (2020).

Based on the choice of the value of 7 clusters, Figure 18 presents the results of the clustering with the k-Means algorithm and RobustScaler scaling, showing the quantity of municipalities in each cluster, region, and UF.

**Figure 18** – Details of clusters with k-Means and RobustScaler scaling.

![Cluster details](image2)

Source: Prepared by the author (2020).
You can see some characteristics of k-means clustering:
- specific clusters are created with extreme value municipalities (municipalities with the highest population, the most students or teachers);
- SP, with the largest population, was in one cluster; while RJ (half the population of SP, but larger number of enrollments), was in another cluster;
- When using a K-value greater than 10, specific clusters are created with minimum value municipalities (municipalities with few schools, for example).
- clusters of a few municipalities (less than 1% of the total) are usually created, which can be considered cluster anomalies.

5.2.4. DBSCAN Clustering

Unlike k-Means, which seeks proximity by the distance between points, DBSCAN uses the local density of the points—that is, it uses the calculation of areas of highest and lowest point density—as the criteria for forming the clusters (ZAKI and MEIRA JR., 2014). It does not require the number of clusters and allows the definition of irregular clusters.

In this work, it was agreed to use certain parameters so that there would not be too much noise (points that are not allocated to any cluster), but that there would be at least 6 clusters (except cluster -1, composed of noise). Based on the parameters eps=0.9 and min_samples=5, Figure 19 presents the results of clustering with the DBSCAN algorithm and StandardScaler scaling, showing the number of municipalities in each cluster, region, and UF.

Figure 19 – Details of clusters with DBSCAN and StandardScaler scaling.

Source: Prepared by the author (2020).
You can see some characteristics of DBSCAN clustering:

- The clusters are not well separated as in k-Means, because the separation of clusters considers bands in which there is a high density;
- the extreme points (municipalities with high population, few students or few teachers) are not allocated to some cluster, but inserted in cluster noise. DBSCAN identifies the outlier points, unlike k-Means (which allocates them to a small cluster size);
- is not the most appropriate algorithm for identifying anomalies in spending, because many points are allocated to few clusters (and the study requires a few similar clusters, not large clusters of similar density).

5.2.5. Agglomerative Clustering

Hierarchical algorithms create a hierarchical structure of nested points according to a clustering strategy and a linking criterion (dissimilarity metric). The clustering strategies can be (ZAKI and MEIRA JR., 2014):

- Agglomerative (bottom-up approach): Each point is a cluster, and pairs of clusters are merged successively as you move up the hierarchy;
- divisive (top-down approach): all points are in a cluster, and the divisions are performed recursively as you move down the hierarchy.

The dissimilarity between clusters is calculated according to the method chosen in the algorithm:

- Single: the distance between 2 clusters is given by the smallest distance between two points (minimum distance, nearest neighbor). It is sensitive to outliers;
- Complete: the distance between 2 clusters is given by the greatest distance between two points (maximum distance, farthest neighbor). Tends to generate very large clusters;
- Average: the distance between 2 clusters is given by the average of the distances between each two points (distance between the centroids). It is less sensitive to outliers, but tends to generate large, globular clusters;
- Ward: Minimizes the sum of differences between points in clusters. Most effective when you have outliers; tends to generate more regular cluster sizes.

The results of hierarchical clustering are usually shown as a tree of groups or dendrograms, from which the number of clusters can be obtained. In the present work, the hierarchical agglomerative algorithm was used. As shown in Figure 20, just to make it possible to obtain the dendrograms below, the dataframe was filtered to contain the records of municipalities with populations of up to 11 thousand inhabitants (50% of the data) and using the Ward method.

Figure 20 – Generating dendrograms with the Ward method.
Despite the suggestions indicated by the dendrograms, the choice of 7 clusters was followed. The following figure shows the graph resulting from the application of Agglomerative Clustering with 7 clusters, with the variables population and number of students. Clearly, the use of RobustScaler brings a good separation of the clusters if compared with MinMaxScaler.

**Figure 21** – Generating a Graph with Clusters Generated by Agglomerative Clustering

Figure 22 presents the results of the application of Agglomerative Clustering with RobustScaler, showing the number of municipalities in each cluster, region, and Federal State.

**Figure 22** – Details of clusters with Aggl. RobustScaler Clustering and Scaling.
Although hierarchical algorithms are simple and effective, with ease in visualizing the clusters in dendrograms and flexibility in choosing the cutoff to determine the number of clusters, there are some drawbacks. Similar to k-Means, the choice of the number of clusters is not trivial - one must decide where to draw the line in the dendrogram to get the number of clusters; and furthermore, dissimilarity methods can produce very different results, and there is no objective criterion to evaluate which method is best.

5.2.6. Validation of the clustering algorithms

Some graphs produced in the jupyter notebook (with log normalized variables) were chosen to be presented in the present work, with the objective of comparing the results of some algorithms used. This comparison not only seeks to demonstrate the consistency of the separability of the clusters, but guided the choice of algorithm for the anomaly detection task.

Figure 23 shows the clusters formed for the algorithms k-Means, DBSCAN and Agglomerative Clustering (AgrC), considering the data of population, number of students enrolled and number of teachers. The clusters formed by k-Means and AgrC are quite similar, but AgrC showed better separability in dense areas; DBSCAN is no longer very suitable because it grouped many points in the same cluster and inserted the extreme points in the cluster noise (which do not appear in the graph).

Figure 23 – Comparison of the results of some clustering algorithms
The following figures compare only the results of the k-Means and Agglomerative Clustering (AgrC) algorithms.

Figure 24 – Comparison of the results of some clustering algorithms

Qtd Escolas x Qtd matrículas x Qtd Docentes
In fact, the algorithm with the best results, considering the objectives of the present work and the generation of graphs with several variables (available in the jupyter notebooks), is Agglomerative Clustering with data scaled with RobustScaler.

Because of the choice for Agglomerative Clustering, a measure of the evaluation of the quality of separation between clusters - the Davies-Bouldin index, which compares the distance between clusters with the size of the clusters themselves (DAVIES and BOULDIN, 1979) - was used in the jupyter notebook. It is used mainly when the data is unlabeled. A value closer to zero means a model with better separation between the clusters.

The best indices were achieved with 2 and 5 clusters (below 0.4); and the index was not satisfactory for 7 clusters (0.69). However, we kept this value, because a larger number of clusters is of interest for local anomaly detection. If there are few clusters, few anomalous points will be identified.

5.3. ANOMALY DETECTION

Outlier or anomaly detection aims to identify unexpected items in the dataset that differ from the norm (GOLDSTEIN and UCHIDA, 2016). In general, such discrepant values are removed prior to the use of data mining algorithms. For the present work, however, the goal is in fact the detection of discrepant spendings in each group of municipalities, which may indicate failure to fill out records, possible irregularities or atypical events that must be justified (a possible construction work in a school, for example).

In this sense, the focus is on the use of unsupervised anomaly detection algorithms, which use only the intrinsic information of the data to detect the points that deviate from the others. We tried to use algorithms based on different detection criteria - distance, similarity and density.
5.3.1. Delineation of outlier detection strategies

The exploratory analysis allowed the identification of global outliers (atypical values when compared to the entire data set). However, there is interest in detecting other types of outliers, listed and represented below.

Table 9 – The types of anomalies

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>Extreme values (low or high value) very different from the other points, easily detectable by statistical techniques, histograms and scatter plots.</td>
</tr>
<tr>
<td>Clusters</td>
<td>Few points together in a cluster, far away from other denser clusters, detectable by applying clustering algorithms such as k-Means.</td>
</tr>
<tr>
<td>Locations</td>
<td>Few points close to the other dense clusters, detectable by specific outlier detection algorithms, such as LOF. These points look normal, and are noticed when analyzing a particular cluster in isolation.</td>
</tr>
</tbody>
</table>


Figure 25 – Graphic representation of the types of anomalies


Anomaly detection algorithms generate two results: a label indicating whether an instance is anomalous or not; and a score that indicates the degree of abnormality (GOLDSTEIN and UCHIDA, 2016). In unsupervised algorithms, scores are more common and allow the classification (ranking) of the most anomalous points. By means of a threshold, the classification can be converted into a label.

5.3.2. The Python Outlier Detection (PyOD) library

To identify the discrepant spendings of municipalities, the Python Outlier Detection (PyOD) library was used, which has a variety of models for anomaly detection in multivariate data (ZHAO, NASRULLAH, and LI, 2019). The following figure presents graphs of some PyOD algorithms (black dots represent outliers); and the table below lists the algorithms applied in the present work.

Table 10 – Anomaly detection algorithms used on the data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle-based Outlier Detection (ABOD)</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>Local Outlier Factor (LOF)</td>
<td>Distance-based</td>
</tr>
<tr>
<td>Cluster-based Local Outlier Factor (CBLOF)</td>
<td>Distance/Density based</td>
</tr>
<tr>
<td>Histogram-based Outlier Detection (HBOS)</td>
<td>Based on statistics</td>
</tr>
<tr>
<td>K Nearest Neighbors (KNN)</td>
<td>Distance-based</td>
</tr>
<tr>
<td>Average KNN (A_KNN)</td>
<td>Distance-based</td>
</tr>
</tbody>
</table>
Isolation Forest (IF)  |  Ensemble
---|---
Feature Bagging (FB)  |  Ensemble

Source: Prepared by author (2020), adapted from ZHAO, NASRULLAH, and LI (2019)

**Figure 26** – Listing some models available in the PyOD library

The ABOD algorithm has the distinction of using angle-based measures. In multidimensional spaces, angles are more stable measures than distance measures, and a good example is the cosine-based similarity measures for text analysis (KRIEGEL et al, 2008). As per the following figure, a point in space is considered an outlier if most of the other points are in similar directions.

**Figure 27** – Anomaly detection in the ABOD algorithm

Source: ZHAO, NASRULLAH, and LI (2019).

The ABOD algorithm has the distinction of using angle-based measures. In multidimensional spaces, angles are more stable measures than distance measures, and a good example is the cosine-based similarity measures for text analysis (KRIEGEL et al, 2008). As per the following figure, a point in space is considered an outlier if most of the other points are in similar directions.


The LOF algorithm was the first to introduce the idea of identifying local anomalies (GOLDS-TEIN and UCHIDA, 2016). The anomaly score is based on a local density measure and the deviation of that measure between a point and that of its neighbors. Normal instances, with similar densities of their neighbors, contain a score close to 1; anomalous instances, with substantially lower local density
than their neighbors, have higher scores. It requires a number for k (the number of neighbors).

The CBLOF algorithm uses clustering to determine the dense areas in the data, create clusters, and calculate their density estimate (GOLDSTEIN and UCHIDA, 2016). The anomaly score of a point is based on the size of the cluster to which it belongs (parameter disabled by default) and the distance from the nearest largest cluster. Consequently, all points in small clusters far from the largest clusters are anomalous (HE et al, 2003) - thus, it locates global and cluster anomalies, but not local ones.

The HBOS algorithm assumes that variables are independent, and calculates the degree of anomaly of a data instance through histograms (GOLDSTEIN and UCHIDA, 2016). It is considered a simple and fast running algorithm, but with lower accuracy.

The KNNN algorithm calculates the anomaly score based on the distance from a point to the k-th nearest neighbor, with three possible methods of calculation - the longest distance, the average (A-KNN) or median of the distances of all k-th nearest neighbors (ZHAO, NASRULLAH and LI, 2019). It therefore detects global anomalies, not local ones.

The Isolation Forest algorithm performs data partitioning using a tree structure. The anomaly score considers how isolated a given point is in the structure, when few partitions are needed for its isolation.

The Feature Bagging algorithm uses multiple detection algorithms (could be LOF, kNN, and ABOD) on different samples from a multivariate dataset, and uses averaging measures, or other combined metrics, to calculate the anomaly score (ZHAO, NASRULLAH, and LI, 2019). It is a technique that seeks to reduce variation between algorithms to improve efficiency and avoid overfitting.

5.3.3. Choice of Cluster to be submitted to the algorithms

Based on the results of the Agglomerative Clustering, cluster 6 was chosen and submitted to the eight detection algorithms chosen, in different data scopes: all attributes (expense groups, expense types, programs, subfunctions and accounting accounts); Own and FUNDEB expenses; and compensation and maintenance expense types. The abnormality scores and labels for each municipality, produced by each detection algorithm, in each data scope, were stored in the municipalities dataframe.

5.3.4. Results of anomaly detection

The figure below displays the characteristics of the chosen cluster, the scatter plot with the indication of the anomalies and the number of anomalous municipalities in each algorithm, in the scope of all attributes.

Figure 28 – Indication and quantities of anomalous municipalities (scope: all attributes)
The graph below presents the scatter plot of Own Expenses with FUNDEB Expenses, and right below it displays the results of the algorithms in the scope of spending groups (i.e., the input to the algorithms is only the Own Expenses and FUNDEB Expenses data). All algorithms (except LOF) considered the extreme points (high value of expenses), above the dotted band 1, as anomalous. Between dotted bands 1 and 2, few algorithms (ABOD, CBLOF, kNN) detected as anomalous some innermost points, located between normal regions. Below the dotted band 3, the LOF and FB algorithms detected as anomalous a large concentration of points very close to the normal points.

Figure 29 – Indication of the outliers in the scope of the spending group
Some considerations are necessary:

- the red line represents the threshold of the statistically established anomaly grade score (score at the 5% percentile) - all points above the line are classified as anomalies;
- The six color layers represent ranges of anomaly score values, with the first layer starting after the threshold value.

5.3.5. Validation of anomaly detection models

It is necessary to assess the reliability of the models generated. There are several known metrics for the validation of supervised algorithms, such as accuracy, precision, confusion matrix, and cross-validation. In the case of anomaly detection algorithms, the traditional methods for evaluating the quality of abnormality scores (scoring) are the ROC (Receiver Operating Characteristic) curve and the PR (Precision-Recall) curve - but only when labels (classes) are available (GOIX, 2016).

In the present work, however, there are no labels - that is, there are no real data on municipalities that presented discrepant spendings in education. Thus, the validation of the generated models, in an unsupervised manner, is not a trivial task. A viable alternative was to compare the statistics and graphs between the normal and anomalous data sets - to ascertain whether such abnormalities are consistent.
The execution of the 8 algorithms, on the entire data set, pointed out 375 anomalous municipalities (cited in at least one algorithm), or 11.5% of the total number of municipalities, with 10 anomalous municipalities in common (cited by all algorithms), listed in figure 30.

A simplified analysis was performed in these ten anomalous municipalities. According to the statistics and histograms of population and numbers of schools, students and teachers in figure 31 - these are municipalities that are above the 75% percentile range when compared to the whole set. In addition, the other histograms of spendings, of some functions and of some accounting accounts, in figures 32 and 33, presented curves more shifted to the right - which may have influenced the abnormality score for certain municipalities.

Then, the municipality of Bombinhas was chosen for an even more detailed analysis (because it has the smallest population), as shown in Figure 34. Comparing some histograms of Bombinhas with all the other municipalities, one can clearly see that there are expenses of atypical values when compared to the same expenses of its peers (the values for Bombinhas are represented by the vertical red line, located at the end of the curves that represent the intervals of the other municipalities). Such values should be presented to the control bodies for the appropriate investigations or measures.

It can be said, therefore, that the generated models are valid and allow the identification of relevant facts.

Figure 30 – The 10 outliers in all detection algorithms
Figure 31 – The 10 anomalous municipalities - Comparison of IBGE and INEP data

Source: Prepared by the author (2020).
## Figure 32 – The 10 Outliers - Comparison of Spending Data

<table>
<thead>
<tr>
<th>Estatísticas dos municípios não-anômalos</th>
<th>Estatísticas dos municípios anômalos</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DespFUNDEB</strong></td>
<td><strong>DespFUNDEB</strong></td>
</tr>
<tr>
<td><strong>count</strong></td>
<td>3233.00</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td>4289156.39</td>
</tr>
<tr>
<td><strong>std</strong></td>
<td>4193133.01</td>
</tr>
<tr>
<td><strong>min</strong></td>
<td>0.00</td>
</tr>
<tr>
<td><strong>25%</strong></td>
<td>1334999.60</td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td>2726222.20</td>
</tr>
<tr>
<td><strong>75%</strong></td>
<td>587849.90</td>
</tr>
<tr>
<td><strong>max</strong></td>
<td>29267721.60</td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020)
Figure 33 – The 10 Anomalous Municipalities - Comparison of sub-functions and G/L accounts

<table>
<thead>
<tr>
<th>Estatísticas dos municípios não anômalos</th>
<th>Estatísticas dos municípios anômalos</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AdmGeral</strong></td>
<td><strong>PlanOrc</strong></td>
</tr>
<tr>
<td>count</td>
<td>Loc Mão de Obra</td>
</tr>
<tr>
<td>mean</td>
<td>Inden e Rest Trab</td>
</tr>
<tr>
<td>std</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>max</td>
<td></td>
</tr>
</tbody>
</table>

Source: Prepared by the author (2020).
Figure 34 – An example of an anomalous county

Source: Prepared by the author (2020)

6. EVALUATION AND IMPLEMENTATION PHASE

The Evaluation phase examines whether the results of the model meet the business objectives and success criteria, and determines the decision whether the model should be deployed or submitted
to new phase iterations. The Deployment phase is all about putting the obtained model into production, including, when applicable, writing the final report.

Detecting anomalies in a database with hundreds of attributes in an unsupervised manner has proven to be a complex task, as there are many challenging issues, such as:

- decide which is the best scaling algorithm for the data, as well as the most appropriate clustering algorithm, given the difficulty of visualizing multidimensional data in only two or three dimensions;
- define the most relevant attributes as input for the algorithms (in CGEBC’s words, all expenses and G/L accounts are indispensable and should not be discarded during data analysis);
- establishing the threshold of the abnormality score in an objective way, since the boundary between normal and abnormal is not precise;
- identify the innermost anomalous points, located in regions with two or more clusters; and
- validate the accuracy of the results in a situation without historical data.

Furthermore, faced with the great diversity of detection algorithms (of different criteria - statistical, distance, density, similarity), it is necessary to choose which to apply and optimize the parameters of each one. Some algorithms still require the definition of specific criteria, such as the number of clusters (CBLOF) and the number of near neighbors (kNN and Average kNN). Due to the limited resources available (personnel and time), all these issues could not be worked on in adequate depth.

On the other hand, it can be said that the present work achieved the main purpose of detecting anomalies in municipalities’ spending on Primary Education (the dataframe of municipalities is updated with the labels and scores of each algorithm), given that there was no previous statistical or data analysis work on the SIOPE by CGEBC - and it was also possible to achieve the objectives defined in item 3.5:

- obtaining at least 1% of federal entities with discrepancies in their declared educational spendings (when considering all attributes) - according to Figure 28, the percentage of each algorithm ranged from 2 to 5%; and
- In fact, some entities were identified as anomalous in all the detection algorithms chosen (see Figure 28, there were ten municipalities identified as anomalous by the eight algorithms applied).

However, it was realized that for this detection activity to be more effective, it is still necessary to establish some more individualized strategies. One example would be to determine the most relevant expenses - in case it was food, one could cross-reference SIOPE data with SigPC databases (FNDE’s accountability database that contains information about PNAE), to detect anomalies in expenses with school meals. Thus, a recommended path for the present work would be to return to the business understanding phase to reformulate more specific objectives, and present the consolidated results in a management dashboard.

Nevertheless, the present work presented a proposal, a feasible and productive path to reach anomalies in municipalities’ expenses with Primary Education - through the concomitant use of data mining, clustering algorithms and algorithms focused on anomaly detection. It is necessary to rea-
pply such paths for anomaly detection in the other education modalities (Kindergarten, High School, etc.). The present work, therefore, is not finished - in fact, it needs to be submitted to new iterations of phases, with reformulations of business and data mining objectives, and to be presented to the Basic Education audit area. One possibility would be the formulation of a consolidated panel, with the indication of the anomalous municipalities for each modality of education, also including, for each case, the attributes involved that influenced a higher score for the anomalies.

7. CONCLUSION

It is understood that exploring public spending data to obtain strategic and value-added knowledge represents an important objective for internal control and for the Public Administration - because it enables, in the case of the SIOPE, the identification of atypical spendings (anomalies) that may indicate possible failures or irregularities in public investments in education. Consequently, we see the opportunity to offer the results of the detection of these anomalous spendings as inputs for a series of activities, namely: planning of control actions (in this case, complementing the Vulnerability Matrix of the municipalities); monitoring and adoption of appropriate measures by the control bodies; and creation of audit trails aimed at monitoring public spending.

Thus, the present work had as its initial objective an extensive exploratory analysis of the SIOPE data, the results of which delimited the scope of the expenses to be investigated (i.e., the municipal expenses paid in Primary Education in the year 2018) and determined the following strategies - the use of the clustering and anomaly detection techniques.

Based on the idea that similar municipalities should also present similar educational expenses, at least in order of magnitude - the clustering sought to group similar municipalities, i.e., similar as to population data; quantities of students, teachers and schools; and IDEB and IDHM indicators. Several runs were performed with different algorithms (k-Means, DBSCAN, MeanShift, and Agglomerative Clustering), different ways of scaling the data (StandardScaler, RobustScaler, and MinMaxScaler), and different parameters (number of clusters, epsilon, number of samples, link criteria, etc.). Some methods for calculating the optimal number of clusters were also used (Elbow, Gap Statistics and Silhouette Coefficient methods). The Davies-Bouldin index was used as a measure of the evaluation of the quality of separation between the clusters, which was interpreted in conjunction with graphs that allowed the consistency of the separability of the clusters to be verified. Despite suggestions (2 or 6 clusters) by the mentioned methods, we defined as seven the ideal number of clusters - because it was already expected the creation of two clusters with only one municipality (SP and RJ), and two other clusters with few municipalities (less than 1%). It was concluded that the Agglomerative Clustering, with data scaled with RobustScaler, presented the best results.

Finally, preference was given to the cluster with the highest number of municipalities to submit it to eight anomaly detection algorithms from the PyOD library. An additional parameter was the definition of which attributes to submit to the algorithms (spending scope) - i.e.: all spending attributes; only groups of own spending and FUNDEB spending; and remuneration and maintenance spending types. The results obtained (classification of the municipality, if anomalous or not; and abnormality score, calculated by each algorithm in each spending scope) were consolidated to the da-
taframe of municipalities. Other metrics were also added, such as the number of times a municipality was marked as anomalous. In the scope of all attributes, it was possible to identify the expenses that influenced the abnormality score (through comparative histograms).

It can be said that it was possible, throughout the work, to identify the global, local, and cluster anomalies - that is, the atypical spendings of each municipality with respect to its similar municipalities.

The set of models generated met the business objectives (indicate the municipalities with discrepancies in their educational expenses) and the success criteria (obtain at least 1% of municipalities with discrepancies in their educational expenses, some of them being pointed out as anomalous in all algorithms) defined in item 3.6. Similarly, it can be said that all the data mining objectives (performing AED, clustering and anomaly detection tasks) were also achieved.

As recommendations and suggestions for future work, there are a multitude of possibilities:

- detect the anomalies in Primary Education spending for the other clusters of municipalities, and reapply the techniques used (AED, clustering, and spending anomaly detection) in the other education modes (Early Childhood Education, Secondary Education, Higher Education);
- perform the same studies for the State SIOPE data;
- perform data analysis considering the per capita spending figures (by dividing the spending by the estimated population or the number of students enrolled);
- create normal profile of spending on education, for each municipality or group of similar municipalities (involves analyzing SIOPE data from years prior to 2018) - to enable more immediate detection of anomalies from this spending profile, every bimester (as it is the periodicity of SIOPE data transmission);
- propose or seek measures of accuracy of the results of unsupervised algorithms;
- consolidate the classifications and scores of the anomaly detection algorithms in a business intelligence platform, to present managerial dashboards, with multiple views of the data, to the audit teams;
- study the feasibility of cross-referencing the SIOPE with other databases of the federal entities, such as the SIAFEM base (this is a system equivalent to the SIAFI, but within the scope of the states and municipalities);
- verify possibilities of applying deep learning techniques for anomaly detection, such as auto-encoders and generative adversarial networks.


KRIEGEL, Hans-Peter, SCHUBERT, Matthias, and ZIMEK, Arthur. **Angle-based outlier detection in high-dimensional data.** In Proceedings of the 14th ACMKDD International Conference on Knowledge Discovery and Data Mining (pp. 444-452). Association for Computing Machinery. Las Vegas, NV, 2008.


