Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

RENATA GUANAES MACHADO

SUBSÍDIO ÀS FISCALIZAÇÕES PÚBLICAS:
Identificação dos Municípios com gastos discrepantes na Educação Básica

Brasília

2020

RENATA GUANAES MACHADO

SUBSÍDIO ÀS FISCALIZAÇÕES PÚBLICAS:
Identificação dos Municípios com gastos discrepantes na Educação Básica

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle, realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Orientador: Prof. MSc. Rodrigo Peres Ferreira

Brasília

2020

REFERÊNCIA BIBLIOGRÁFICA

MACHADO, Renata Guanaes. **Subsídio às fiscalizações públicas:** Identificação dos Municípios com gastos discrepantes na Educação Básica. 2020. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília, DF. 128 fl.

CESSÃO DE DIREITOS

NOME DO AUTOR: Renata Guanaes Machado

TÍTULO: Subsídio às fiscalizações públicas: Identificação dos Municípios com gastos discrepantes na Educação Básica

GRAU/ANO: Especialista/2020

_____

Renata Guanaes Machado

renata.guanaes@cgu.gov.br

Ficha catalográfica

Machado, Renata Guanaes

Subsídio às fiscalizações públicas: Identificação dos Municípios com gastos discrepantes na Educação Básica / Renata Guanaes Machado; orientador, Rodrigo Peres Ferreira, 2020.

128 p.

Monografia (especialização) – Escola Superior do Tribunal de Contas da União, Curso de Especialização em Análise de Dados para o Controle, Brasília, 2020.

Inclui referências.

1. Mineração de Dados. 2. Detecção de Anomalias. 3. Despesas Públicas. 4.Educação Básica. I. Ferreira, Rodrigo Peres. II. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle. III. Título

RENATA GUANAES MACHADO

SUBSÍDIO ÀS FISCALIZAÇÕES PÚBLICAS: Identificação dos Municípios com gastos discrepantes na Educação Básica

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle, realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Brasília, 25 de março de 2020.

Banca Examinadora:

_____

Prof. Rodrigo Peres Ferreira, MSc.

Orientador

_____

Prof. Edans Flavius de Oliveira Sandes, Dr.

Instituto Serzedello Corrêa - TCU

Ao meu eterno amor, a única luz que me guia, me fortalece e me engrandece como ser humano, **Arton Perez Peixoto**, por sempre me incentivar a explorar todo o meu potencial em quaisquer jornadas que me arrisco a enfrentar.

AGRADECIMENTOS

Sou grata às pessoas que me proporcionaram não só adquirir novos conhecimentos e habilidades, mas também muitos momentos felizes. Reitero às pessoas e instituições, listadas abaixo, os meus mais sinceros agradecimentos.

Primeiramente, agradeço ao meu esposo – ouso dizer, a pessoa mais importante em toda a minha vida – pelo apoio incondicional e irrestrito, por aceitar as minhas várias ausências e, principalmente, por profundamente entender o meu incansável anseio em caminhar pela longa estrada do conhecimento.

Ao meu ex-chefe, Leonardo Jorge Sales, por ter me apoiado nos trâmites burocráticos para que eu pudesse participar desta especialização. A ele e ao Rodrigo Peres, sou muito grata por terem me concedido esta excepcional e única oportunidade de aprendizado.

Ao meu orientador e colega de trabalho, mais uma vez, Rodrigo Peres, que aceitou o desafio de acompanhar a saga deste trabalho, e que soube passar tranquilidade e dicas valiosas.

Aos meus colegas da especialização, em especial, ao Marcus Vinícius Borela de Castro, Cláudio Augusto Grunewald Soares e Ricardo de Farias Santos, pelas dicas de trabalho, pelas amizades, pelas risadas e pelos bons cafés.

Aos professores do curso que, em sua grande maioria, planejaram com entusiasmo e dedicação suas aulas e exercícios extraclasse. Gostaria de agradecer, em especial, ao professor Edans Flavius de Oliveira Sandes, que aceitou o convite de compor a banca examinadora. Agradeço, mais uma vez, a todos os professores, por contribuírem para a minha formação.

Ao ISC (Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa) e seus colaboradores, pela infraestrutura e pelo ambiente de ensino, ambos impecáveis.

Ambição é o caminho para o sucesso. Persistência é o veículo no qual se chega lá.
(William Warren Bradley, 1943 -)

Esforço contínuo – não a inteligência ou a força – é a chave para liberar o nosso potencial.
(Winston Churchill, 1874 - 1965)

Pegue um método e tente. Se falhar, admita-o francamente e tente outro. Mas, por todos os meios, tente alguma coisa.
(Franklin Delano Roosevelt, 1882 - 1945)

RESUMO

Na Controladoria-Geral da União (CGU), os gastos públicos do Poder Executivo Federal são constantemente monitorados pelo Observatório da Despesa Pública (ODP), unidade de inteligência pertencente à Secretaria de Combate à Corrupção (SCC) – responsável pela produção de informações estratégicas que apoiam a tomada de decisão no controle interno, bem como possibilitam meios para o combate à corrupção e à má gestão dos recursos públicos. Com a disponibilização do Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE), que contém as receitas e despesas dos entes federativos com a educação, o presente trabalho tem como objetivo incluir mais uma temática à atuação do ODP, procedendo-se ao emprego de técnicas de mineração de dados para um estudo mais aprofundado das despesas registradas neste sistema. Levando-se em consideração algumas fragilidades existentes no SIOPE, apontadas por órgãos de controle interno, e de inúmeros desvios na aplicação de recursos destinados à educação (a exemplo do FUNDEB), delimitou-se o escopo do trabalho para a detecção de despesas atípicas realizadas pelos municípios com o Ensino Fundamental, no ano de 2018 – que podem constituir tão somente eventos ocasionais, específicos (como obras em escolas, por exemplo) ou representar indícios de falhas ou irregularidades nos investimentos públicos em educação. Qualquer que seja a situação, justificativas sobre tais gastos anômalos devem ser apresentadas. Para isso, análises exploratórias levaram ao estabelecimento das estratégias seguintes: a clusterização de municípios e a detecção de anomalias através do uso da biblioteca *Python Outlier Detection* (PyOD). Com base na afirmação de que municípios com dados de população e de indicadores educacionais em mesma ordem de grandeza devem ter despesas semelhantes, aplicou-se alguns algoritmos de detecção de anomalias em um grupo de municípios semelhantes. Os resultados alcançados (classificação de cada município, se anômalo ou não; e pontuação do grau de anomalia) podem ser agregados ao planejamento das ações de controle e, ainda, subsidiar a adoção de providências cabíveis por parte das demais instâncias de controle existentes, como o Ministério da Educação e conselhos de controle social.

Palavras-chave: Mineração de dados. Análise exploratória de dados. Clusterização de municípios. Detecção de anomalias. Despesas públicas. Educação Básica. FUNDEB. SIOPE.

ABSTRACT

At Brazil´s Office of the Comptroller General (CGU), public spending of the Federal Executive Branch is continuously monitored by the Public Spending Observatory (ODP), an intelligence unit belonging to the Anti-Corruption Secretariat (SCC). The ODP is responsible for the production of strategic information to support the decision-making for internal control, as well as to provide tools to fight against corruption and misuse of public resources. This paper proposes a new theme to be monitored by ODP's team: the education spending analysis through the data available of the Education Public Budgets Information System (SIOPE), which comprises the education's revenues and expenses of federative entities. It will be achieved by performing data mining techniques for a more in-depth study of the education spending recorded in SIOPE system. Considering some well-known weaknesses in SIOPE's data quality, already pointed out by internal control agencies, and numerous irregularities in resource allocation for education, this paper's scope is delimited in detecting atypical expenses incurred by the municipalities in Elementary Education, in 2018. These atypical expenses may be occasional events (such as works that are under implementation in schools) or may represent failure or fraud possibilities. In whatever situation, justifications for such anomalous expenses must be provided. Initially, exploratory analyses led to the following strategies defined: clustering municipalities and detecting anomalies through using the Python Outlier Detection (PyOD) library. Some anomaly detection algorithms were applied in a group of similar municipalities, since municipalities with similar population and educational ratings should also have similar expenses. Thus, the results achieved by algorithms (anomaly label and score for each municipality) can be added to the planning of control actions and, further, can support the adoption of appropriate measures by the other existing control instances, such as the Ministry of Education and social control councils.

Keywords: Data mining. Exploratory data analysis. Clustering of municipalities. Anomaly detection. Public spending. Elementary Education.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A_KNN – *Average K-Nearest Neighbors* (médio k-ésimo vizinhos mais próximos)

AAC – Auditoria Anual de Contas

ABOD – *Angle-based Outlier Detection* (Detecção de anomalia baseada em ângulo)

ANEB – Avaliação Nacional da Educação Básica

CACS – Conselhos de Acompanhamento e Controle Social do FUNDEB

CAUC – Serviço Auxiliar de Informações para Transferências Voluntárias

CBLOF – *Cluster-based Local Outlier Factor* (Fator de anomalia local baseado em agrupamento)

CGEBC – Coordenação-Geral de Auditoria das Áreas de Educação Básica, Direitos Humanos e Desenvolvimento Social

CGU – Controladoria-Geral da União

DBSCAN – *Density-Based Spatial Clustering of Applications with Noise* (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído)

HBOS – *Histogram-based Outlier Detection* (Detecção de anomalia baseado em histograma)

LDB – Lei de Diretrizes e Bases da Educação Nacional

LOF – *Local Outlier Factor* (Fator de anomalia local)

FB – *Feature Bagging* (Recurso de empacotamento)

FEF – Fiscalização em Entes Federativos

FNDE – Fundo Nacional de Desenvolvimento da Educação

FUNDEB – Fundo de Manutenção e Desenvolvimento da Educação Básica

FUNDEF – Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério

IBGE – Instituto Brasileiro de Geografia e Estatística

IDEB – Índice de Desenvolvimento da Educação Básica

IF – *Isolation Forest* (Floresta de Isolamento)

KNN – *K-Nearest Neighbors* (k-ésimo vizinhos mais próximos)

MDE – Manutenção e Desenvolvimento do Ensino

MEC – Ministério da Educação

ODP – Observatório da Despesa Pública

PDDE – Programa Dinheiro Direto na Escola

PNAE – Programa Nacional de Alimentação Escolar

PNATE – Programa Nacional de Apoio ao Transporte do Escolar

STN – Secretaria do Tesouro Nacional

TCU – Tribunal de Contas da União

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A corrupção pública custa, todos os anos, muitos milhões aos governos em todo o mundo, sendo o seu combate um desafio para o setor público e para a sociedade (THESING, 2019). Para alguns autores, a má gestão do dinheiro público pode acarretar perdas ainda piores do que os atos de corrupção (ANGELICO, 2012; AGUIAR, 2019). Nesse contexto, o uso conjunto de tecnologias de ponta, de métodos de inteligência e de ferramentas de análise de dados, nos últimos anos, provou-se ser um poderoso e efetivo aliado não só para enfrentar a corrupção, mas também para minimizar ou evitar o desperdício dos recursos públicos.

A despesa pública, por si só, é um ato complexo. Desta forma, o Observatório da Despesa Pública (ODP), unidade de ciência de dados pertencente à Secretaria de Combate à Corrupção (SCC) na Controladoria-Geral da União (CGU), tem como principais objetivos o monitoramento de gastos da Administração Pública do Poder Executivo Federal e a produção de informações estratégicas como apoio à tomada de decisão no controle interno.

Os temas trabalhados pelo ODP incluem: compras e licitações públicas; programas de governo; benefícios sociais; diárias e uso do cartão corporativo; convênios e transferências; entre outros. Com relação aos métodos e ferramentas utilizadas pelo ODP, pode-se mencionar: processamento analítico de dados, técnicas estatísticas, *big data*, *business intelligence*, inteligência artificial, mineração de dados e aprendizagem de máquina (*machine learning*). Em especial, o emprego das mais variadas técnicas de mineração de dados tem trazido reconhecimento e importantes benefícios à CGU, como a descoberta de conhecimento de alto valor agregado a partir das grandes bases de dados governamentais, a potencialização da capacidade de análise e a modernização do controle interno.

Pode-se citar alguns exemplos de trabalhos desenvolvidos pelo ODP, em conjunto com as equipes de auditores federais (e premiados por diversas instituições), como: a análise automatizada de licitações e contratos públicos, a criação do modelo de risco de fornecedores, a qualificação de atos do Diário Oficial da União, a malha fina de convênios, análise de vínculos, entre outros. Tais trabalhos permitem a geração de alertas sobre situações atípicas, a identificação de indícios de irregularidades, a detecção de fraudes e a verificação da má gestão dos recursos públicos – servem, portanto, de insumos diretos para: execução das ações de auditoria, fiscalização e correição; deflagração de operações especiais (em caso de maior gravidade); e, mais importante, para a elaboração de planos de prevenção da corrupção.

Nesse momento, torna-se fundamental a definição de alguns termos. A Ciência de Dados é um conjunto de princípios fundamentais que norteiam a extração de conhecimento a partir de dados; Mineração de Dados é a extração de conhecimento propriamente dita, por meio de tecnologias que incorporam tais princípios (PROVOST e FAWCETT, 2016), sendo a aprendizagem de máquina uma dessas tecnologias, na forma de algoritmos de indução (afirmar uma verdade generalizada a partir da observação de alguns elementos).

1.2 PROBLEMA E JUSTIFICATIVA

Cada vez mais, e de forma mais facilitada, os dados de diversos assuntos e de diferentes órgãos são disponibilizados às equipes de controle interno na CGU – como os dados do Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE), sob a gestão do Fundo Nacional de Desenvolvimento da Educação (FNDE).

No SIOPE, os estados e municípios registram as receitas e despesas realizadas com a educação pública – as quais são, posteriormente, transmitidas às demais instâncias de controle para fins de validação das informações. Uma vez validadas, o FNDE gera os indicadores de gestão e de conformidade às disposições legais, como, por exemplo, o cumprimento da aplicação de 25% das receitas na Manutenção e Desenvolvimento do Ensino (MDE).

Entre as despesas declaradas no SIOPE, o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB) merece especial atenção. O FUNDEB é um fundo especial, com destinação exclusiva para a ampliação da educação básica, com padrão adequado de qualidade, a todos os cidadãos.

Não obstante, alguns relatórios produzidos pela CGU e pelo Tribunal de Contas da União (TCU) apontaram falhas nos relatórios produzidos pelo SIOPE, que não evitaram os desvios na aplicação dos recursos do FUNDEB em diversos municípios. Estes problemas se encontram mais detalhados no item 3.4 do presente trabalho, após a descrição da contextualização sobre o SIOPE, o FUNDEB e as atribuições específicas da CGU enquanto instância de controle.

Diante desses fatos – a disponibilidade dos dados do SIOPE, até então inexplorados pelo ODP, e a existência concreta de desvios dos recursos do FUNDEB – surgiu o seguinte problema de pesquisa: a partir dos registros no SIOPE, como identificar possíveis indícios de irregularidades nos gastos públicos dos municípios com a educação básica?

Por fim, a justificativa do presente trabalho é a possibilidade de gerar conhecimento de valor estratégico no tema da educação, o qual poderá conferir uma maior qualidade de dados

ao SIOPE, bem como propiciar subsídios complementares aos trabalhos desempenhados pelas instâncias de controle.

1.3 OBJETIVOS GERAIS E ESPECÍFICOS

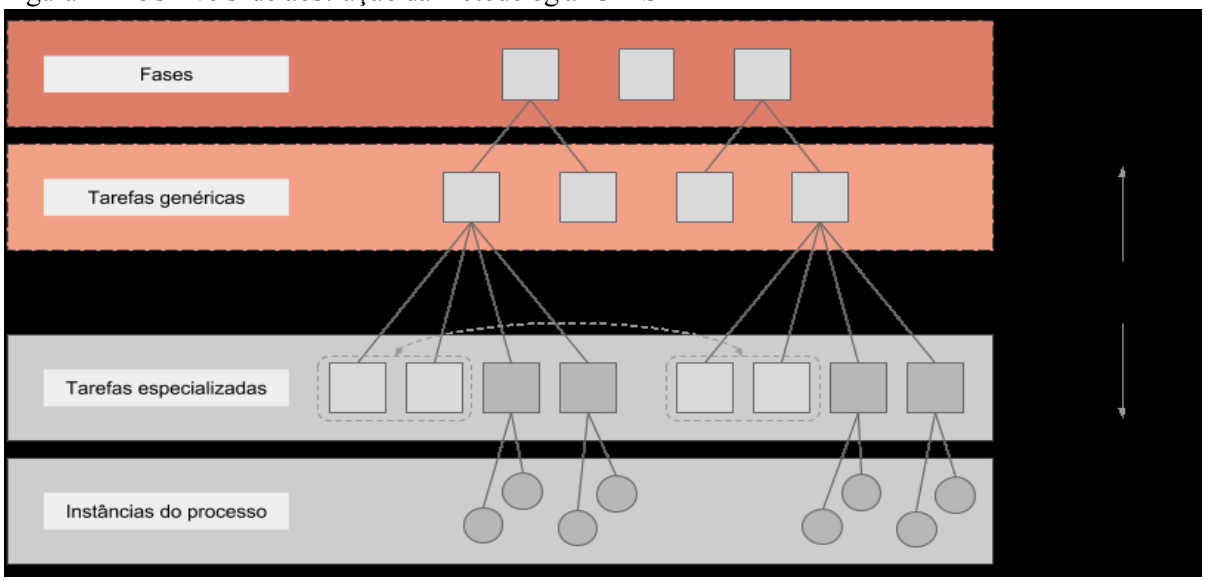
Tendo em vista a oportunidade de explorar os dados do SIOPE, definiu-se o seguinte objetivo geral: "Propor mecanismos para a identificação de discrepâncias nos gastos públicos realizados pelos municípios na educação básica, com o intuito de produzir um relatório com informações de valor agregado para as instâncias de controle", o qual proporciona os objetivos específicos abaixo:

- contextualização acerca da legislação sobre as políticas e os investimentos na educação;
- entendimento sobre a atuação das instâncias de controle;
- entendimento do sistema SIOPE;
- leitura de relatórios de auditoria e fiscalização, para entender as falhas do SIOPE e os desvios na aplicação dos recursos federais em educação básica;
- realizar análises exploratórias nos dados do SIOPE, a fim de levantar hipóteses iniciais sobre os dados;
- escolher as técnicas de mineração de dados a serem utilizadas para a identificação de discrepâncias nos gastos públicos.

2 METODOLOGIA UTILIZADA

O presente trabalho foi realizado seguindo-se a metodologia de referência *Cross-Industry Standard Process for Data Mining* (CRISP-DM), comumente utilizada para projetos de mineração de dados (CHAPMAN, 2000). Conforme demonstrado na Figura 1, o CRISP-DM é um modelo de processo hierárquico, composto de tarefas estruturadas em quatro níveis de abstração, a saber: fases, tarefas genéricas, tarefas específicas e instâncias de processo.

Figura 1 – Os níveis de abstração da metodologia CRISP-DM



Fonte: CHAPMAN et al (2000).

O primeiro nível de abstração (mais geral) se refere ao ciclo de vida de um projeto de mineração de dados, conforme apresentado na Figura 2, composto das seguintes fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. As setas, não rigorosamente sequenciais, indicam as dependências mais importantes e frequentes entre as fases.

A fase de entendimento do negócio – de primordial importância para o sucesso do projeto – consiste em identificar o contexto institucional, entender o problema de negócio que a organização espera resolver e levantar as necessidades mais prioritárias. Em seguida, deve-se determinar os objetivos de negócio e seus critérios de sucesso; verificar os recursos disponíveis em termos de pessoal, dados, hardware e software; determinar os objetivos da mineração de dados a partir do objetivo de negócio (incluindo as técnicas a serem utilizadas, como predição, classificação, agrupamento, entre outras); e, finalmente, criar o plano de projeto.

A fase de entendimento dos dados envolve a coleta, a verificação da qualidade e a exploração dos dados com estatística descritiva, de forma a realizar descobertas e detectar possíveis correlações.

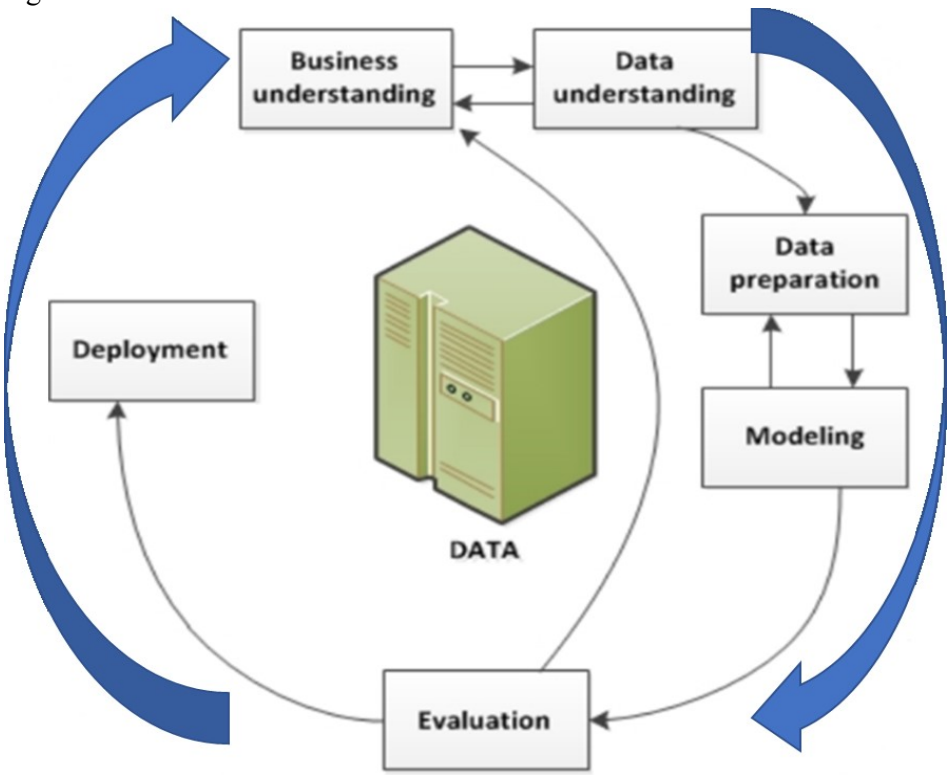
A fase de preparação dos dados realiza tratamentos sobre os dados, a fim de torná-los adequados para a aplicação dos algoritmos de mineração na próxima fase. Tais tratamentos incluem: limpeza dos dados, substituição de valores omissos, seleção de atributos relevantes, redução de dimensionalidade, criação de atributos derivados, integração de dados externos, pivoteamento de dados, entre outros.

A fase de modelagem compreende a seleção e aplicação de técnicas (com ajustes de parâmetros) para criar modelos e encontrar padrões ou descobertas de conhecimento – exemplos dessas técnicas são classificação, agrupamento e regressão. Procedimentos de teste devem ser executados (como, por exemplo, matriz de confusão e medidas de acurácia) para verificar a qualidade e validade dos resultados.

A fase de avaliação objetiva analisar se os resultados do modelo atendem aos objetivos de negócio e aos critérios de sucesso definidos anteriormente, bem como se o modelo pode ser implantado ou se é necessário aprimorá-lo através de novas iterações das fases anteriores.

A fase de implantação resume-se em colocar o modelo obtido em efetiva produção, incluindo, quando aplicável, a confecção de relatório final com os resultados alcançados. Refere-se, assim, à aplicação de modelos de mineração de dados no processo decisório.

Figura 2 – O ciclo de vida do CRISP-DM



Fonte: IBM (2019).

3 FASE DE ENTENDIMENTO DO NEGÓCIO

Esta fase consiste em levantar problemas existentes para, em seguida, estabelecer os objetivos de negócio, seus critérios de sucesso e os objetivos da mineração de dados. Para o presente trabalho, as principais atividades realizadas estão resumidas abaixo:

a) interpretação da legislação sobre as diretrizes e os investimentos na educação nacional, bem como sobre a atuação dos principais órgãos de governo envolvidos (Ministério da Educação, FNDE, CACS e CGU);
b) reunião com o responsável pelo SIOPE no FNDE, que apresentou as principais características do sistema;
c) leitura de relatórios de fiscalização da CGU nos investimentos em educação básica, a fim de identificar problemas relacionados com aplicação dos recursos federais;
d) acesso ao Relatório de Auditoria Anual de Contas no FNDE, elaborado pela CGU, que cita as falhas no monitoramento da aplicação dos recursos pelo FNDE.

Desta forma, faz-se necessário uma prévia contextualização sobre o sistema SIOPE; o FUNDEB como a principal fonte de financiamento federal para a educação básica; e o papel fiscalizador da CGU sobre os recursos federais repassados a estados e municípios.

3.1 O SIOPE PARA MONITORAMENTO DOS GASTOS NA EDUCAÇÃO

O SIOPE, instituído pela Portaria Ministerial MEC nº. 06, de 20 de junho de 2006, e operacionalizado pelo FNDE[1], surgiu como resposta a uma demanda, em 2003, do então Ministro da Educação, Cristovam Buarque, que almejava identificar o quanto se investia na educação pública brasileira. Conforme publicado no sítio do FNDE:

> *"O SIOPE é uma ferramenta eletrônica instituída para coleta, processamento, disseminação e acesso público às informações referentes aos orçamentos de educação da União, dos estados, do Distrito Federal e dos municípios, sem prejuízo das atribuições próprias dos Poderes Legislativos e dos Tribunais de Contas."* (BRASIL, FNDE, 2019).

Esse sistema possibilita que entes federativos registrem, bimestralmente, as receitas e despesas realizadas com a educação pública, incluindo a remuneração de profissionais do magistério e as despesas custeadas com recursos de programas federais relacionados, como o Programa Nacional de Alimentação Escolar (PNAE) e Programa Nacional de Apoio ao Transporte do Escolar (PNATE). Embora os dados sejam de natureza declaratória, há uma série

[1] O Ministério da Educação, por meio da Portaria nº 952/2007, transfere a gestão das atividades operacionais do SIOPE ao FNDE, e este passa a ser o gestor do sistema.

de regras de integridade, embutidas no sistema, que checam os dados lançados[2] antes da transmissão aos CACS[3] e ao FNDE. Por exemplo, o sistema informa se há omissão de alguma receita, proveniente de impostos, que se encontra lançada nos sistemas da STN; ou quando há alunos na Educação Infantil, declarados no Censo Escolar (INEP), sem o registro de despesas nesta mesma modalidade de ensino.

Após a validação dos dados, o sistema indica se cada ente federativo atinge os percentuais legalmente estabelecidos – tendo em vista que o artigo 212 da Constituição Federal (BRASIL, 1988) decreta que os estados e municípios devem aplicar o mínimo de 25% da receita de impostos e transferências no MDE. Ainda, a Lei nº 9.394/96 (BRASIL, 1996) – Lei de Diretrizes e Bases da Educação Nacional (LDB) – especifica, em seu artigo 70, as despesas elegíveis com recursos do MDE:

> *Art. 70. Considerar-se-ão como de manutenção e desenvolvimento do ensino as despesas [...]:*
> *I - remuneração e aperfeiçoamento do pessoal docente e demais profissionais da educação;*
> *II - aquisição, manutenção, construção e conservação de instalações e equipamentos necessários ao ensino;*
> *III – uso e manutenção de bens e serviços vinculados ao ensino;*
> *IV - levantamentos estatísticos, estudos e pesquisas visando precipuamente ao aprimoramento da qualidade e à expansão do ensino;*
> *V - realização de atividades-meio necessárias ao funcionamento dos sistemas de ensino;*
> *VI - concessão de bolsas de estudo a alunos de escolas públicas e privadas; [...]*
> *VIII - aquisição de material didático-escolar e manutenção de programas de transporte escolar.* (BRASIL, 1996)

Há penalidades impostas às unidades que não utilizem o SIOPE. A inserção, atualização e transmissão dos dados aos CACS e ao FNDE são obrigatórias e com prazos definidos[4], sob pena de bloqueio de repasses dos recursos financeiros (transferências voluntárias e de convênios com o Governo Federal) e de situação irregular no Serviço Auxiliar de Informações para Transferências Voluntárias (CAUC).

---

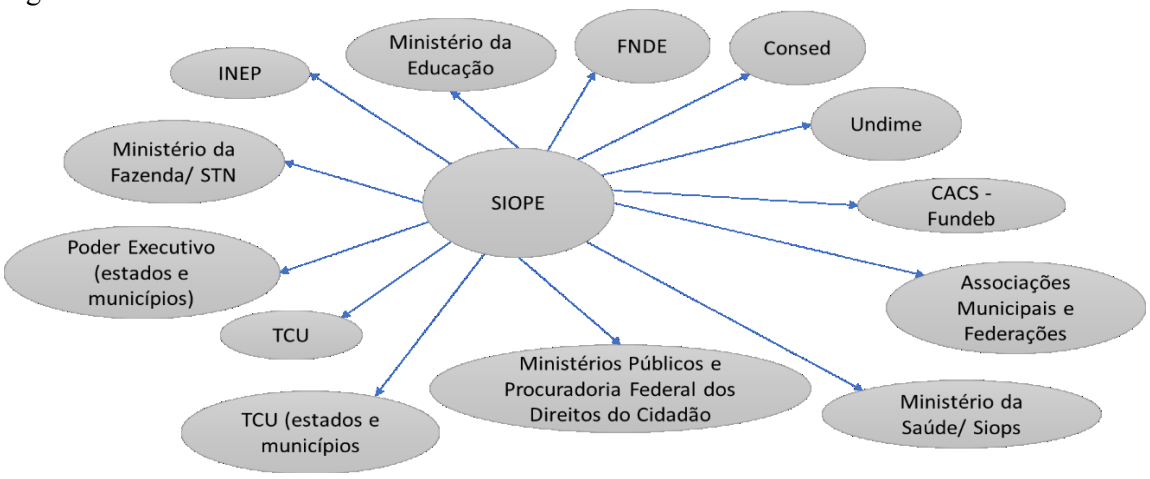
[2] Na versão do SIOPE de 2018, há cerca de 120 críticas, compostas por informações oficiais disponibilizadas pela Secretaria do Tesouro Nacional (STN), pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), pela Agência Nacional do Petróleo (ANP) e pelo FNDE. Os dados lançados devem, portanto, corresponder às informações dos demonstrativos contábeis publicados pelos entes.

[3] O Conselho de Acompanhamento e Controle Social do FUNDEB (CACS), colegiado formado por representações sociais variadas (sindicato de professores, servidores da Educação, alunos e pais de alunos, etc.), se responsabiliza em acompanhar e controlar a distribuição, a transferência e a aplicação dos recursos do Fundo, no âmbito das esferas federal, estadual, distrital e municipal (adaptado de BRASIL, 2007). Para atender a este dispositivo legal, cabe ao CACS validar os dados consolidados no SIOPE, após confirmação pelo Dirigente de Educação.

[4] Embora haja a obrigatoriedade, por parte dos entes, de informar como aplicou as verbas federais – oficialmente falando, o SIOPE não é um sistema de prestação de contas, mas sim o Sistema de Gestão de Prestação de Contas (SigPC), sob gestão do FNDE. Mais informações em: https://www.fnde.gov.br/fnde_sistemas/sigpc-acesso-publico.

Finalmente, o sistema permite a geração periódica de indicadores da gestão educacional, subsidiando a definição, implementação e monitoramento de políticas públicas educacionais por diversos atores, conforme mostra a Figura 3.

Figura 3 – Rede de Parceiros do SIOPE



Fonte: BRASIL, MEC (2018).

De todo modo, o SIOPE é uma importante ferramenta para o acompanhamento e a avaliação dos gastos públicos em educação – tanto pelo Ministério da Educação/FNDE e órgãos de controle[5], como pelos gestores educacionais e CACS. Ademais, as informações registradas, em conjunto com indicadores e relatórios consolidados, são disponibilizadas pelo FNDE em página na internet[6] para acesso pelos cidadãos, assegurando a transparência e o controle social dos recursos públicos destinados à educação.

3.2 DESPESAS COM O FUNDEB

Há no SIOPE o registro de uma importante fonte de recursos: trata-se do FUNDEB. Esse fundo foi criado em 2006 por meio da Emenda Constitucional nº. 53/2006[7], em substituição ao Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (FUNDEF), e encontra-se regulamentado pela Lei nº 11.494/2007 (BRASIL, 2007). Todos os recursos do FUNDEB devem ser aplicados, exclusivamente, na educação básica[8] (particularmente, na valorização do magistério) – esse montante atingiu, conforme a Figura 4, o valor de R$ 150 bilhões em 2018. Sua vigência encerra-se em 2020, mas há projetos de lei em tramitação para torná-lo permanente.

---

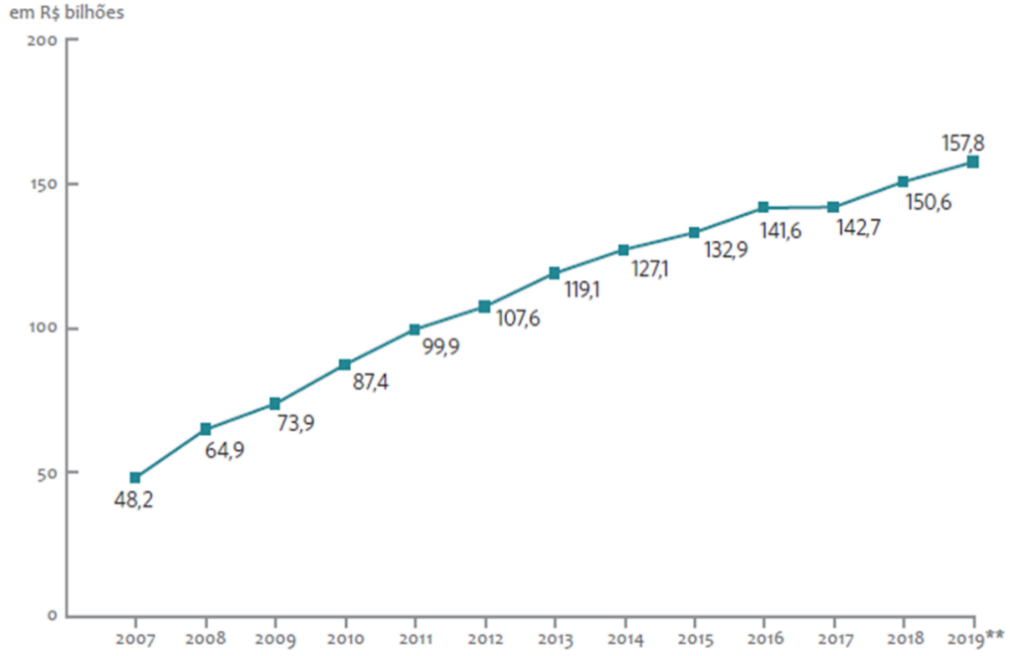
[5] São órgãos de controle: Controladoria-Geral da União, Tribunais de Contas da União, dos estados, Distrito Federal e municípios, os Ministérios Públicos federal, estadual e distrital.

[6] Dados das despesas e planilhas de remuneração dos profissionais do magistério se encontram em https://www.fnde.gov.br/index.php/fnde_sistemas/siope/relatorios/arquivos-dados-analiticos.

[7] Regulamentado pela Lei nº 11.494/2007 e pelo Decreto nº 6.253/2007.

[8] No Brasil, a educação básica compreende a Educação Infantil, o Ensino Fundamental e o Ensino Médio.

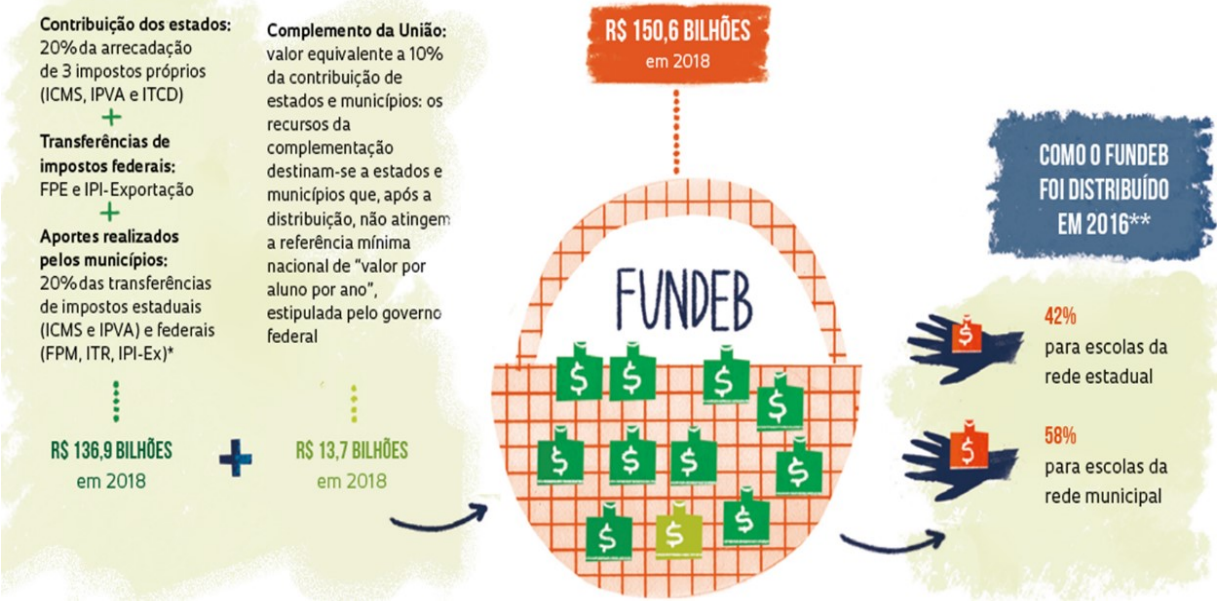
Figura 4 – Evolução dos Investimentos do FUNDEB



Fonte: JEDUCA (2019).

O FUNDEB é um fundo especial e redistributivo, de natureza contábil e de âmbito estadual. Conforme a Figura 5, é formado por 20% dos recursos provenientes dos impostos e transferências dos estados e municípios, e por parcela financeira de recursos federais (a título de complementação). Posteriormente, valores são recalculados (com base no custo por aluno, calculado pelo FNDE, e o número de alunos matriculados, levantado pelo INEP) e redistribuídos aos entes, de forma proporcional e igualitária.

Figura 5 – Composição e redistribuição do FUNDEB



Fonte: QUEIROZ (2020).

Essa redistribuição procura garantir o valor mínimo nacional por aluno/ano a cada ente federativo, contribuindo para diminuir as desigualdades de recursos entre as redes de ensino e universalizar a oferta de ensino a todos os cidadãos.

3.3 O PAPEL DA CONTROLADORIA-GERAL DA UNIÃO

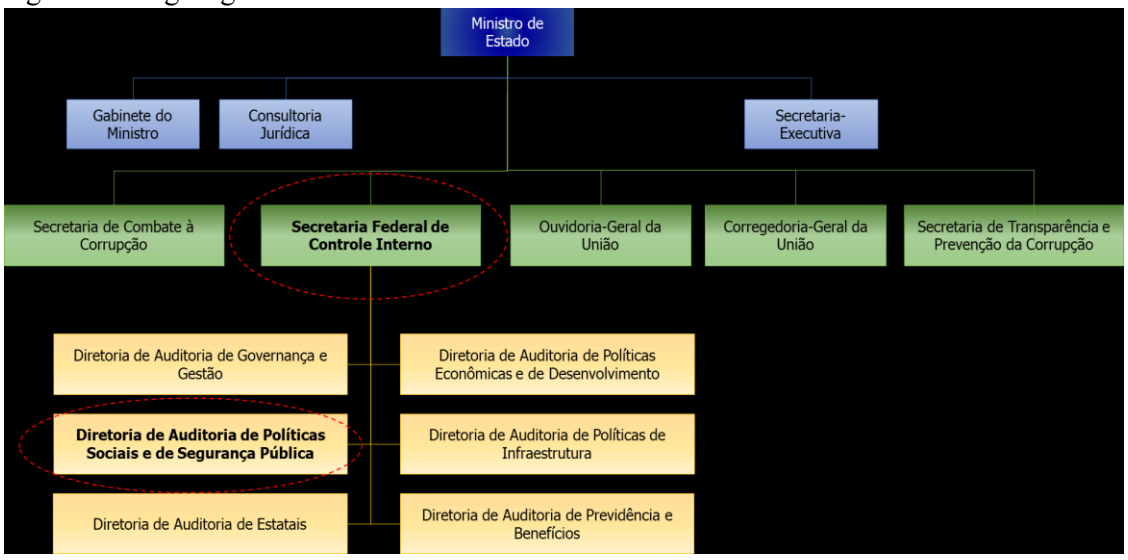
Na esfera federal, a CGU é o órgão central do Sistema de Controle Interno (SCI), do Sistema de Correição e do Sistema de Ouvidoria do Poder Executivo (BRASIL, CGU, 2020), e a ela compete desenvolver funções de defesa do patrimônio público, de controle interno, de auditoria pública, de correição e de ouvidoria, além de ações para a promoção da transparência; para o incentivo à integridade pública; e para a prevenção e combate à corrupção.

O Regimento Interno da CGU (BRASIL, CGU, 2020b) dispõe sobre a sua estrutura organizacional, cabendo à Secretaria Federal de Controle Interno (SFC) exercer as competências de órgão central do SCI do Poder Executivo Federal e, ainda:

> *XIV - fiscalizar e avaliar a execução dos programas de governo [...];*
>
> *XVI - realizar auditorias sobre a gestão dos recursos públicos federais [...];*
>
> *XVIII - apurar atos ou fatos ilegais ou irregulares praticados por agentes públicos ou privados na utilização de recursos públicos federais [...].*

Conforme o organograma na Figura 6, o presente trabalho procurou subsidiar os esforços da SFC/ Diretoria de Auditoria de Políticas Sociais e de Segurança Pública (DS), mais especificamente, da Coordenação-Geral de Auditoria das Áreas de Educação Básica, Direitos Humanos e Desenvolvimento Social (CGEBC), responsável pela Educação Básica.

Figura 6 – Organograma da CGU



Fonte: Elaborada pelo autor (2020), adaptado de BRASIL, CGU (2020b).

Entre as atividades desempenhadas no controle interno, está a Avaliação dos Programas de Governo (AEPG) e a Fiscalização em Entes Federativos (FEF), que verificam a regularidade da aplicação de recursos públicos federais repassados aos estados e municípios. Estes instrumentos passam por constantes melhorias, como a criação da Matriz de Vulnerabilidade (BRASIL, CGU, 2019e) em 2015, uma ferramenta de análise de risco composta por doze indicadores[9] – esta gera um *ranking* de municípios com maiores fragilidades na aplicação dos recursos, e do qual são selecionados aqueles a serem fiscalizados.

Com relação aos recursos federais provenientes do FUNDEB para os municípios, a CGU realiza ações de fiscalização, em campo, nos municípios escolhidos pela Matriz de Vulnerabilidade ou por sorteios públicos. Essas ações resultam em Relatórios de Avaliação, com análises detalhadas sobre atos de execução operacional e financeira do FUNDEB - como, por exemplo, sobre a adequação da folha de pagamentos dos profissionais do magistério ou sobre a correta execução de processos de contratação de fornecimento de bens e serviços destinados às escolas.

Alguns relatórios – em especial, os resultantes da FEF em dois municípios do estado do Piauí[10], produzidos em 2016 (BRASIL, CGU, 2019a e 2019b); e em um município do estado do Maranhão[11], produzido em 2019 (BRASIL, CGU, 2020a) – apontaram para irregularidades na aplicação dos recursos FUNDEB[12], repassados às prefeituras municipais, entre elas:

- Realização de despesas inelegíveis com recursos do FUNDEB, ou despesas incompatíveis com o objetivo do FUNDEB;
- Pagamento de profissionais sem comprovada atuação na docência;
- Realização de gastos sem apresentação de documentos comprobatórios;
- Pagamentos efetuados por serviços não recebidos;
- Superfaturamento por quantidade na aquisição de bens, móveis, equipamentos e materiais permanentes com recursos do FUNDEB;
- Superfaturamento por quantidade na execução de obras de construção, reforma e ampliação de escolas com recursos do FUNDEB.

---

[9] Os indicadores são baseados em informações públicas, extraídas de sistemas do Governo Federal; no histórico de resultados das ações de controle já executadas na região; no histório de análises de Tomada de Contas Especial existentes; em denúncias recebidas; em índices de desenvolvimento econômico e social, entre outros critérios.
[10] Relatório 201602218 – Município de Luzilândia/PI e Relatório 201602219 – Município de Boqueirão do Piauí.
[11] Relatório de Avaliação - Municipio de Mata Roma/MA (FUNDEB) - Exercícios 2017 e 2018.
[12] Recentemente, o G1 - Portal de Notícias (ARCOVERDE e TOLEDO, 2019) divulgou que a CGU identificou uso irregular do FUNDEB em todo o país, no valor de R$ 51 milhões.

Pode-se afirmar que metodologias e instrumentos para o controle interno auxiliam no processo de seleção dos municípios e possibilitam, quando em campo, a detecção das irregularidades existentes. Não obstante, com a disponibilidade das informações de receitas e despesas no SIOPE (no período de 2005 a 2019) e com o uso concomitante de cruzamento de bases de dados, de técnicas estatísticas e de algoritmos de mineração de dados, é possível agregar valor aos métodos existentes – como a Matriz de Vulnerabilidade – no sentido de indicar perspectivas não exploradas, como, por exemplo:

- correlações entre os gastos com educação e índices educacionais do INEP (evolução de matrículas, frequência e taxa de evasão escolar, notas IDEB);
- perfil padrão de gastos por modalidade de ensino e categoria de município;
- indicação de entes federativos com discrepâncias nos seus gastos educacionais declarados, que podem ser falhas ou indícios de irregularidades se não justificadas;
- perfil de remuneração dos profissionais, indicando a média do piso salarial de cada ente federativo.

3.4 IDENTIFICAÇÃO DE PROBLEMAS OU DESAFIOS

O artigo 212 da Constituição Federal (BRASIL, 1988) e a LDB (BRASIL, 1996) fomentam a descentralização da educação básica e a municipalização do ensino fundamental, concedendo autonomia para que estados e municípios organizem e mantenham as instituições oficiais de seus sistemas de ensino[13]. Entretanto, essa descentralização administrativa, em um país de grande extensão territorial, gera um sistema educacional de proporções colossais – mais de 180 mil escolas de educação básica e 48 milhões de matrículas, conforme o Censo Escolar de 2019 (BRASIL, INEP, 2020). Ainda, o substancial volume de recursos financeiros (em bilhões de reais), investido pelo Governo Federal, resulta em um modelo de financiamento de difícil monitoramento e, consequentemente, em uma maior complexidade dos mecanismos de controle interno.

Além da CGU como órgão de controle interno, há ainda a atuação do FNDE, que realiza o monitoramento das aplicações dos recursos do FUNDEB em estados e municípios por meio do SIOPE (BRASIL, MEC, 2018); e a atuação dos CACS, que são incumbidos de desempenhar o acompanhamento e controle social sobre a distribuição, o planejamento e a aplicação dos recursos do FUNDEB (BRASIL, CGU, 2019d), bem como de analisar as contas

[13] Os municípios devem atuar, prioritariamente, na Educação Infantil e no Ensino Fundamental; e estados, nos Ensinos Fundamental e Médio. À União cabe providenciar a assistência técnica e financeira aos entes federativos, entre outras atribuições constitucionais.

informadas no SIOPE pelos entes federativos. Esclarece-se que "monitorar" não tem o mesmo sentido de "fiscalizar", conforme nota emitida pelo MEC:

> *"A fiscalização e o controle quanto à aplicação dos recursos do Fundeb [...] competem aos tribunais de contas locais e ao Ministério Público dos estados, resguardada a competência do Ministério Público Federal, para os estados que recebem o aporte federal de recursos. Ao MEC, por meio do FNDE, compete o monitoramento quanto à aplicação, que é feito por meio do SIOPE [...]. Porém, o Siope é um sistema de monitoramento cuja base é declaratória. Significa que a fiscalização e o controle só são exercidos diretamente para fins de realização de auditoria, inspeção e eventual punição, pelos tribunais de contas locais e pelo Ministério Público."* (ARCOVERDE e TOLEDO, 2019)

Todavia, o Relatório de Auditoria Anual de Contas (AAC), sobre a prestação de contas do FNDE como unidade auditada (BRASIL, CGU, 2019c), apontou para as seguintes constatações sobre o SIOPE como sistema de controle da aplicação dos recursos do FUNDEB:

a) a atuação das diversas instâncias de controle adicionais (MEC, FNDE, CACS) não são suficientes para evitar perdas, desvios e/ou fraudes nas aplicações dos recursos do FUNDEB, detectadas e documentadas em diversos relatórios de auditoria elaborados pela CGU (BRASIL, CGU, 2019a, 2019b, 2020a);
b) a apresentação dos relatórios gerados pelo SIOPE em seu sítio: "*não permitem perceber a situação da educação nos entes, exigindo trabalho especializado para tratamento dos dados, compreensão das informações e criação de parâmetros que dêem significado aos números. É a comparação dos gastos e dos resultados entre entes federativos semelhantes que permite qualificar a atuação do controle*" (BRASIL, CGU, 2019c).
c) os CACS não dispõem das informações gerenciais adequadas para a sua atuação como instância de controle.

Em levantamento realizado pela CGU/CGEBC sobre os Acórdãos da Corte de Contas, verificou-se que o TCU, no Acordão nº 618/2014 – Plenário (BRASIL, TCU, 2020), avaliou o SIOPE quanto à confiabilidade das informações sobre os gastos com MDE e listou as seguintes constatações, dentre outras:

a) impossibilidade de atestar que as informações prestadas pelos entes federados no SIOPE refletem os gastos em MDE, e que a despesa de pessoal é fidedigna;
b) validação de informações antes da transmissão dos dados ocorre somente em itens de receitas – itens relativos às despesas não possuem o mesmo nível de controle.

Diante dos fatos mencionados, os principais problemas identificados são: difícil monitoramento dos investimentos em educação; alta complexidade dos mecanismos de controle

interno; relatórios gerenciais do SIOPE inadequados para uso pelos CACS; e ausência de validação prévia de informações no SIOPE para itens de despesa. Assim, uma das justificativas do presente trabalho é melhorar a qualidade da informação nos relatórios gerenciais do SIOPE, por meio de análises exploratórias nas despesas que extraiam conhecimento de valor agregado à tomada de decisão dos envolvidos com o controle interno.

3.5 TRABALHOS DA CGU/CGEBC EM ANDAMENTO

Há uma demanda específica na CGEBC: aferir a confiabilidade dos dados existentes no SIOPE. Esse trabalho se resume em:

a) criar categorias genéricas de despesas para agrupar determinadas contas contábeis;
b) estabelecer grupos de estados e grupos de municípios semelhantes;
c) em cada grupo, determinar o estado/ município de referência (aquele com o maior valor IDEB), juntamente com intervalos para valores de referência;
d) em cada grupo, identificar os gastos por aluno para cada categoria de despesa;
e) em cada grupo, identificar valores que sejam incompatíveis, ou seja, fora dos limites dos valores de referência; e
f) finalmente, a partir das incompatibilidades, quantificar a confiabilidade do SIOPE.

Pretende-se a criação de uma ferramenta, esboçada na figura abaixo, na qual será possível comparar qualquer estado ou município com seu respectivo ente de referência – com a finalidade subsidiar o CACS do FUNDEB na avaliação das despesas declaradas.

Figura 7 – Elaboração de ferramenta para comparação das despesas dos entes

Município de Interesse:_____ | Município de Referência: (automático) ➡ | Maior Ideb dentre os municípios de mesmo Grupo.

EI	EF	EM	EE	EP	ES	Total
Ideb			Ideb ⬇	Ideb ⬆	Tx. Ev.	

Grupos de Despesas	Va/Al
Remuneração Professores	1500
Remuneração outros Profissionais	900
Formação Profissional	50
Material Didático	150
Alimentação Escolar	200
Transporte Escolar	250
Manutenção	100
Investimento	50
Outros	50
Total	3250

Etapa de Interesse						
<<	<	Ideb	>	>>	Ideb ⬇	Ideb ⬆	Tx. Ev.

Grupos de Despesas	Va/Al	Valores Referência
Remuneração Professores	1600	1300 a 1900
Remuneração outros Profissionais	800	700 a 1250
Formação Profissional	200	0 a 250
Material Didático	250	100 a 350
Alimentação Escolar	250	150 a 400
Transporte Escolar	100	50 a 150
Manutenção	150	50 a 250
Investimento	50	0 a 200
Outros	50	-
Total	3450	

Fonte: Elaborada pela CGU/ CGEBC (2020).

3.6 OBJETIVOS DE NEGÓCIO E DA MINERAÇÃO DE DADOS

Dado todo o contexto até o momento, deu-se prioridade ao problema: "Relatórios gerenciais do SIOPE inadequados para uso pelos CACS" e formulou-se o seguinte objetivo de negócio: realização de técnicas de análise exploratória e de mineração de dados nas despesas vinculadas à educação básica, com o intuito de produzir um relatório com informações de valor agregado para as instâncias de controle. Como consequência, este relatório poderá fornecer subsídios complementares à Matriz de Vulnerabilidade para os trabalhos da CGU, bem como poderá melhorar o acompanhamento da aplicação dos recursos pelo FNDE e CACS.

Entretanto, devido às limitações de tempo e recursos, o mencionado relatório foi restringido para indicar os municípios com discrepâncias nos seus gastos educacionais declarados – ou seja, com despesas inconsistentes em comparação a um determinado padrão, podendo estas serem indícios de falhas ou de irregularidades nos gastos públicos. Como critério de sucesso, espera-se obter ao menos 1% de entes federativos com discrepâncias nos seus gastos educacionais declarados[14], e que alguns entes sejam identificados como anômalos em todos os algoritmos de detecção a serem utilizados.

Com relação à mineração de dados, foram estabelecidos os seguintes objetivos:

a) Realizar análises exploratórias nos dados para gerar novos conhecimentos;
b) Utilizar ao menos três técnicas de clusterização para a criação de grupos de municípios semelhantes;
c) Em um cluster de municípios semelhantes, utilizar ao menos cinco algoritmos de detecção de anomalias nas despesas desses municípios.

3.7 REFORMULAÇÃO DOS OBJETIVOS

Nas primeiras iterações do projeto, foram estabelecidos os objetivos de analisar as despesas do SIOPE Estadual e Municipal (a fim de comparar os gastos municipais com relação à UF), incluindo os salários dos professores. Entretanto, diante da complexidade do trabalho, foi necessário restringir o escopo de análise para as despesas pagas[15] pelos municípios no

[14] Não há alguma razão científica para a escolha do valor de 1% de entes federativos anômalos, apenas o fato de que, em 2020, a CGU sorteou cerca de 1% dos municípios brasileiros para serem fiscalizados (https://www.gov.br/cgu/pt-br/assuntos/noticias/2020/02/cgu-sorteia-60-municipios-para-fiscalizar-em-2020).

[15] O SIOPE apresenta, para cada conta contábil, os seguites valores: Despesas Orçadas, Despesas Empenhadas, Despesas Liquidadas e Despesas Pagas. Decidiu-se utilizar o campo Despesas Pagas para as análises de dados.

Ensino Fundamental[16], no ano de 2018. Ademais, a análise do módulo que trata da remuneração dos profissionais do magistério foi excluída pelas razões abaixo:

a) granularidade da informação (a nível de escola) diferente das demais informações de despesas (a nível de município);
b) ausência de informações (CPF, nome do profissional ou código da escola) e baixa confiabilidade dos dados (remunerações e cargas horárias inválidas) em boa parte dos registros;
c) objetivos de negócio divergentes – enquanto as despesas com educação tem relação com o uso de algoritmos de detecção de anomalias, os dados de remuneração dos profissionais implica em criar tipologias de auditoria (por exemplo: pagamentos indevidos a pessoas falecidas).

4 FASE DE ENTENDIMENTO E PREPARAÇÃO DOS DADOS

Essas fases envolvem a coleta, exploração e tratamentos sobre os dados; e ocorreram simultaneamente ao longo da execução deste trabalho. Ao final do capítulo, tem-se os seguintes *dataframes*[17]: *dataframe* de despesas (dados tabulares) para uso por análises exploratórias, no qual cada registro representa dados de uma despesa; e *dataframe* de municípios para uso pelas técnicas de clusterização, no qual cada registro representa um município com o seu vetor de características. Vale ressaltar que ambos os dataframes contêm apenas as despesas municipais.

4.1 ENTENDIMENTO DOS DADOS DO SIOPE MUNICIPAL

São descritos os componentes existentes no SIOPE para familiarização com o domínio da aplicação: grupos de despesas, programas, subfunções da educação e contas contábeis.

4.1.1 Programas Vinculados

Trata-se de recursos do FNDE repassados aos entes federativos e, naturalmente, este campo é preenchido somente para os registros do grupo de Despesas Vinculadas. No SIOPE Municipal para o Ensino Fundamental em 2018, estão registrados os seguintes programas:

- PNAE, PNATE, PDDE;
- Vinculadas a Contribuição Social do Salário-Educação;
- Outras Transferências de Recursos do FNDE;
- Transferências de Convênios – Educação;
- Outros Recursos Destinados à Educação;
- Ação Judicial FUNDEF – Precatórios.

[16] Inclui as modalidades Ensino de Jovens e Adultos e Educação Especial no contexto do Ensino Fundamental.
[17] *Dataframe* é uma estrutura de dados bidimensional semelhante a uma planilha de dados, com colunas representando atributos e linhas representando registros.

4.1.2 Grupos de Despesas

As despesas são classificadas em grupos, descritos na Tabela 1.

Tabela 1 - Descrição de cada Grupo de Despesa

Despesas próprias custeadas com impostos e transferências	São despesas vinculadas aos recursos próprios de cada ente, ou seja, tem como fonte o Tesouro do Estado, do Distrito Federal ou do Município.
	Não poderão ser consideradas as despesas com convênios, recursos transferidos pelo FNDE, royalties de petróleo e indenizações.
	Algumas despesas[18] devem cumprir o percentual mínimo de 25% no MDE.
Despesas efetuadas com os recursos do FUNDEB	Fundo especial formado por parcela de recursos federais e por recursos provenientes dos impostos e transferências dos entes federativos.
	Devem ser empregados exclusivamente em ações de manutenção e desenvolvimento da Educação Básica pública. No caso de municípios, somente são disponibilizadas as modalidades de ensino Educação Infantil (creche e pré-escola) e Ensino Fundamental.
	Mínimo de 60% destinados à remuneração dos profissionais do magistério.
	Máximo de 40% destinados nas demais ações de MDE.
	Dos 25% das receitas de impostos e transferências destinadas ao MDE, 20% de algumas comporão o FUNDEB.
Despesas custeadas com recursos vinculados	Despesas custeadas com recursos de programas federais de educação como: PNAE, PNATE, PDDE, entre outros.
	Recursos legalmente vinculados a uma finalidade específica.
	Recursos recebidos provenientes de transferências constitucionais, como: Salário-Educação, alimentação (merenda), transporte, Programa Dinheiro Direto na Escola, de transferências voluntárias (convênios) firmados com as unidades federativas, recursos provenientes de royalties do petróleo, etc.
	Não contam como MDE.

Fonte: Elaborada pelo autor (2020).

4.1.3Subfunções da Educação estruturadas em Pastas e SubPastas

A inclusão de despesas no SIOPE atende ao modelo orçamentário brasileiro, utilizando a classificação funcional e programática[19]. Conforme a Portaria nº 42, de 14 de abril de 1999, do Ministério do Planejamento, a função representa o maior nível de agregação das diversas áreas de despesas que competem ao setor público (BRASIL, MPDG, 1999). Diz respeito, portanto, à área de ação do governo (educação, saúde, previdência social, etc.). Por tratar da função educação, a Tabela 2 lista as subfunções da educação[20] cadastradas no sistema.

---

[18] Exemplos dessas despesas são: remuneração do pessoal do magistério, aquisição de material didático e concessão de bolsas de estudos. Demais despesas elegíveis estão detalhadas no item 3.1 (art. 70 da LDB).

[19] http://www.lrf.com.br/mp_op_classificacao_funcional_programatica.html

[20] Esta lista é não exaustiva. O SIOPE cadastra o maior número possível de subfunções (algumas são solicitadas pelos próprios entes, por constarem em seus balanços) e de contas de despesas. Sempre existem as opções "outras", caso no balanço do ente conste determinado item não previsto no Plano de contas contábeis.

Tabela 2 - Subfunções da Função Educação no SIOPE Municipal.

Subfunção típica da educação *modalidade de ensino*	361 - Ensino Fundamental (EF) 362 – Ensino Médio (EM) 363 – Ensino Profissional (Qualif. para o Trabalho) 364 – Ensino Superior (ES) 365 – Educação Infantil (EI) 366 – Educação de Jovens e Adultos (EJA) 367 – Educação Especial (EE)	Pode ser Pasta Pai/ Pasta
Subfunção de apoio administrativo (de infraestrutura)	121 – Planejamento e Orçamento 122 – Administração Geral 123 – Administração Financeira 125 – Normatização e Fiscalização 126 – Tecnologia da Informação 128 – Formação de Recursos Humanos 131 – Comunicação Social	Apenas Pasta
Subfunção Outras *considerada no cálculo do MDE*	331 – Proteção e Benefícios ao Trabalhador (*) 722 – Telecomunicações (Educação a Distância) 782 – Transporte Escolar (*) 841 – Refinanciamento da Dívida Interna (*) 842 – Refinanciamento da Dívida Externa (*) 843 – Serviço da Dívida Interna (*) 844 – Serviço da Dívida Externa (*) 846 – Outros Encargos Especiais (*)	Apenas Pasta
Subfunção Outras *NÃO considerada no cálculo do MDE*	242 – Assistência ao Portador de Deficiência 243 – Assistência à Criança e ao Adolescente 271 – Previdência Básica 272 – Previdência do Regime Estatutário 273 – Previdência Complementar 274 – Previdência Especial 306 – Alimentação e Nutrição - Merenda Escolar (*) 392 – Difusão Cultural 695 - Turismo 812 – Desporto Comunitário 813 – Lazer	Apenas Pasta

(*) São também consideradas subfunções de apoio administrativo e se localizam como subpastas abaixo de alguma modalidade de ensino (Pasta Pai de numeração 361 a 365).

Fonte: Elaborada pelo autor (2020), adaptado de BRASIL, MEC (2018).

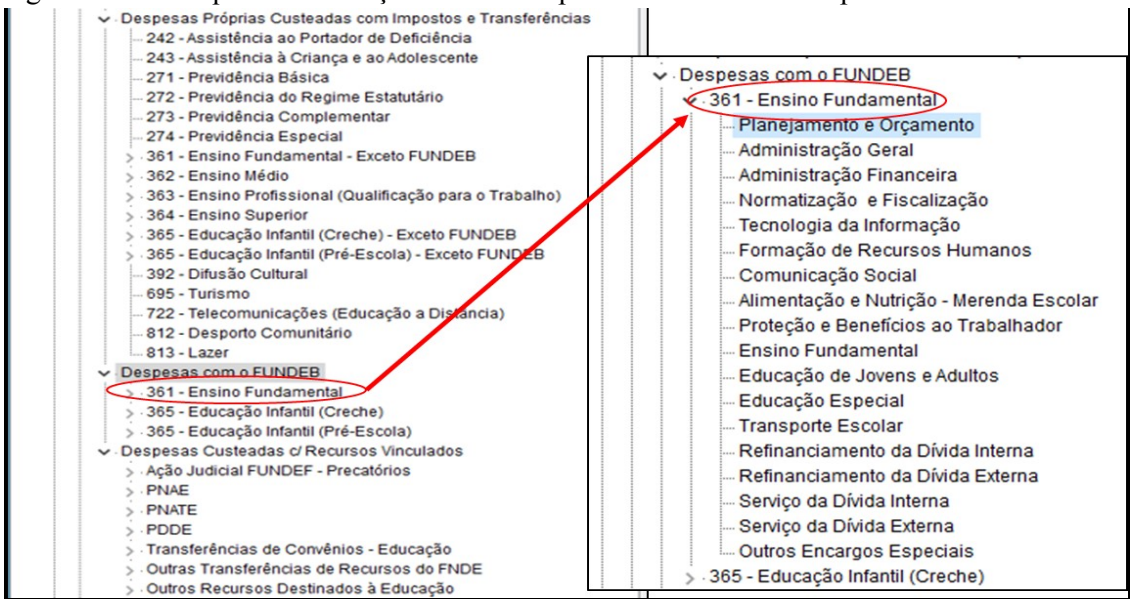
Na alocação de uma dada despesa, após a escolha do grupo de despesa, seleciona-se uma ou mais subfunções que a representem, conforme demonstrado na Figura 8. Por tratar-se de uma hierarquia de subfunções, convencionou-se pelo uso dos campos Pasta_Pai e Pasta[21] - a Pasta_Pai trata-se de uma subfunção típica (a modalidade de ensino) ou do item "Despesas Próprias Custeadas com Impostos e Transferência" – referentes às subfunções 242, 243, etc. que são despesas da escola que independem de uma modalidade de ensino. A Pasta é somente uma outra subfunção, subordinada à Pasta_Pai.

Na Figura 8, abaixo de subfunções típicas de numeração 361 a 365 (modalidades de ensino) estão desdobradas as subfunções de apoio administrativo. Estas são atividades-meio que favorecem o desenvolvimento das atividades escolares e influenciam, indiretamente, para

[21] Esta convenção pode ser encontrada no banco de dados e nos cadernos *jupyter* desenvolvidos para o trabalho.

a execução das subfunções típicas – naturalmente, são despesas rateadas pelo número de matrículas de cada modalidade de ensino, subsidiando os cálculos do custo por aluno e facilitando a apuração, em nível nacional, do quanto se gasta em cada uma destas áreas (BRASIL, MEC, 2018).

Figura 8 - Hierarquia de Subfunções utilizadas para classificar uma despesa



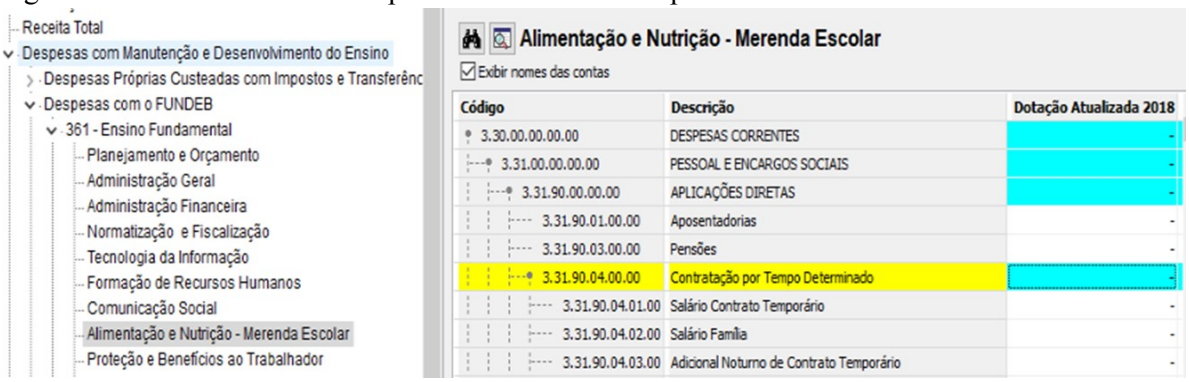
Fonte: BRASIL, MEC (2018).

4.1.4 Contas Contábeis (Natureza e Elemento da Despesa)

O Plano de Contas Contábeis no SIOPE segue as determinações do Manual Técnico de Orçamento (MTO), base para a elaboração dos Orçamentos Fiscal e da Seguridade Social da União (BRASIL, MPDG, 1999). Conforme a Figura 9[22], após a seleção do grupo de despesa e das subfunções (no lado esquerdo), são apresentadas, no lado direito, as contas contábeis hierarquizadas em analíticas (todas as contas analíticas se encontram na cor em branco) e sintéticas (todas as contas sintéticas se encontram na cor em azul, desabilitados para preenchimento). Escolhe-se, enfim, a conta mais relacionada à despesa sendo registrada. A cor amarela é a conta que está, no momento, selecionada pelo usuário na aplicação.

[22] Embora a Figura 9 esteja exibindo apenas a coluna Dotação Atualizada, a aplicação, na verdade, disponibiliza as demais colunas para preenchimento: Despesas Empenhadas, Despesas Liquidadas e Despesas Pagas.

Figura 9 - Uso de conta contábil para classificar uma despesa



Fonte: BRASIL, MEC (2018).

A Figura 10 exibe as contas contábeis sintéticas de maior agregação e um exemplo de várias contas analíticas para uma conta sintética.

Figura 10 – Plano de Contas Contábeis sintéticas e analíticas

Conta Contabil	Descricao_Conta_Contabil
3.30.00.00.00.00	DESPESAS CORRENTES
3.31.00.00.00.00	PESSOAL E ENCARGOS SOCIAIS
3.31.90.00.00.00	APLICAÇÕES DIRETAS
3.31.91.00.00.00	Aplicações Diretas - Operações Intra-Orçamentárias
3.33.00.00.00.00	OUTRAS DESPESAS CORRENTES
3.33.50.00.00.00	Transferências a Inst. Privadas s/ Fins Lucrativos
3.33.90.00.00.00	APLICAÇÕES DIRETAS
3.33.91.00.00.00	Aplicações Diretas - Operações Intra-Orçamentárias
3.40.00.00.00.00	DESPESAS DE CAPITAL
3.44.00.00.00.00	INVESTIMENTOS
3.44.90.00.00.00	APLICAÇÕES DIRETAS
3.45.00.00.00.00	INVERSÕES FINANCEIRAS
3.45.90.00.00.00	APLICAÇÕES DIRETAS
3.45.91.00.00.00	Aplicações Diretas - Operações Intra-Orçamentárias

Conta Contabil	Descricao_Conta_Contabil	Nivel
3.30.00.00.00.00	DESPESAS CORRENTES	0
3.31.00.00.00.00	PESSOAL E ENCARGOS SOCIAIS	1
3.31.90.00.00.00	APLICAÇÕES DIRETAS	2
3.31.90.01.00.00	**Aposentadorias**	**3**
3.31.90.03.00.00	**Pensões**	**3**
3.31.90.04.00.00	**Contratação por Tempo Determinado**	**3**
3.31.90.04.01.00	Salário Contrato Temporário	4
3.31.90.04.02.00	Salário Família	4
3.31.90.04.03.00	Adicional Noturno de Contrato Temporário	4
3.31.90.04.05.00	Adicional de Periculosidade Contrato Temporário	4
3.31.90.04.06.00	Adicional de Insalubridade Contrato Temporário	4
3.31.90.04.07.00	Adicional de Atividade Penosa Contrato Temporário	4
3.31.90.04.10.00	Serviço Extraordinário Contrato Temporário	4
3.31.90.04.12.00	Férias Vencidas/Proporcionais Contrato Temporário	4
3.31.90.04.13.00	13º Salário Contrato Temporário	4
3.31.90.04.14.00	Férias - Abono Constituicional - Contrato Temporário	4
3.31.90.04.15.00	Obrigações Patronais - Contrato por Tempo Determinad	4
3.31.90.04.16.00	Férias Pagamento Antecipado Contrato Temporário	4
3.31.90.04.17.00	Indenização	4
3.31.90.04.99.00	Outras Vantagens Contrato Temporário	4
3.31.90.05.00.00	**Outros Benefícios Previdenciários**	**3**

Fonte: Elaborada pelo autor (2020).

Por exemplo, a conta "Obrigações Patronais", é uma conta de despesa corrente (não contribui para formação de um bem de capital) para gastos de pessoal de contratação por tempo determinado (a fim de atender à necessidade temporária de excepcional interesse público), e de aplicação direta (a unidade orçamentária do ente utiliza diretamente os recursos consignados no orçamento, sem transferi-los a entidades públicas ou privadas).

4.2 PREPARAÇÃO DOS DADOS – *DATAFRAME* DE DESPESAS

4.2.1 Coleta dos dados

No segundo semestre de 2019, dados do SIOPE Estadual e Municipal foram recebidos em formato *dump* e carregados em ambiente com banco de dados *Oracle*. Em seguida, foram copiados para outro servidor com o banco de dados *SQL Server*, através de uma conexão *linked server* (funcionalidade do *SQL Server* que permite acessar dados de um servidor externo).

Posteriormente, as tabelas e os arquivos de criação de objetos foram estudados para a compreensão do modelo de dados e a geração de *scripts*[23] de seleção dos registros. O escopo da seleção se restringiu às tabelas com as despesas dos estados e municípios no ano de 2018, e com os atributos mais relevantes (Tabela 3). Foi necessário realizar o pivoteamento (converter linhas para colunas) no campo Tipo de Despesa, que geraram quatro novos campos (Dotação Atualizada, Despesas Empenhadas, Despesas Liquidadas e Despesas Pagas).

Os critérios de seleção resultaram em cerca de 2 milhões de registros de despesas para o ano de 2018, no âmbito municipal. Os dados selecionados foram inseridos em novas tabelas no *SQL Server,* conforme abaixo:

- DESPESAS_ESTADOS_PROPRIAS_2018
- DESPESAS_ESTADOS_FUNDEB_2018
- DESPESAS_ESTADOS_VINC_2018
- DESPESAS_ESTADOS_OUTRAS_2018
- DESPESAS_MUNIC_PROPRIAS_2018
- DESPESAS_MUNIC_FUNDEB_2018
- DESPESAS_MUNIC_VINC_2018

Estas tabelas foram acessadas no ambiente *Jupyter Notebook* (PROJECT JUPYTER, 2019) e consolidadas em diferentes *dataframes* (um com despesas estaduais; outro, com municipais[24], que é escopo do presente trabalho), por meio do emprego de linguagem *Python*

[23] Para se ter uma ideia da complexidade do modelo de dados do SIOPE Municipal, consultar os scripts de seleção dos dados criados nos APÊNDICE A, APÊNDICE B e APÊNDICE C.
[24] Seleção das tabelas no APÊNDICE D.

(PYTHON, 2001) e de uma diversidade de bibliotecas, entre elas: a biblioteca Pandas para análise de dados (THE PANDAS PROJECT, 2019), a biblioteca PYODBC para conexão ao banco de dados e a biblioteca SEABORN (SEABORN, 2019) para a geração de gráficos. Para uso dos algoritmos de mineração de dados (clusterização), foram utilizados os módulos da biblioteca de aprendizagem de máquina de código aberto SCIKIT-LEARN (PEDREGOSA, 2011).

Tabela 3 - Campos do modelo de dados do SIOPE relevantes para o trabalho

CodUF	Código da UF
NomeUF	Nome da UF
SigUF	Sigla da UF
CodMunicipio	Código do município
NomeMunicipio	Nome do município
Classif_Pasta	Equivale ao Grupo de Despesa (Próprias, FUNDEB, Vinculadas).
NomePrograma	Nome do Programa do FNDE (somente para despesas vinculadas).
CodPasta_Pai	Código da Pasta Pai (representa a modalidade de ensino).
NomePasta_Pai	Nome da Pasta Pai (representa a modalidade de ensino).
CodPasta	Código da Pasta (subfunção da educação).
NomePasta	Nome da Pasta (subfunção da educação).
CodCC	Número da conta contábil, sem pontos. Ex. 34490523400.
Cod_CC_f	Número da conta contábil com pontos que separam grupos de dois dígitos. Ex. 3.44.90.52.34.00
NomeCC	Nome da conta contábil. Ex. Máquinas, Utensílios e Equip. Diversos.
Cod_NomeCC_f	Campo que junta o Nome ao Código da Conta Contábil - necessário porque há contas contábeis diferentes que apresentam o mesmo nome.
Dotação Atualizada (DA)	Dotação prevista no Orçamento (mais as suplementações, menos as anulações registradas).
Desp. Empenhadas (DE)	Despesa originária de ato emanado de autoridade competente que cria para o Estado uma obrigação de pagamento.
Desp. Liquidadas (DL)	Verificação do direito adquirido pelo credor, com base em documentos comprobatórios da entrega do material ou da prestação de serviço.
Desp. Pagas (DP)	Consiste na quitação do bem adquirido ou do serviço contratado.

Fonte: Elaborada pelo autor (2020).

As análises exploratórias foram realizadas em ambos os *dataframes*, mas o presente trabalho, a partir do próximo item, detalha as atividades de preparação dos dados executadas apenas no *dataframe* de despesas municipais.

4.2.2 Limpeza de dados

Poucos procedimentos de limpeza foram necessários. Primeiramente, verificou-se os registros registros que contivessem campos nulos e procedeu-se às ações registradas na

Tabela 4.

Tabela 4 - Procedimentos de limpeza dos dados

DA, DE, DL, DP	Campos nulos foram preenchidos com valor ZERO (se deve a alguns tipos de despesa não preenchidos pelos entes federativos – consequência do pivoteamento na coluna "Tipo de Despesa").
Nome Programa	Este campo se aplica somente para Despesas Vinculadas. Demais registros (despesas próprias e FUNDEB) com valor nulo foram atualizados para "Não se aplica".
CodPasta Pai	Campos nulos foram atualizados para "Não se aplica", pois são referentes às subfunções que não se encaixam em alguma modalidade de ensino. Embora nulo, o campo NomePasta_Pai está preenchido com o valor "Despesas Próprias Custeadas com Impostos e Transferências".

Fonte: Elaborada pelo autor (2020).

4.2.3 Inclusão de colunas: "Classificação" e "Tipo de Gasto"

A coleta de dados trouxe todas as contas contábeis sintéticas (contas agrupadas) e analíticas (contas específicas nas quais as despesas são inseridas pelos usuários). Desta forma, tornou-se necessário indicar a classificação das contas em sintéticas e analíticas, para que o *dataframe* de despesas contivesse somente as analíticas que seriam submetidas à análise de dados. Na Figura 11, são apresentados os quantitativos de contas analíticas e contas sintéticas.

Figura 11 – Quantitativo de registros de contas analíticas e contas sintéticas[25]

```
Lista de Grupos de Despesa:  3 valores
 ['Desp proprias - não EF', 'Desp FUNDEF', 'Desp com Recursos Vinculados']

Contagem de contas analíticas (unique):  476
Contagem de contas sintéticas (unique):  40

Considerando-se todos os grupos de Despesas dos Municípios:
Total de registros de despesas dos Municipios:  2026697 despesas
Total de registros de despesas Analíticas dos Municipios:  923778 despesas
Total de registros de despesas Sintéticas dos Municipios:  1102919 despesas
```

Fonte: Elaborada pelo autor (2020).

Ademais, em virtude de um grande número de contas contábeis (mais de 300 contas), cada conta contábil analítica foi classificada, manualmente, em uma categoria de gasto mais genérico, com o intuito de consolidar as contas em 9 grandes grupos de tipo de gasto: Remuneração, Formação, Didático, Alimentação, Transporte, Manutenção, Investimentos, Conveniadas e Outros. Essa classificação foi realizada pela equipe da CGU/ CGEBC, sendo validada por um gestor do SIOPE no FNDE. Um exemplo do resultado da inclusão desta nova coluna (Tipo de Gasto) é mostrado na

[25] A figura apresenta o termo "Desp proprias – não EF" (descrição literal encontrada na base de dados) como um Grupo de Despesa. Esse grupo representa, na verdade, o grupo de Despesas Próprias cujos gastos não foram aplicados no Ensino Fundamental. Outro termo é "Desp FUNDEF", que representa o próprio FUNDEB (que antes era denominado FUNDEF, conforme explicado no item 3.2).

Figura 12.

Figura 12 – Exemplo de registros com as colunas Classificação e Tipo Gasto

	SigUF	NomeMunicipio	GrupoDespesa	NomePasta	CodCC_f	NomeCC	DE	Classificação	Tipo de Gasto
373596	RJ	Angra dos Reis	Desp proprias - não EF	Educação Especial	3.31.90.13.00.00	Obrigações Patronais	5326.10	Sintético	Não se aplica
487390	RJ	Angra dos Reis	Desp proprias - não EF	Educação Especial	3.31.90.13.01.00	FGTS	1775.40	Analítico	1. Remuneração
430904	RJ	Angra dos Reis	Desp proprias - não EF	Educação Especial	3.31.90.13.99.00	Outras obrigações patronais	3550.70	Analítico	1. Remuneração

Fonte: Elaborada pelo autor (2020).

Com resultado da classificação, A Figura 13 exibe a contagem de despesas analíticas em cada grupo de Tipo de Gasto.

Figura 13 – Contagem de despesas analíticas em cada Tipo de Gasto

index	Total registros	Total de registros em %
6. Manutenção	409755	44.36
1. Remuneração	264333	28.61
7. Investimentos	99349	10.75
4. Alimentação	47067	5.10
5. Transporte	39084	4.23
3. Didático	33816	3.66
9. Outros	18912	2.05
8. Conveniadas	6763	0.73
2. Formação	4699	0.51

Fonte: Elaborada pelo autor (2020).

4.2.4 Inclusão de dados adicionais

As tentativas iniciais de modelagem nas despesas não produziram bons resultados. Na verdade, não faz muito sentido comparar as despesas de todos os municípios sem o devido cuidado com as particularidades de cada município (municípios maiores, com mais e alunos, gastam muito mais e vice-versa). Em vista disso, houve um retorno à fase de preparação de dados e estabeleceu-se novas estratégias – a inclusão de dados externos que caracterizassem os municípios para fins de clusterização. A

Tabela 5 resume os dados externos adicionados ao *dataframe* de despesas.

Tabela 5 – Dados de fontes externas adicionados ao estudo das despesas municipais

Dados do IBGE	Dados complementares de Regiões e Municípios: código do município IBGE, regiões (região, mesoregião e microregião) e população nos 5.570 municípios brasileiros (BRASIL, IBGE, 2019).
Dados do SIOPE e INEP	Dados referentes ao contexto educacional de cada município: 1) quantitativos de matrículas, por modalidade de ensino (subfunções 361 a 367), para cada município (dados do INEP previamente cadastrados no SIOPE); 2) quantitativos de escolas e de professores (dados do INEP); 3) Índice de Desenvolvimento da Educação Básica (IDEB); e 4) Taxa de evasão (2014/2015) para dependência administrativa pública.
Dados do PNUD	Dados referentes ao Índice de Desenvolvimento Humano Municipal (IDHM), para cada município: IDHM Educação, IDHM Longevidade e IDHM Renda.

Fonte: Elaborada pelo autor (2020).

Os quantitativos de matrículas, por modalidade de ensino e para cada município, são dados fornecidos pelo INEP e cadastrados no SIOPE. Com relação aos quantitativos de escolas e professores, as condições de seleção foram obtidas do documento "Filtros da Educação Básica" do INEP, que trata de instruções para a utilização dos Microdados do Censo da Educação Básica - 2018 (BRASIL, INEP, 2019e), e executadas na base de dados do INEP, existente na CGU[26].

O Índice de Desenvolvimento da Educação Básica (IDEB) (BRASIL, INEP, 2019a), elaborado pelo MEC, é um indicador de qualidade dos ensinos fundamental e médio, abrangendo as redes pública e privada, sendo resultado do cruzamento do desempenho (Prova Brasil e Avaliação Nacional da Educação Básica - ANEB) com o rendimento escolar (aprovação). Foram acrescentados ao *dataframe* de despesas os seguintes indicadores:

- IDEB Anos Iniciais (**IDEB_AI**) e IDEB Anos Finais (**IDEB_AF**): se referem às notas IDEB – Ensino Fundamental (EF) - Anos Iniciais[27] e Anos Finais[28] (referência 2017) de cada município (apenas a nota das escolas urbanas da rede municipal); e
- IDEB Ensino Médio (**IDEB_EM**): se refere às notas IDEB - Ensino Médio[29] (referência 2017) de cada município (apenas a nota das escolas da rede estadual).

A taxa de evasão representa a proporção de alunos que, em 2014 estavam matriculados na série k (etapa de ensino seriada do ensino fundamental ou médio), e em 2015 não estavam

[26] Outras informações complementares com relação à obtenção destes dados, como os nomes dos campos, a elaboração dos *scripts* com os critérios de seleção em banco de dados, podem ser visualizados no APÊNDICE E.
[27] Notas IDEB EF – Anos Iniciais acessadas em (BRASIL, INEP, 2019b)
[28] Notas IDEB EF – Anos Finais acessadas em (BRASIL, INEP, 2019c)
[29] Notas IDEB EF – Ensino Medio acessadas em (BRASIL, INEP, 2019d)

matriculados. No momento em que se buscou estes dados, a taxa mais recente disponibilizada era com relação ao ano de 2014 para o ano de 2015.

Com relação ao Índice de Desenvolvimento Humano Municipal (IDHM), gerado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) – trata-se de medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda, para avaliar o desenvolvimento dos municípios brasileiros (BRASIL PNUD, 2019)[30].

- IDHM Educação (**IDHM_E**): média geométrica do subíndice de frequência de crianças e jovens à escola, com peso de 2/3, e do subíndice de escolaridade da população adulta, com peso de 1/3.
- IDHM Longevidade (**IDHM_L**): obtido a partir do indicador Esperança de vida ao nascer (valores mínimo e máximo são 25 e 85 anos, respectivamente).
- IDHM Renda (**IDHM_R**): obtido a partir do indicador Renda per capita (valores mínimo e máximo são 8,00 e 4.033,00)[31].

4.2.5 Filtro dos dados para contexto ao Ensino Fundamental

Manteve-se no *dataframe* de despesas apenas os registros do Ensino Fundamental.

Figura 14 – Quantidade de registros de despesas analíticas em cada modalidade de ensino

	GrupoDespesa	CodPasta_Pai	NomePasta_Pai	Total Registros
0	Desp FUNDEF	361	Ensino Fundamental	131431
3	Desp com Recursos Vinculados	361	Ensino Fundamental	171823
9	Desp proprias - não EF	361	Ensino Fundamental - Exceto FUNDEB	273025
4	Desp com Recursos Vinculados	362	Ensino Médio	7214
10	Desp proprias - não EF	362	Ensino Médio	5155
5	Desp com Recursos Vinculados	363	Ensino Profissional	1668
11	Desp proprias - não EF	363	Ensino Profissional (Qualificação para o Trabalho)	3276
6	Desp com Recursos Vinculados	364	Ensino Superior	3983
12	Desp proprias - não EF	364	Ensino Superior	5705
1	Desp FUNDEF	365	Educação Infantil (Creche)	39364
2	Desp FUNDEF	365	Educação Infantil (Pré-Escola)	45082
7	Desp com Recursos Vinculados	365	Educação Infantil (Creche)	39159
8	Desp com Recursos Vinculados	365	Educação Infantil (Pré-Escola)	38131
13	Desp proprias - não EF	365	Educação Infantil (Creche) - Exceto FUNDEB	74784
14	Desp proprias - não EF	365	Educação Infantil (Pré-Escola) - Exceto FUNDEB	74960
15	Desp proprias - não EF	Não se aplica	Despesas Próprias Custeadas com Impostos e Transferências	9018

Fonte: Elaborada pelo autor (2020).

[30] Os dados brutos do IDHM foram obtidos em http://www.atlasbrasil.org.br/2013/data/rawData/atlas2013_dadosbrutos_pt.xlsx.
[31] Preços de agosto de 2010.

Entende-se que cada modalidade de ensino utiliza subfunções e contas contábeis diferenciadas. É provável que o Ensino Médio, por exemplo, tenha muito mais gastos com a infraestrutura de laboratórios (Física, Química e Biologia) do que o Ensino Fundamental e a Educação Infantil (que não terá esse tipo de gasto). Consequentemente, os estudos devem ser realizados, de forma separada, por modalidade de ensino - e o presente trabalho se limitou aos estudos das despesas aplicadas no Ensino Fundamental (registros de campo CodPasta_Pai igual ao valor 361) – lembrando-se que são incluídas as modalidades Ensino de Jovens e Adultos e Educação Especial no contexto do Ensino Fundamental.

4.2.6 Resumo do *dataframe* de despesas

Após a realização de todos os tratamentos, tem-se um *dataframe* de despesas municipais municipais executadas para o Ensino Fundamental, no ano de 2018 – suas principais características estão resumidas na

Figura 15. Alguns comentários se fazem necessários:

- Há o total de 4.989 municípios, pois nem todos os municípios havia entregue a declaração das contas no momento do recebimento do sistema SIOPE pela CGU;
- Alguns campos inseridos inicialmente no *dataframe* ("Custo da educação por aluno" e "Despesa da educação com Professor, por Aluno") foram descartados por apresentarem inconsistências que prejudicariam a eficácia dos algoritmos de mineração de dados;
- Os valores zero existentes para as variáveis "IDHM dos Municípios", "Número de matrículas no EF", "IDEB – EF Anos Iniciais", "IDEB – EF Anos Finais" e "Custo da educação por Aluno" significam a ausência de dados (não houve valores registrados dessas variáveis para determinados municípios);
- Os valores zero existentes para a variável Taxa de Evasão significam que realmente não houve evasão de alunos.

Figura 15 – Resumo dos atributos do *dataframe* de despesas municipais

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

Total de colunas do dataframe:  50 colunas

Colunas do dataframe de despesas:
 ['CodUF' 'NomeUF' 'SigUF' 'CodMun' 'CodIBGE' 'CodIBGE_Completo'
 'NomeMunicipio' 'Regiao' 'MesoRegiao' 'NomeMesoRegiao' 'MicroRegiao'
 'NomeMicroRegiao' 'NomePrograma' 'GrupoDespesa' 'Tipo de Gasto'
 'CodPasta_Pai' 'NomePasta_Pai' 'CodPasta' 'NomePasta' 'CodCC' 'CodCC_f'
 'NomeCC' 'Cod_NomeCC' 'Cod_NomeCC_f' 'DA' 'DE' 'DL' 'DP' 'Vlr_FUNDEB_STN'
 'Pop_estimada' 'IDHM' 'IDHM_E' 'IDHM_L' 'IDHM_R' 'QtdEscolas'
 'QtdDocentes' 'NUM_MATR_361' 'NUM_MATR_362' 'NUM_MATR_363' 'NUM_MATR_365'
 'NUM_MATR_365_1' 'NUM_MATR_365_2' 'NUM_MATR_366' 'NUM_MATR_367' 'IDEB_AI'
 'IDEB_AF' 'IDEB_EM' 'TxEvasao_EF' 'CustoAluno' 'DespesaProf']


Regiões cujos Municipios entregaram a declaração ao SIOPE:  5 regiões
 ['CENTRO OESTE', 'NORDESTE', 'NORTE', 'SUDESTE', 'SUL']

Estados cujos Municipios entregaram a declaração ao SIOPE:  4989 Municípios em  26  estados.
 ['Acre', 'Alagoas', 'Amapa', 'Amazonas', 'Bahia', 'Ceara', 'Espirito Santo', 'Goias', 'Maranhao',
'Mato Grosso', 'Mato Grosso do Sul', 'Minas Gerais', 'Para', 'Paraiba', 'Parana', 'Pernambuco', 'Pia
ui', 'Rio Grande do Norte', 'Rio Grande do Sul', 'Rio de Janeiro', 'Rondonia', 'Roraima', 'Santa Cat
arina', 'Sao Paulo', 'Sergipe', 'Tocantins']

Lista de Programas:  9 valores
 ['Não se aplica', 'PNAE', 'Vinculadas a Contribuição Social do Salário-Educação', 'PNATE', 'Outras
Transferências de Recursos do FNDE', 'Transferências de Convênios - Educação', 'Outros Recursos Dest
inados à Educação', 'PDDE', 'Ação Judicial FUNDEF - Precatórios']

Lista de Grupos de Despesa:  3 valores
 ['Desp proprias - não EF', 'Desp FUNDEF', 'Desp com Recursos Vinculados']

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

Lista de Pastas:  21 valores
 ['Administração Financeira', 'Administração Geral', 'Alimentação e Nutrição - Merenda Escolar', 'Co
municação Social', 'Despesas Custeadas com Recursos de Royalties de Petróleo e de Indenizações', 'Ed
ucação Especial', 'Educação de Jovens e Adultos', 'Ensino Fundamental', 'Ensino Fundamental - Exceto
FUNDEB', 'Formação de Recursos Humanos', 'Normatização  e Fiscalização', 'Normatização e Fiscalizaçã
o', 'Outros Encargos Especiais', 'Planejamento e Orçamento', 'Proteção e Benefícios ao Trabalhador',
'Refinanciamento da Dívida Externa', 'Refinanciamento da Dívida Interna', 'Serviço da Dívida Extern
a', 'Serviço da Dívida Interna', 'Tecnologia da Informação', 'Transporte Escolar']

Lista de valores de Tipo de Gasto da despesa:  9 valores
 ['1. Remuneração', '2. Formação', '3. Didático', '4. Alimentação', '5. Transporte', '6. Manutençã
o', '7. Investimentos', '8. Conveniadas', '9. Outros']

Variação da população dos Municípios:   781 a 12.252023  milhões
Variação do indice IDHM dos Municípios:   0.0 a  0.862
Variação da quantidade de escolas nos municípios:   1 a 1539
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 224567.0
Despesa da educação com Professor, por Aluno:  1401.95 a 186627.96
```

Fonte: Elaborada pelo autor (2020).

Tabela 6 – Descrição dos campos presentes no dataframe de despesas

CodUF, NomeUF, SigUF	Informações da UF.
CodMun, CodIBGE, CodIBGE_Completo, NomeMunicipio	Informações do Município.
Regiao, MesoRegiao, NomeMesoRegiao, MicroRegiao, NomeMicroRegiao	Informações da Região.
NomePrograma	Nome do Programa do FNDE (somente para despesas vinculadas).
GrupoDespesa	Próprias, FUNDEB ou Vinculadas.
Tipo de Gasto	Classificação da conta contábil em um tipo de gasto mais genérico.
CodPasta_Pai e NomePasta_Pai	Representa a modalidade de ensino.
CodPasta e NomePasta	Representa a subfunção da educação.
CodCC	Número da conta contábil, sem pontos. Ex. 34490523400.
CodCC_f	Número da conta contábil com pontos que separam grupos de dois dígitos. Ex. 3.44.90.52.34.00
NomeCC	Nome da conta contábil. Ex. Máquinas, Utensílios e Equip. Diversos.
Cod_NomeCC	Campo que junta o Nome ao Código da Conta Contábil - necessário porque há contas contábeis diferentes que apresentam o mesmo nome.
Cod_NomeCC_f	Campo que junta o Nome ao Código da Conta Contábil formatado.
DA, DE, DL, DP	Tipo da Despesa (Dotação Atualizada, Despesa Empenhada, Despesa Liquidada e Despesa Paga).
Vlr_FUNDEB_STN	Campo adicionado ao dataframe por solicitação da CGEBC (fora do escopo do presente trabalho).
Pop_estimada	População estimada do município (fonte: IBGE).
IDHM (*)	IDHM (Fonte: PNUD).
IDHM_E (*)	IDHM Educação (Fonte: PNUD).
IDHM_L (*)	IDHM Longevidade (Fonte: PNUD).
IDHM_R (*)	IDHM Renda (Fonte: PNUD).
QtdEscolas	Quantidade de escolas (Fonte: INEP).
QtdDocentes	Quantidade de professores (Fonte: INEP).
NUM_MATR_361 (*)	Número de alunos matriculados no Ensino Fundamental (Fonte: INEP).
NUM_MATR_362	Número de alunos matriculados no Ensino Medio (Fonte: INEP).
NUM_MATR_363	Número de alunos matriculados no Ensino Profissional (Fonte: INEP).
NUM_MATR_365	Número de alunos matriculados no Ensino Superior (Fonte: INEP).
NUM_MATR_365_1	Número de alunos matriculados na Educação Infantil (creche) (Fonte: INEP).
NUM_MATR_365_2	Número de alunos matriculados na Educação Infantil (pré-escola) (Fonte: INEP).
NUM_MATR_366	Número de alunos matriculados na Educação de Jovens e Adultos (Fonte: INEP).
NUM_MATR_367	Número de alunos matriculados na Educação Especial (Fonte: INEP).
IDEB_AI (*)	Nota IDEB no Ensino Fundamental - Anos Iniciais (Fonte: INEP).
IDEB_AF (*)	Nota IDEB no Ensino Fundamental - Anos Finais (Fonte: INEP).
IDEB_EM (*)	Nota IDEB no Ensino Médio (Fonte: INEP).

TxEvasao_EF	Taxa de Evasão no Ensino Fundamental (Fonte: INEP). Valores zero significam que não houve evasão de alunos.
CustoAluno (*)	Valor do Custo por Aluno (cálculo realizado pelo FNDE). Não será utilizado no presente trabalho.
DespesaProf	Valor das Despesas por Professor (cálculo realizado pelo FNDE). Não será utilizado no presente trabalho.

(*) Valores zero significam valores ausentes, não informados.

Fonte: Elaborada pelo autor (2020).

4.3 PREPARAÇÃO DOS DADOS – *DATAFRAME* DE MUNICÍPIOS

A execução de determinados algoritmos de mineração de dados pressupõe um formato de dados de entrada no qual cada objeto seja representado por um vetor de atributos (ou vetor de características). Em virtude disso, outras preparações foram realizadas nos dados.

4.3.1 Pivoteamento dos dados

Para a geração de vetores de características, foi realizado um pivoteamento nos dados do *dataframe* de despesas – o resultado é um novo *dataframe* de municípios (com um maior número de colunas), onde cada linha é um município e as colunas são todos os seus atributos.

Tabela 7 – Criação de novas colunas após o pivoteamento dos dados

Grupo Despesa	Despesas Próprias, Despesas FUNDEB, Despesas Vinculadas.
Tipo de Gasto	tgRemun (Remuneração), tgFormacao (Formação), tgDidatico (Material Didático), tgAlim (Alimentação), tgTransp (Transporte), tgManut (Manutenção), tgInvest (Investimentos), tgConv (Conveniadas), tgOutros (Outros)
Nome do Programa	Ação Judicial FUNDEF – Precatórios, Outros Recursos Destinados à Educação, Outras Transf Recursos do FNDE, Transf Convênios – Educação, PDDE, PNAE, PNATE, Vincul a Contrib Social do Salário-Educação.
Nome Pasta (subfunção)	Administração Financeira, Administração Geral, Planejamento e Orçamento, Alimentação e Nutrição - Merenda Escolar, Transporte Escolar, Comunicação Social, Tecnologia da Informação, Despesas Custeadas com Recursos de Royalties de Petróleo e de Indenizações, Educação Especial, Educação de Jovens e Adultos, Ensino Fundamental, Ensino Fundamental - Exceto FUNDEB, Formação Recursos Humanos, Normatização e Fiscalização, Outros Encargos Especiais, Proteção e Benefícios ao Trabalhador, Refinanciamento da Dívida Externa, Refinanciamento da Dívida Interna, Serviço da Dívida Externa, Serviço da Dívida Interna
Conta Contábil	de 3.31.90.01.00.00 a 3.46.00.00.00.00 (mais de 300 contas contábeis analíticas)

Fonte: Elaborada pelo autor (2020).

4.3.2 Consolidação de Contas Contábeis

O *dataframe* de municípios resultante do pivoteamento apresentou um grande número de colunas (387), principalmente de contas contábeis. As análises exploratórias de dados, realizadas inicialmente, comprovou um alto número de valores zerados para muitas dessas contas, o que motivou a substituição de algumas contas analíticas pela conta sintética equivalente, pois não se pretendeu analisar as contas contábeis em um nível muito detalhado.

Em outras palavras, basta identificar os municípios que possuem anomalias em gastos relacionados com, por exemplo, Obrigações Patronais (composta de 14 contas analíticas, conforme listado na Figura 16), e não em cada uma dessas 14 contas.

Figura 16 – Exemplo de uma conta contábil sintética a ser considerada no *dataframe*

Conta Contabil	Descricao_Conta_Contabil	Niveis	Analit/ Sint
3.31.90.13.00.00	**Obrigações Patronais**	**3**	**Sintético**
3.31.90.13.01.00	FGTS	4	Analítico
3.31.90.13.02.00	Contribuições Previdenciárias - INSS	4	Analítico
3.31.90.13.03.00	Contribuições Previdenciárias no Exterior	4	Analítico
3.31.90.13.04.00	Contribuição de Salário-Educação	4	Analítico
3.31.90.13.08.00	Plano de Seguridade Social do Servidor - Pessoal Ativo	4	Analítico
3.31.90.13.09.00	Seguros de Acidentes do Trabalho	4	Analítico
3.31.90.13.11.00	FGTS - PDV	4	Analítico
3.31.90.13.13.00	Sesi/Sesc Ativo Civil	4	Analítico
3.31.90.13.14.00	Multas Indedutíveis	4	Analítico
3.31.90.13.15.00	Multas Dedutíveis	4	Analítico
3.31.90.13.17.00	Juros	4	Analítico
3.31.90.13.18.00	Contribuição para o PIS/PASEP S/Folha Pagto	4	Analítico
3.31.90.13.40.00	Encargos de Pessoal Requisitado de Outros Entes	4	Analítico
3.31.90.13.99.00	Outras Obrigações Patronais	4	Analítico

Fonte: Elaborada pelo autor (2020).

A lista abaixo apresenta as contas contábeis nas quais se fez a substituição das contas analíticas pela sintética equivalente, gerando um novo *dataframe* com menos de 200 colunas.

- Contas (3.31.90.04.**) em Contratação por Tempo Determ. (3.31.90.04.00.00)
- Contas (3.31.90.11.**) em Vencimentos e Vantagens Fixas - Pessoal Civil (3.31.90.11.00.00)
- Contas (3.31.90.13.**) em Obrigações Patronais (3.31.90.13.00.00)
- Contas (3.33.50.43.**) em Subvenções Sociais (3.33.50.43.00.00)
- Contas (3.33.90.30.**) em Material de Consumo (3.33.90.30.00.00)
- Contas (3.33.90.36.**) em Outros Serviços de Terceiros - PF (3.33.90.36.00.00)
- Contas (3.33.90.39.**) em Serviços de Terceiros - PJ (3.33.90.39.00.00)
- Contas (3.33.90.47.**) em Obrigações Tribut. e Contributivas (3.33.90.47.00.00)
- Contas (3.44.90.51.**) em Obras e Instalações (3.44.90.51.00.00)
- Contas (3.44.90.52.**) em Equipam. e Material Permanente (3.44.90.52.00.00)

4.3.3 Resumo do *dataframe* de municípios

Após a realização do pivoteamento e da consolidação de algumas contas contábeis, tem-se um *dataframe* de municípios com suas características: informações do IBGE (UF, região, população total); informações do INEP (quantidade de matrículas, escolas e professores; notas IDEB; taxa de evasão); informações do PNUD (IDHM); e informações de despesas, tipos de gastos, subfunções e contas contábeis.

Figura 17 – Resumo dos atributos do *dataframe* de municípios

```
Total de registros do dataframe df_consol:  4989 registros
Total de Municipios:  4989 Municípios

Total de colunas do dataframe df_consol:  175 colunas

Colunas do dataframe consolidado  :
 ['CodUF' 'NomeUF' 'SigUF' 'CodMun' 'CodIBGE' 'CodIBGE_Completo'
 'NomeMunicipio' 'Regiao' 'MesoRegiao' 'NomeMesoRegiao' 'MicroRegiao'
 'NomeMicroRegiao' 'Pop_estimada' 'IDHM' 'IDHM_E' 'IDHM_L' 'IDHM_R'
 'QtdEscolas' 'QtdDocentes' 'NUM_MATR_361' 'IDEB_AI' 'IDEB_AF'
 'TxEvasao_EF' 'CustoAluno' 'DespesaProf' 'DespFUNDEB' 'DespVinc'
 'DespProp' 'tgRemun' 'tgFormacao' 'tgDidatico' 'tgAlim' 'tgTransp'
 'tgManut' 'tgInvest' 'tgConv' 'tgOutros'
 'Ação Judicial FUNDEF - Precatórios' 'Outras Transf Recursos do FNDE'
 'Outros Recursos Destinados à Educação' 'PDDE' 'PNAE' 'PNATE'
 'Transf Convênios - Educação'
 'Vincul a Contrib Social do Salário-Educação' 'AdmFinanc' 'AdmGeral'
 'MerEscolar' 'ComunSocial' 'DespCusteadasRecRoyPetrIndeniz'
 'EducEspecial' 'EducJA' 'EnsFund' 'EnsFund_exc' 'FormRH' 'NormatFisc1'
 'NormatFisc2' 'OutrosEE' 'PlanOrc' 'ProtBenefTrab' 'RefinDivExt'
 'RefinDivInt' 'ServDivExt' 'ServDivInt' 'TI' 'TranspEsc'
 '3.31.90.01.00.00 - Aposentadorias' '3.31.90.03.00.00 - Pensões'
 '3.31.90.05.00.00 - Outros Benefícios Previdenciários'
 '3.31.90.07.00.00 - Contribuição a Entidades Fechadas de Previdência'] ... e demais contas contábeis.
```

Fonte: Elaborada pelo autor (2020).

5 FASE DE MODELAGEM

Essa fase compreende a seleção e aplicação de técnicas para criar modelos e descobrir conhecimentos. No presente trabalho, Análises Exploratórias de Dados (AED)[32] foram realizadas para a descoberta de fatos relevantes. Em seguida, foram utilizadas as técnicas de Clusterização (para criar agrupamentos de municípios semelhantes) e Detecção de Anomalias (para identificar, em um determinado cluster, as despesas anômalas).

[32] Análises exploratórias são geralmente consideradas como parte da Fase de Entendimento dos Dados. Para este trabalho, julgou-se mais apropriado tratá-las como parte da Fase de Modelagem, por duas razões: a) por terem sido executadas após extensa preparação dos dados; e b) por também terem orientado a escolha do uso de clusterização antes da detecção de anomalias.

5.1 ANÁLISE EXPLORATÓRIA DE DADOS

A AED objetiva resumir e visualizar os dados antes de se criar modelos, permitindo entender as suas propriedades, inspecionar as suas características qualitativas e descobrir novos padrões (LEEK, 2015). No presente trabalho, procedeu-se à AED com estatísticas descritivas e plotagem de gráficos, a fim de conhecer os dados e sintetizar suas características mais relevantes; de detectar padrões ocultos; e de identificar correlações entre as variáveis.

No *dataframe* de despesas, a intenção foi explorar extensivamente as despesas pagas dos municípios com o Ensino Fundamental, no ano de 2018, de forma a:

- ter uma noção de ordem de grandeza dos montantes das despesas com a educação;
- detectar comportamentos dos totais de despesas pagas por grupo de despesa, subfunção, tipo de gasto e por contas contábeis (apenas as vinte maiores); e
- preliminarmente, verificar a existência de despesas de valores anormais.

No *dataframe* de municípios, o objetivo foi o estudo mais específico dos comportamentos das outras variáveis, não somente do total das despesas pagas. Como exemplos destes estudos, pode-se citar:

- Como é a distribuição da população, das quantidades de escolas e professores? E dos indicadores do IDHM e do IDEB?
- Como se comporta cada grupo de despesa?
- Quais as correlações existentes entre as variáveis?
- Toda a AED realizada em ambos os *dataframes* se encontra em cadernos *jupyter*[33], que poderão ser consultados, caso necessário[34]. Por serem extensos, no presente trabalho são apresentadas apenas as principais constatações.

[33] Caderno "AnaliseExploratoria_MDE_EF".
[34] Todos os cadernos podem ser consultados em: https://github.com/tatuchag/AnomaliasEducacao.

5.2 ANÁLISES EXPLORATÓRIAS – *DATAFRAME* DE DESPESAS

5.2.1 Estatísticas e distribuição dos valores das despesas

Inicialmente, foram analisados todos os tipos de despesas, a saber: Dotação Atualizada (DA), Despesas Empenhadas (DE), Despesas Liquidadas (DL) e Despesas Pagas (DP).

Figura 18 – As estatísticas de todos os tipos de despesas

```
Total de registros de despesas EF:  576279

Soma dos valores das despesas EF orçadas (2018):  125.8  bilhões
Soma dos valores das despesas EF empenhadas (2018):  113.18  bilhões
Soma dos valores das despesas EF liquidadas (2018):  109.86  bilhões
Soma dos valores das despesas EF pagas (2018):  105.16  bilhões
```

	DA	DE	DL	DP
count	576279.00	576279.00	576279.00	576279.00
mean	218300.77	196396.24	190639.97	182484.08
std	5263645.55	4375759.86	4334616.45	4176964.69
min	-942000.00	-5220.20	-5220.20	-7217.20
25%	0.00	0.00	0.00	0.00
50%	2000.00	2803.20	2561.30	2265.60
75%	33812.65	34309.60	32047.70	29699.50
max	1841619334.70	1841599096.40	1841599096.40	1841599096.40

```
Média DA:  218300.8 ;  Mediana DA:  2000.0
Média DE:  196396.2 ;  Mediana DE:  2803.2
Média DL:  190640.0 ;  Mediana DL:  2561.3
Média DP:  182484.1 ;  Mediana DP:  2265.6

Maximo DA:  1841.62  milhões
Maximo DE:  1841.6  milhões
Maximo DL:  1841.6  milhões
Maximo DP:  1841.6  milhões
```

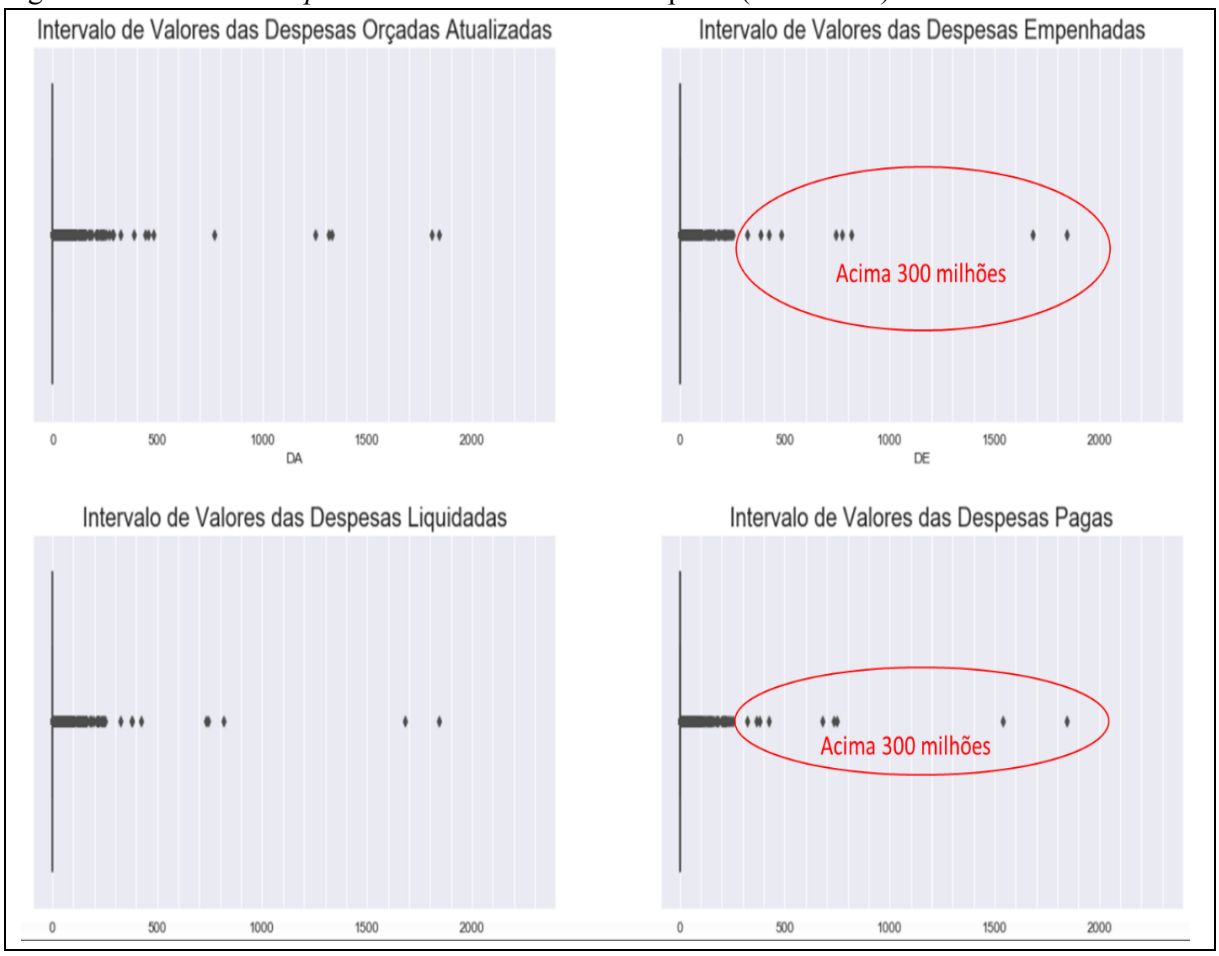
Fonte: Elaborada pelo autor (2020).

Em uma primeira análise, as estatísticas mostraram:

- todos os tipos de despesas apresentam o mesmo comportamento – ou seja, as mesmas grandezas nas médias, medianas, desvios padrão e máximos;
- um alto percentual de valores zero (mais de 25%) em todos os tipos de despesas;
- médias elevadas (180 a 220 mil) e medianas muito baixas (2 a 2,8 mil) indicam a presença de valores muito elevados que aumentam demais o valor da média;
- o desvio padrão elevado (na grandeza de milhões) indica que há muitos dados espalhados e afastados da média.

As medidas estatísticas (média, mediana e desvio padrão) não informam muito sobre a distribuição das despesas. Desta forma, gráficos *boxplots* e histogramas das despesas em diferentes medidas (em valores brutos e em valores log) foram criados, nas figuras a seguir, para uma melhor visualização sobre a distribuição dos valores.

Figura 19 – Gráficos *Boxplots* com os intervalos das despesas (em milhões)



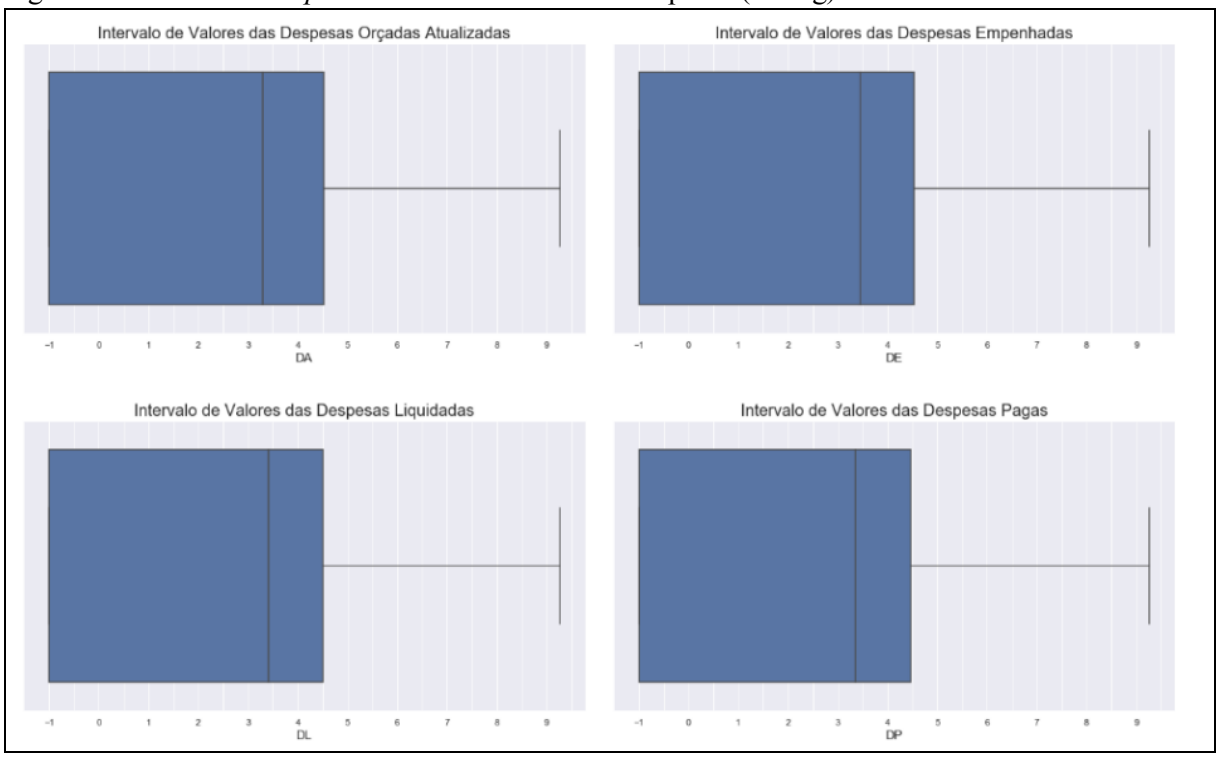
Fonte: Elaborada pelo autor (2020).

Na Figura 19, nota-se que todos os tipos de despesas apresentam, novamente, o mesmo comportamento: concentração de valores no intervalo de 0 a 300 milhões, e alguns poucos valores acima de 300 milhões (cerca de 9 a 11 registros, dependendo do tipo de despesa). Estes poucos pontos representam *outliers* realmente extremos. Como há despesas de valor muito elevado e os valores dos quartis são muito baixos, não é possível visualizar o formato do diagrama de caixa, e todos os pontos acima do terceiro quartil são considerados valores discrepantes (*outliers*).

Os gráficos *boxplots* das despesas com valores em notação logarítmica, na Figura 20, minimizam os efeitos dos *outliers* e possibilitam a detecção de padrões dos gastos de valores menores. Mais uma vez, os comportamentos são similares para todos os tipos de despesa, com uma boa parcela de valores zerados[35] e uma maior concentração no intervalo de 0 a $10^{4,5}$ - ou seja, 75% dos registros variam no intervalo de zero a 30 mil.

[35] Valores zero são representados pelos valores negativos, pois log10(0,1) ~= -1.

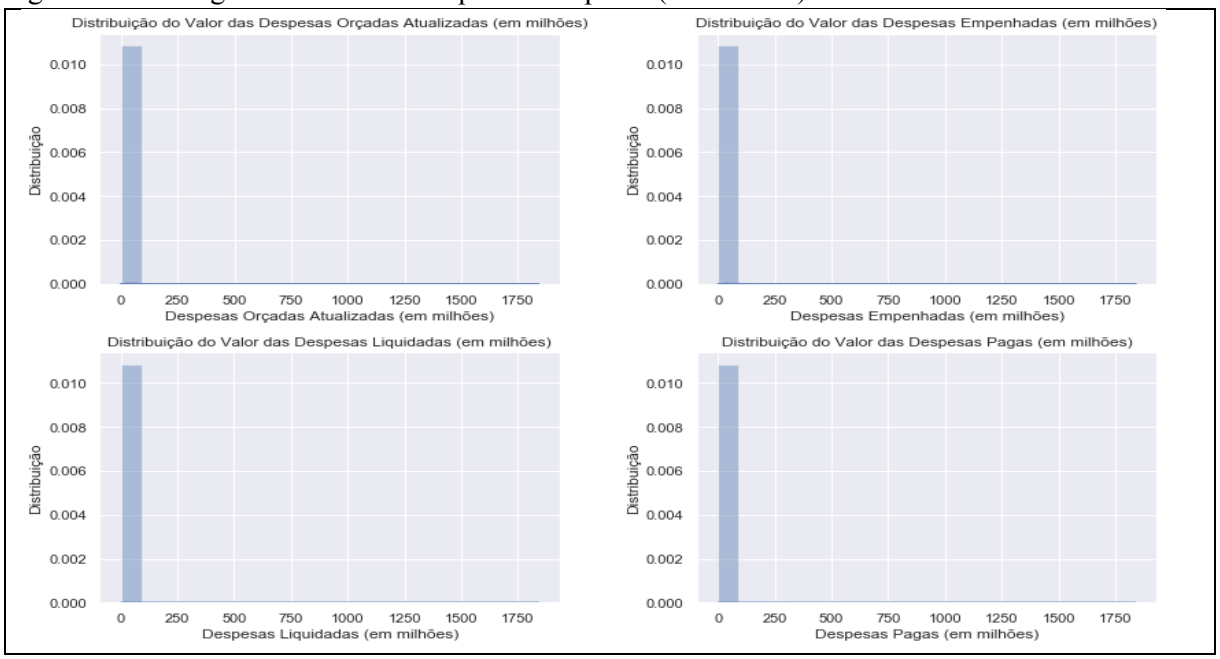
Figura 20 – Gráficos *Boxplots* com os intervalos das despesas (em log)



Fonte: Elaborada pelo autor (2020).

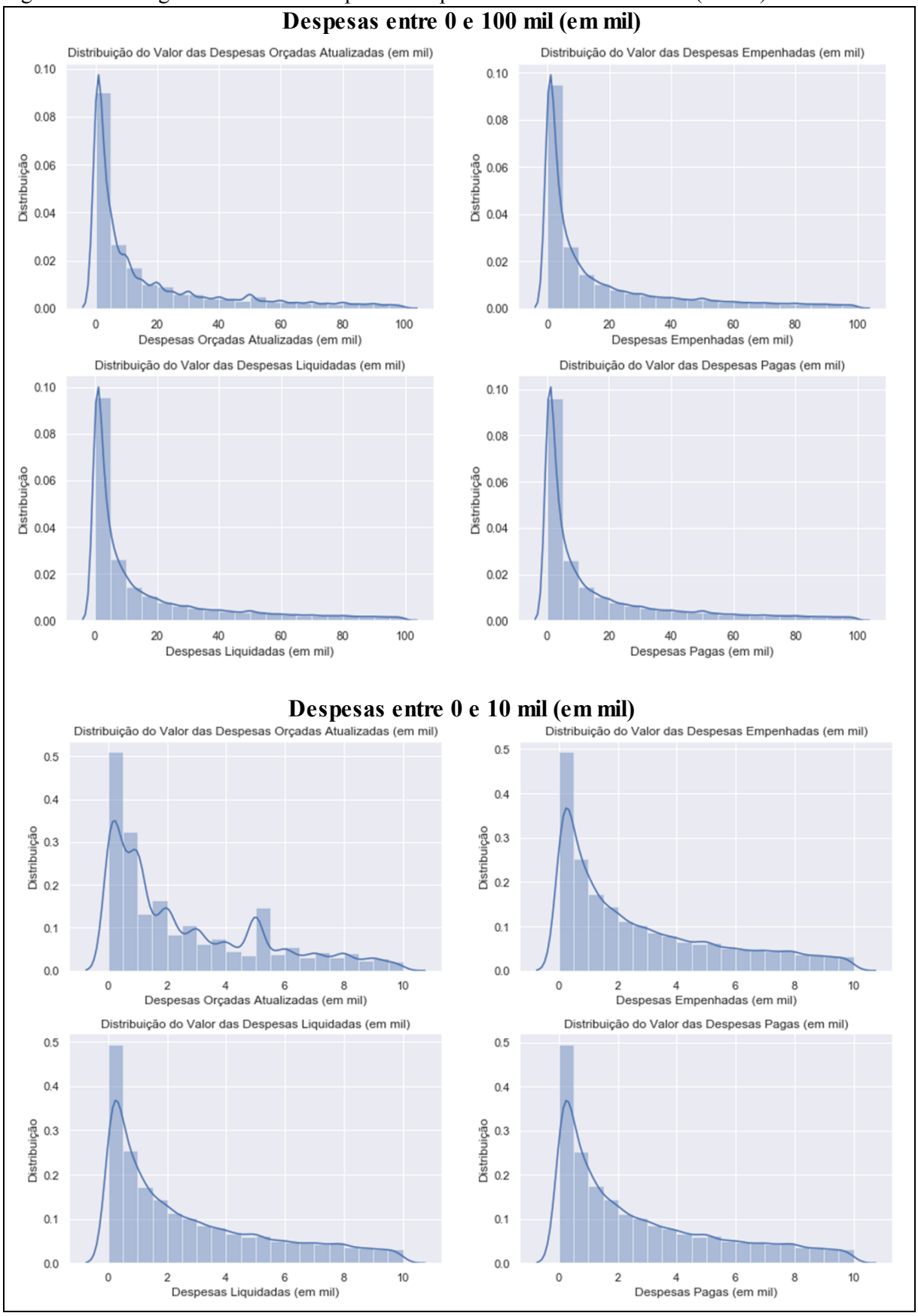
Os histogramas na Figura 21 confirmam a grande variância nos dados, com alta concentração de valores próximos de zero e poucos valores muito altos. Mostram claramente a grande amplitude e a falta de simetria entre os valores de despesas, com dados mais concentrados no lado esquerdo dos gráficos. A Figura 22 apresenta outros histogramas para diferentes intervalos de valores de despesas.

Figura 21 – Histogramas de todos os tipos de despesas (em milhões)



Fonte: Elaborada pelo autor (2020).

Figura 22 – Histogramas de todos os tipos de despesas – diferentes intervalos (em mil)



Fonte: Elaborada pelo autor (2020).

A Figura 23 apresenta histogramas das despesas normalizadas em log, sendo considerados todos os valores (gráficos à esquerda) e valores filtrados para acima de R$ 10 (gráficos à direita). O intuito foi demostrar que é possível obter uma curva que se aproxima de uma distribuição normal quando os valores são normalizados e acima de um determinado valor.

Figura 23 – Histogramas de todos os tipos de despesas - em log e acima de valor R$ 10

Fonte: Elaborada pelo autor (2020).

As estatísticas e os gráficos de distribuição dos valores de despesas permitem identificar os *outliers* de grandes valores - as anomalias globais dentro do conjunto de todas as despesas. Entretanto, é necessário o uso de técnicas alternativas para a identificação das discrepâncias de pequenos valores – ou seja, as anomalias locais.

Em virtude de os tipos de despesas apresentarem comportamentos semelhantes e, conforme interesse da área de negócio, procedeu-se à escolha das DP para análises posteriores, detalhadas nos próximos itens.

5.2.2 Agrupamento das despesas pagas (DP)

Resumidamente, do total do valor de R$ 113 bilhões empenhados com o Ensino Fundamental em 2018, são R$105 bilhões os gastos efetivamente pagos pelos municípios[36].

Figura 24 – Resumo dos tipos de despesas

```
Total de registros de despesas EF:  576279

Soma dos valores das despesas EF orçadas (2018):  125.8  bilhões
Soma dos valores das despesas EF empenhadas (2018):  113.18  bilhões
Soma dos valores das despesas EF liquidadas (2018):  109.86  bilhões
Soma dos valores das despesas EF pagas (2018):  105.16  bilhões
```

Fonte: Elaborada pelo autor (2020).

Nota-se, na Figura 25 que, embora haja uma quantidade bem maior de registros de despesas próprias (47%) do que em despesas FUNDEB (23%), os gastos com o FUNDEB (R$58,4 bilhões) ultrapassam os gastos com despesas próprias (R$35,2 bilhões). O montante das despesas com recursos vinculados (R$11,5 bilhões), por outro lado, representa um percentual baixo de participação no total nacional (apenas 11%).

Figura 25 – Despesas pagas agrupadas por grupo de despesa

```
Total de registros de despesas analíticas:  576279
Soma dos valores das despesas pagas (2018):  105161.74  milhões

Soma dos valores por Grupo de Despesa (em milhões):
```

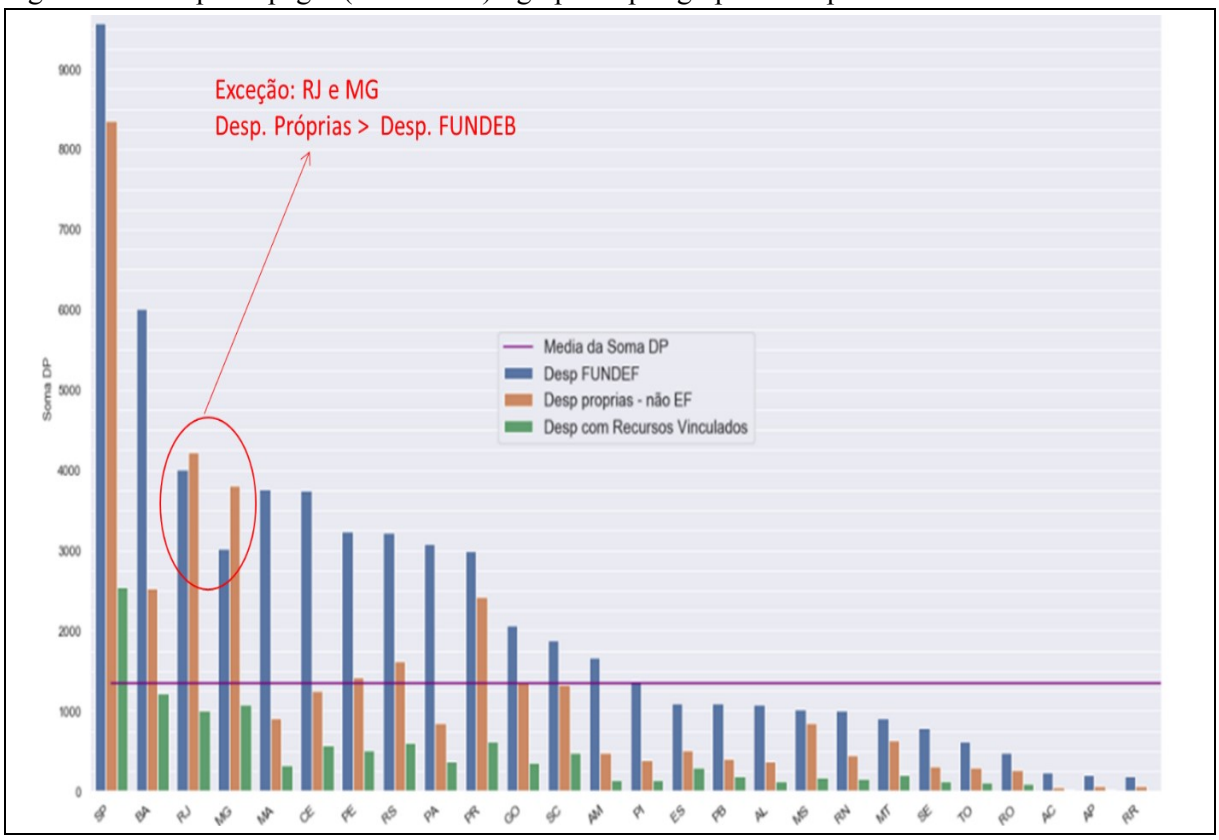
	index	Total registros	Total em %	Soma DP	Soma em %
0	Desp FUNDEF	131431	22.81	58436.56	55.57
1	Desp com Recursos Vinculados	171823	29.82	11518.64	10.95
2	Desp proprias - não EF	273025	47.38	35206.55	33.48

Fonte: Elaborada pelo autor (2020).

Ao se agrupar os montantes das DP dos municípios por UF[37], percebe-se, conforme a Figura 26, que todos os estados apresentam uma menor participação de recursos vinculados no total das despesas, e que gastos com FUNDEB são maiores que os gastos em Despesas Próprias (exceto nos estados do RJ e MG). Ainda, o perfil de gasto do estado de ES que, embora seja da Região SUL (SP, RJ e MG com os maiores montantes em DP), apresenta comportamento similar com estados da região Nordeste (como PB, AL e RN).

[36] Deve-se lembrar que as DP são referentes aos 4.989 municípios que entregaram a declaração de dados no SIOPE.
[37] Código disponível no APÊNDICE F. Os demais gráficos seguem códigos similares.

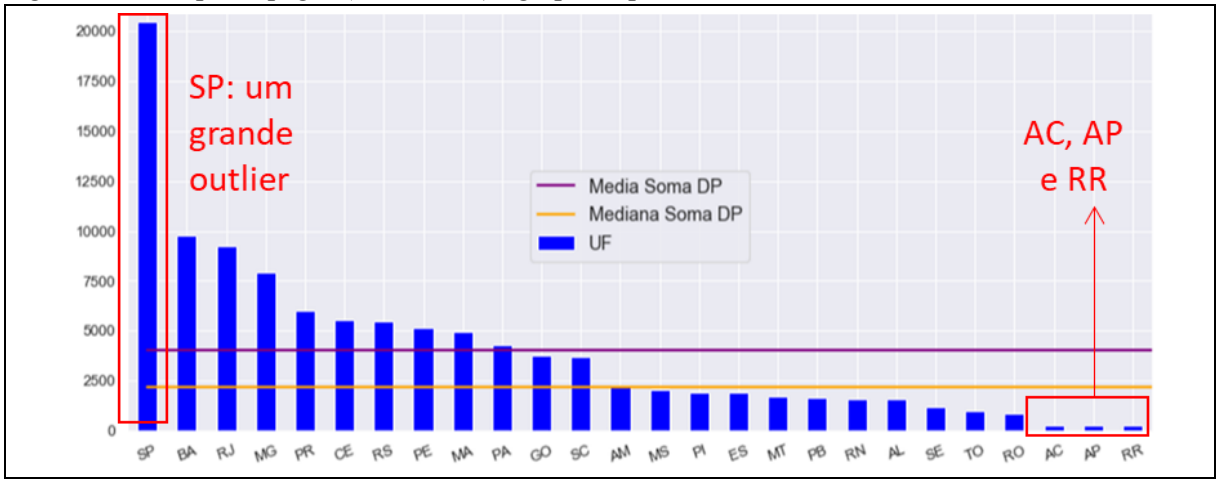
Figura 26 – Despesas pagas (em milhões) agrupadas por grupo de despesa e UF



Fonte: Elaborada pelo autor (2020).

O estado de São Paulo (SP), naturalmente por ser o estado mais populoso do país, absorve a maior parcela – 20% do valor total das DP (aproximadamente de 20,5 bilhões[38], de acordo com a Figura 27) – e contém os maiores montantes em todos os grupos de despesas. Os estados do Norte (RO, AC, AP e RR), menos populosos, apresentam os menores montantes.

Figura 27 – Despesas pagas (em milhões) agrupadas por UF

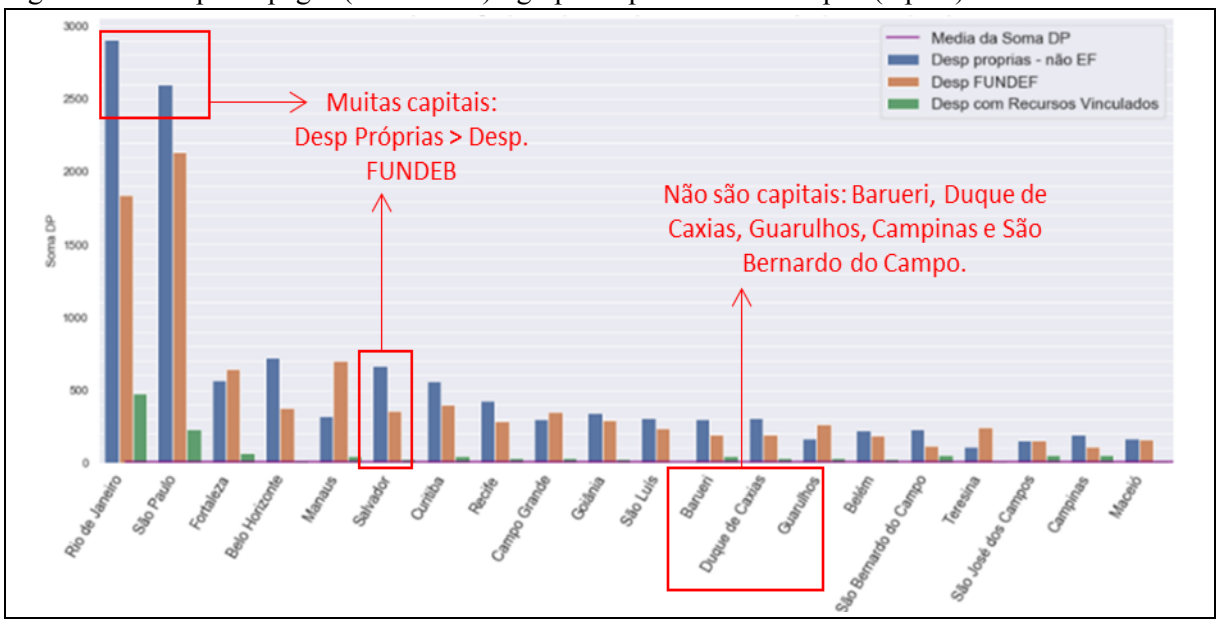


Fonte: Elaborada pelo autor (2020).

[38] A tabela com a soma dos valores por Estado se encontra no caderno *jupyter* "AnaliseExploratoria_MDE_EF", no item "Soma dos valores por Estado".

Com relação à lista dos 20 municípios com os maiores totais de DP, apresentado na Figura 28 – há o predomínio das capitais, embora haja municípios (que não sejam capitais, como Guarulhos e Campinas) com valores maiores que as demais capitais que não aparecem na lista. Muitas capitais (RJ, SP, BH) não seguem o padrão de utilizar mais gastos com FUNDEB – os maiores gastos ocorrem em Despesas Próprias.

Figura 28 – Despesas pagas (em milhões) agrupadas por GD e município (top 20)

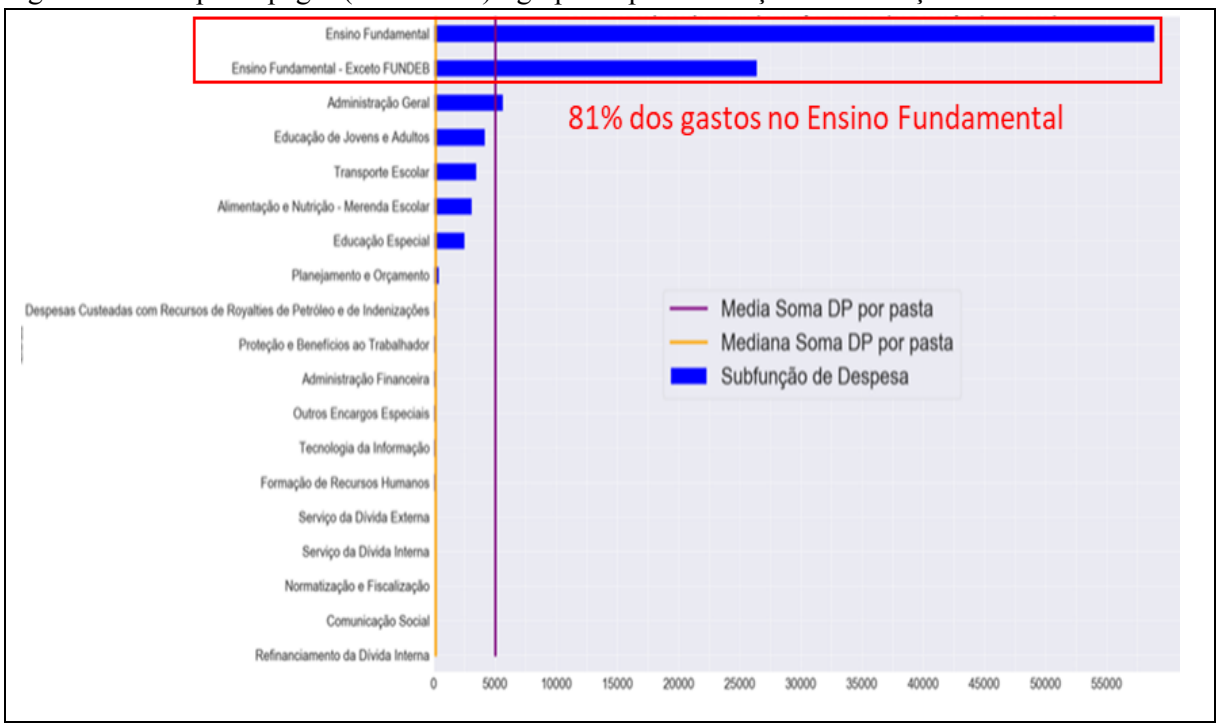


Fonte: Elaborada pelo autor (2020).

Dos 105 bilhões gastos com a Educação Fundamental, gastou-se 81% com o Ensino Fundamental (85,3 bilhões) e o restante foi distribuído nas outras subfunções – conforme mostra a Figura 29. Desta forma, a maioria das despesas se concentra na atividade fim - a modalidade de Ensino Fundamental financiada pelo FUNDEB – e nas devidas funções correlatas ao ensino (ex. Administração Geral, Transporte Escolar e Merenda).

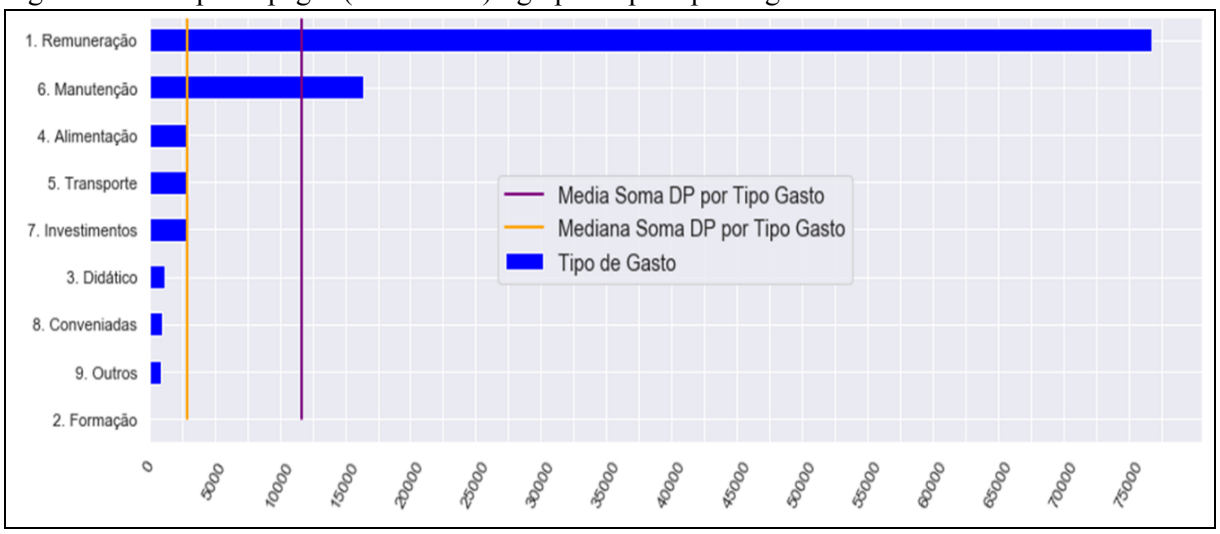
Conforme a Figura 30, ao se agrupar as despesas pagas por tipo de gasto, pode-se visualizar que grande parte dos gastos (73%) ocorre na remuneração dos profissionais do magistério, dos auxiliares e dos profissionais de apoio. O tipo remuneração representa vencimentos e demais obrigações patronais como FGTS e contribuições previdenciárias. Em seguida, maiores são os gastos com manutenção, como: materiais de consumo, materiais de conservação de bens imóveis, instalações e veículos, entre outros. Além disso, vale ressaltar que pouco se gasta na formação dos profissionais, como incentivo à qualificação, licença capacitação e auxílio financeiro a pesquisadores.

Figura 29 – Despesas pagas (em milhões) agrupadas por subfunção da Educação



Fonte: Elaborada pelo autor (2020).

Figura 30 – Despesas pagas (em milhões) agrupadas por tipo de gasto

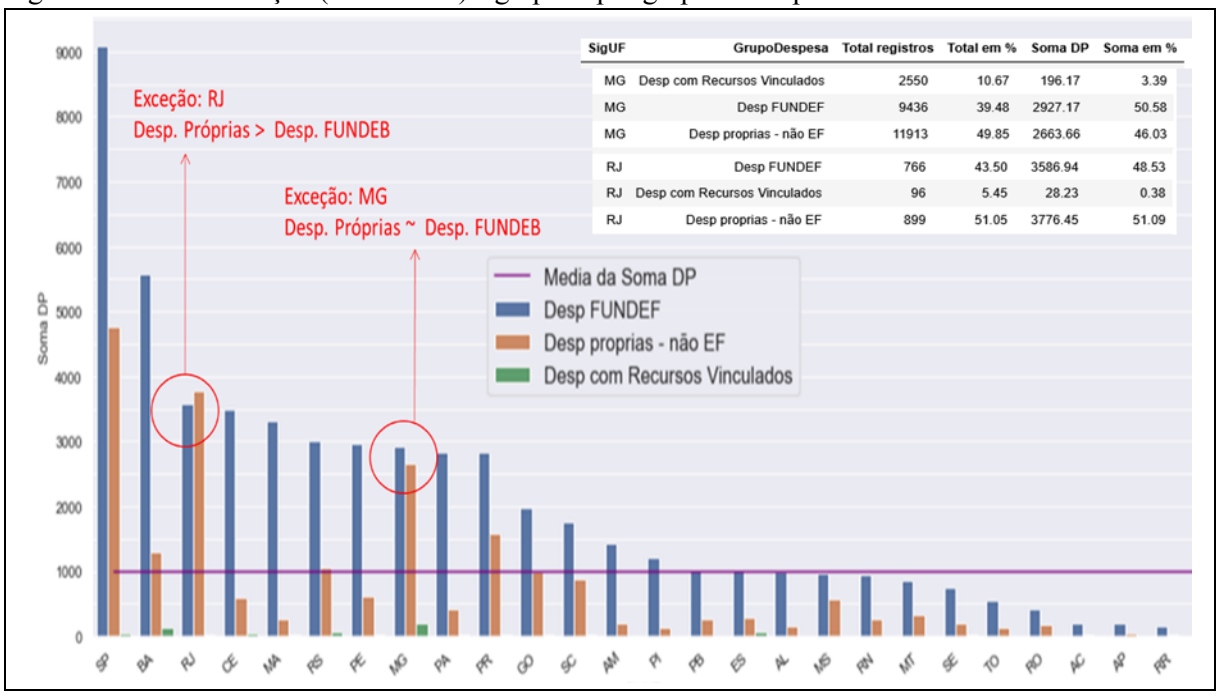


Fonte: Elaborada pelo autor (2020).

Ao se analisar, na Figura 31, o detalhamento dos valores do tipo Remuneração por grupo de despesa e UF, há claramente um padrão que ocorre com todos os estados: a maior parte da remuneração é predominantemente custeada com recursos do FUNDEB, sendo que a participação do FUNDEB varia de 60 a 90% do total[39]. Duas exceções são o estado do RJ (com 48% proveniente do FUNDEB) e MG (com 50% proveniente do FUNDEB).

[39] Pode-se visualizar os detalhes de valores e percentuais no caderno *jupyter* "AnaliseExploratoria_MDE_EF", elaborado para fundamentar as constatações descritas no presente trabalho.

Figura 31 – Remuneração (em milhões) agrupada por grupo de despesa e UF

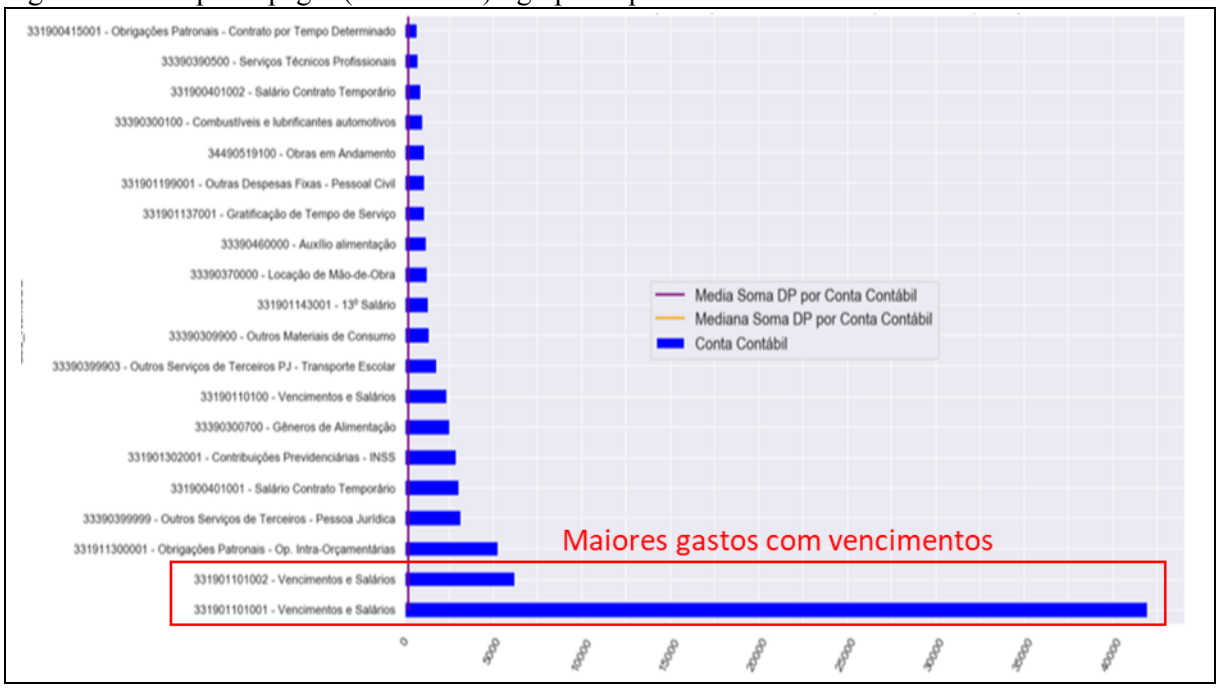


SigUF	GrupoDespesa	Total registros	Total em %	Soma DP	Soma em %
MG	Desp com Recursos Vinculados	2550	10.67	196.17	3.39
MG	Desp FUNDEF	9436	39.48	2927.17	50.58
MG	Desp proprias - não EF	11913	49.85	2663.66	46.03
RJ	Desp FUNDEF	766	43.50	3586.94	48.53
RJ	Desp com Recursos Vinculados	96	5.45	28.23	0.38
RJ	Desp proprias - não EF	899	51.05	3776.45	51.09

Fonte: Elaborada pelo autor (2020).

O agrupamento das despesas por contas contábeis, na Figura 32, reforça a predominância dos gastos com remuneração (gasta-se mais, naturalmente, com profissionais do magistério; e menos com os demais profissionais).

Figura 32 – Despesas pagas (em milhões) agrupadas pelas maiores contas contábeis



Fonte: Elaborada pelo autor (2020).

Nota-se que a conta contábil "Vencimentos e Salários" possui duas numerações. Diante disso, uma observação se faz necessária: o SIOPE foi estruturado de forma a separar contas contábeis de "Pessoal e Encargos Sociais" – uma para os profissionais do Magistério e outra para demais profissionais. Desta forma, a conta 3.31.00.00.00.00 - Pessoal e Encargos Sociais possui um dígito a mais para as contas abaixo dela:

- 3.31.00.00.00.00**.1** = Pessoal e Encargos Sociais - Profissionais do Magistério
- 3.31.00.00.00.00**.2** = Pessoal e Encargos Sociais - Outros Profissionais de Educação
- 3.31.00.00.00.00 = Pessoal e Encargos Sociais (demais pastas que não se referem a uma modalidade de ensino).

5.2.3 Gráficos de dispersão das despesas por UF e subfunção

Pode-se resumir algumas principais constatações discutidas nos itens anteriores:

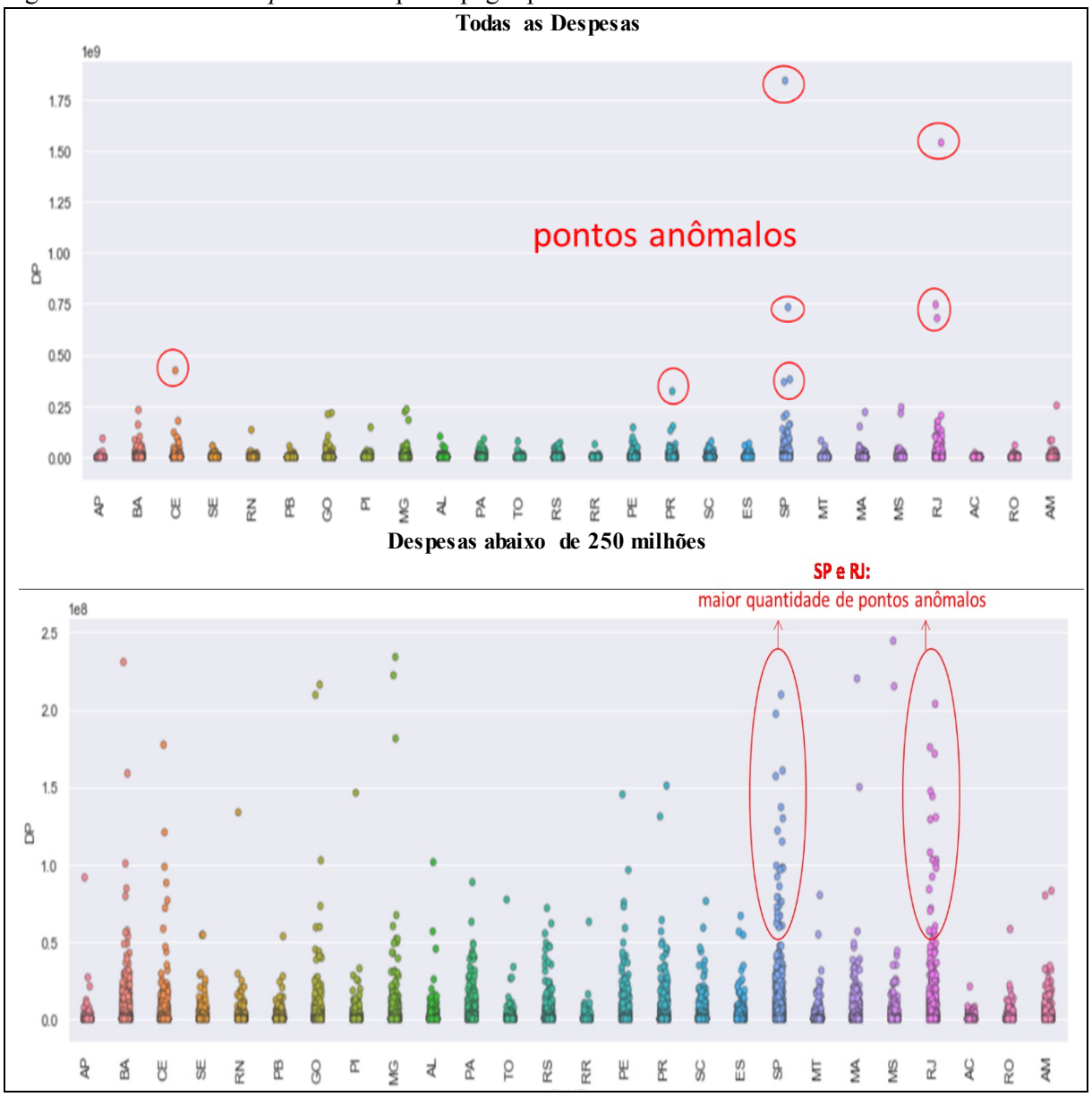
- Os gráficos de distribuição (*boxplots*), apresentados na Figura 19, indicam alguns registros bastante anômalos, que são as despesas acima de 300 milhões; e
- os diversos gráficos em barras (Figura 26 a Figura 32), que exibem as DP agrupadas por diversas variáveis, mostram a predominância de despesas FUNDEB aplicadas no Ensino Fundamental e na remuneração dos profissionais do magistério.

As constatações acima se referem a despesas de valores consolidados. Como complemento a estas constatações, são gerados alguns gráficos de dispersão das despesas por variáveis específicas (UF e por subfunção)[40], considerando-se todas as despesas e as despesas abaixo de 250 milhões (para amenizar os efeitos de *outliers* extremos).

Na Figura 33, o segundo gráfico de dispersão das despesas (até 250 milhões) por UF mostra que, apesar de se descartar os *outliers* extremos (presentes no primeiro gráfico), ainda há valores de despesas que se afastam da concentração dos demais valores restantes – sendo em RJ e SP os municípios com as dispersões mais espaçadas, com uma maior quantidade de pontos que se afastam dos valores mais concentrados. Quase todos os estados apresentam municípios com pontos que se afastam das áreas concentradas: alguns estados com mais pontos (SP, RJ), outros com menos pontos (AP, SE, PB, AC, AM, RR).

[40] Código disponível no APÊNDICE G. Os demais gráficos seguem códigos similares.

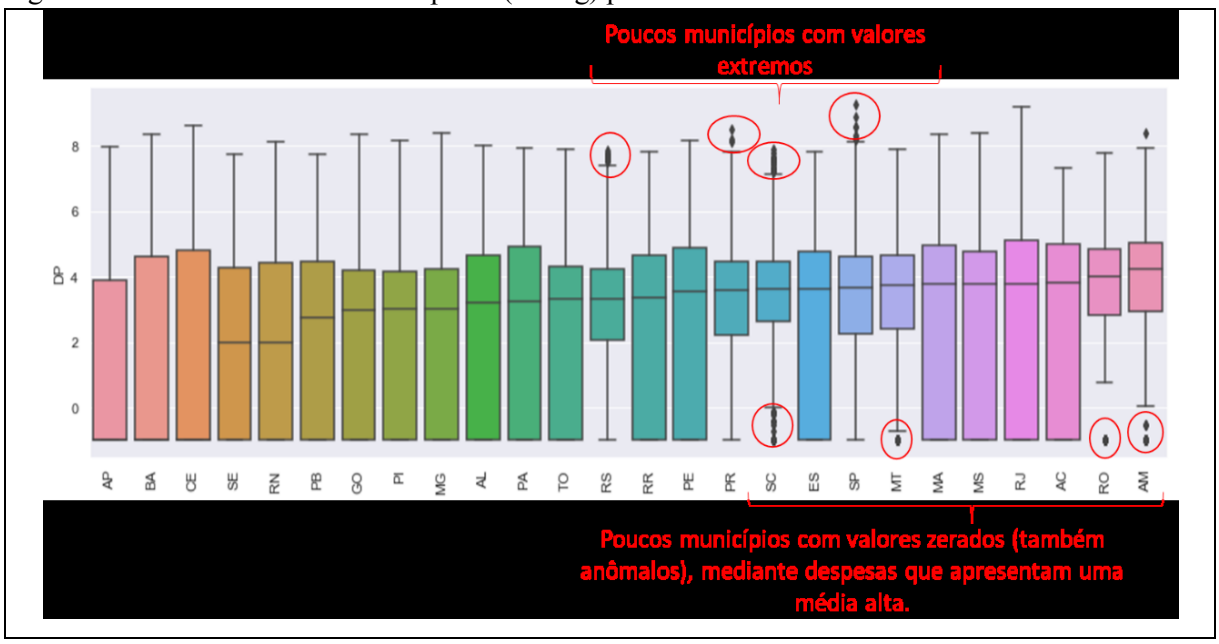
Figura 33 – Gráficos *StripPlots* - despesas pagas por UF



Fonte: Elaborada pelo autor (2020).

A fim de se explorar outras perspectivas de visualização, na Figura 34 são apresentados gráficos *boxplots* de todas as DP, em notação logarítmica, para cada estado. Há a presença de despesas abaixo de 100 milhões (10^8 = 100 milhões) que são anômalas considerando-se o escopo dentro de um estado (RS e SC, por exemplo). Além disso, pode-se perceber valores baixos que também são anômalos porque ocorrem para poucos municípios em um mesmo estado (SC, MT, RO e AM). Em outras palavras, há alguns municípios com gastos específicos bem discrepantes em relação aos demais municípios de um mesmo estado.

Figura 34 – Gráficos *BoxPlots* - despesas (em log) por UF

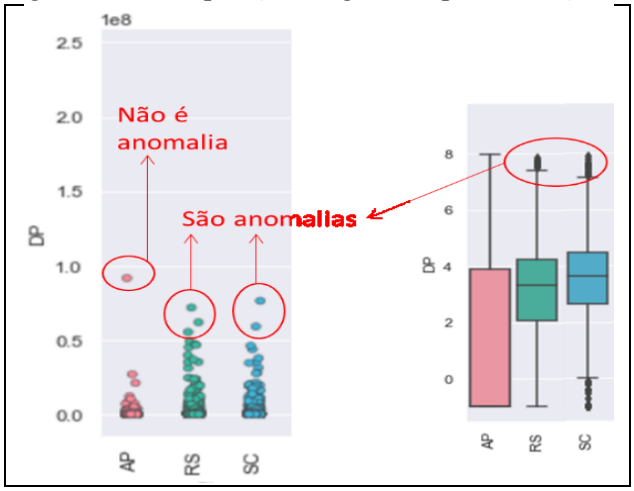


Fonte: Elaborada pelo autor (2020).

Vale enfatizar que os casos que, aparentemente, não apresentam anomalias se deve à distância interquartil muito grande, o que provavelmente é consequência de se ter um número alto de despesas com valor zero.

Conforme se pode notar através de alguns estados exemplos na Figura 35, é preciso analisar os gráficos de dispersão (lado esquerdo) em conjunto com os gráficos *boxplots* (lado direito) – estes últimos é que confirmam se os pontos afastados no gráfico de dispersão são realmente anômalos. Como exemplo, pode-se visualizar que o estado do AP aparenta ter um ponto anômalo (ponto de 100 milhões, bem distante dos demais) no gráfico de dispersão - entretanto, esse mesmo ponto não surge como anômalo no gráfico *boxplot*. Com isso, evidencia-se, novamente, a necessidade de outras técnicas mais eficazes para a detecção de anomalias.

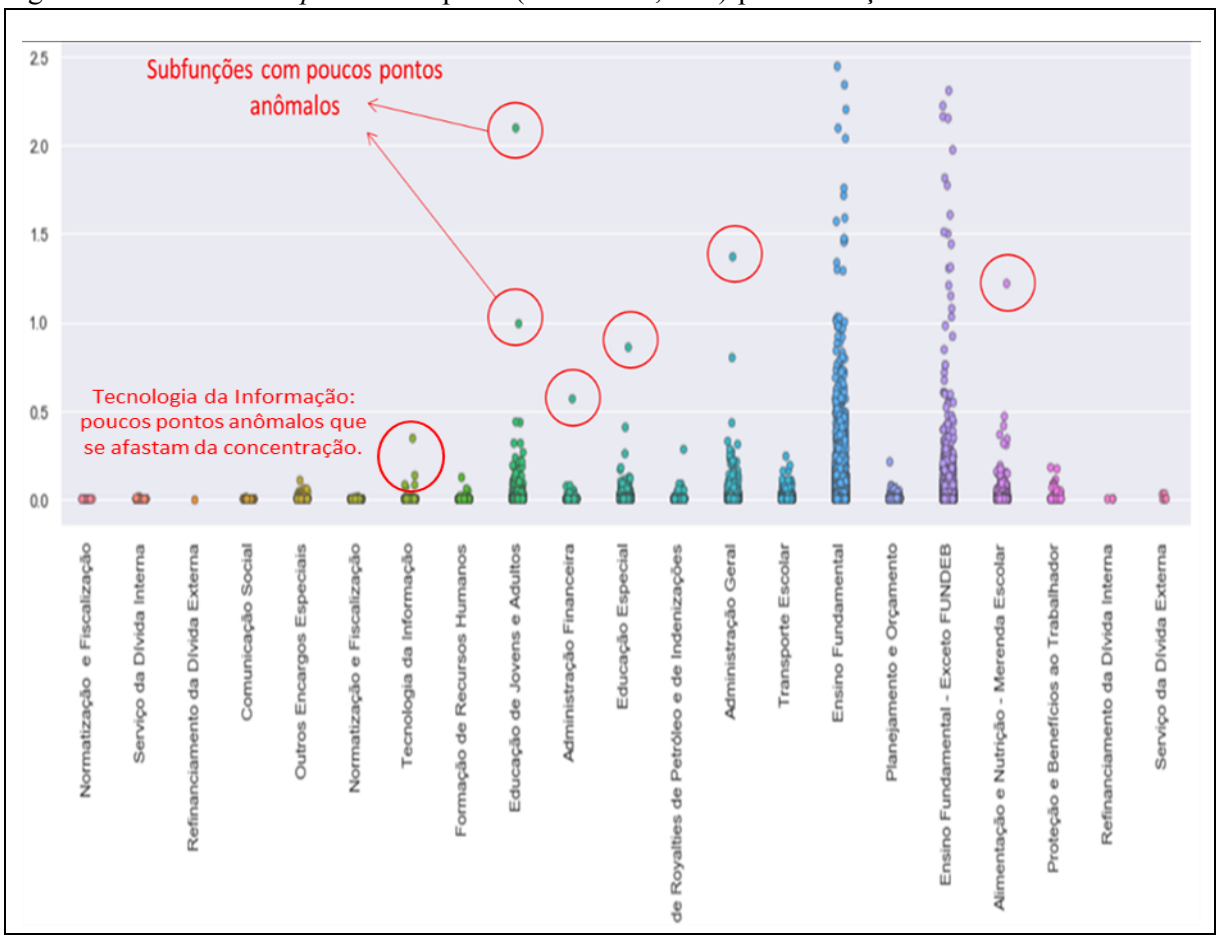
Figura 35 – Comparação de gráficos para detecção de despesas anômalas



Fonte: Elaborada pelo autor (2020).

O gráfico de dispersão das despesas por subfunção, na Figura 36, mostra algumas subfunções com valores de despesas que se afastam da concentração dos demais valores restantes. Alguns pontos relevantes chamam a atenção – como, por exemplo, os poucos pontos de grande valor em Educação de Jovens e Adultos, em Educação Especial, na Administração Geral, na Administração Financeira e na Merenda Escolar.

Figura 36 – Gráficos *StripPlots* - despesas (abaixo de 0,25 bi) por subfunção

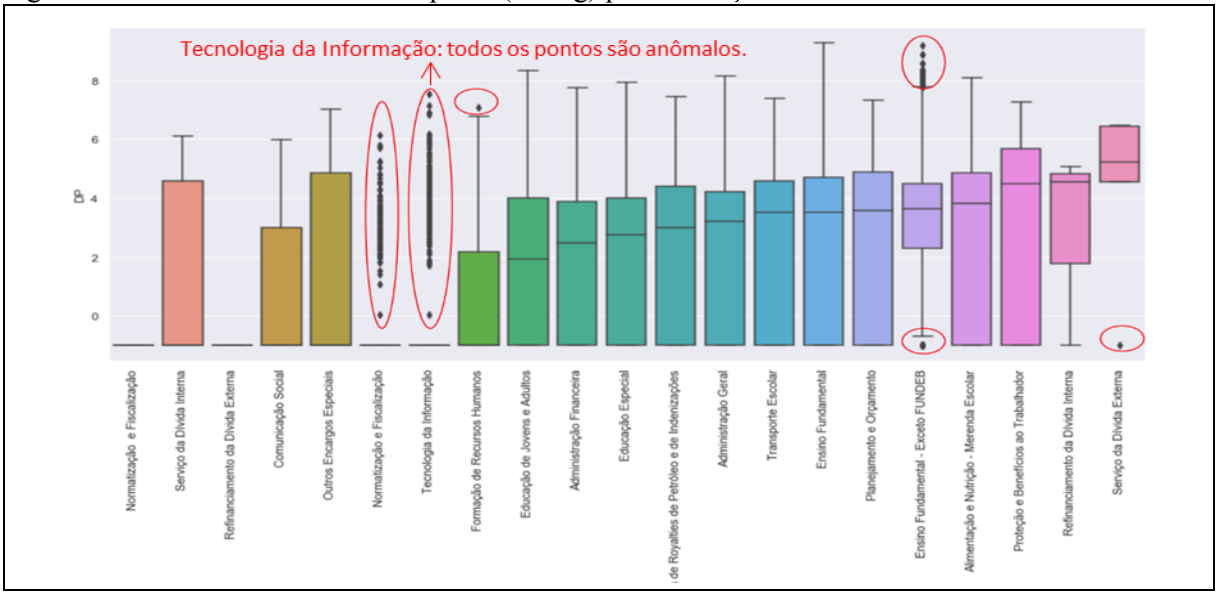


Fonte: Elaborada pelo autor (2020).

A Figura 37 apresenta os *boxplots* de todas as DP, em notação logarítmica, para cada subfunção. Os pontos que aparentam ser anômalos no gráfico de dispersão acima não o são no gráfico *boxplot* abaixo. Como exemplo, a subfunção Tecnologia da Informação apresenta poucos pontos afastados no gráfico de dispersão, mas o gráfico *boxplot* praticamente aponta todos os pontos como anômalos (pode ocorrer quando muitos municípios não preenchem valores, resultando em medianas e quartis próximos do valor zero). Pode-se concluir que, em determinadas situações, deve-se usar outras técnicas para a detecção mais eficaz de anomalias – não somente o uso de técnicas estatísticas e gráficos de visualização. Mais uma vez, deve-se lembrar a distância interquartil muito grande de certas subfunções (consequência de se ter um

número alto de despesas com valor zero), que resulta em não apontamento de anomalias pela técnica de gráficos *boxplot*.

Figura 37 – Gráficos *BoxPlots* - despesas (em log) por subfunção



Fonte: Elaborada pelo autor (2020).

5.2.4 Identificação preliminar de despesas anômalas através da AED

Com o uso das estatísticas e de gráficos de visualização, chegou-se em algumas despesas aparentemente anômalas, descritas na Tabela 8.

Tabela 8 – Identificação preliminar de anomalias nas despesas

Motivo Anomalia	**Ferramentas que permitiram a detecção**	**Qtd**
Despesas acima de 0,25 bilhão	gráficos *boxplots* das despesas (Figura 19) gráficos *stripPlots* com despesas x UF (Figura 33)	11
DP com Educação de Jovens e Adultos (acima 50 milhões)	gráficos *stripPlots* com despesas x Subfunção (Figura 36)	2
DP com Educação Especial (acima de 50 milhões)		1
DP com Administração Geral (acima de 50 milhões)		2
DP com Administração Financeira (acima de 50 milhões)		1
DP com Conveniadas[41] (acima de 100 milhões)	gráficos *StripPlots* com despesas x Tipo de Gasto (em cadernos *jupyter*)	2
DP com Alimentação[30] (acima de 100 milhões)	gráficos *StripPlots* despesas x Tipo de Gasto (em cadernos *jupyter*)	1

Fonte: Elaborada pelo autor (2020).

[41] Conforme gráficos de dispersão no caderno *jupyter* " AnaliseExploratoria_MDE_EF", item "StripPlots com Valores das Despesas abaixo de 0.25 bilhão – Tipo de Gasto", que não estão descritos no presente trabalho.

Detectar anomalias apenas com as técnicas estatísticas e gráficos de dispersão não se mostrou uma tarefa de eficácia adequada, pois o foco destas ferramentas acaba sendo em anomalias de valores extremos; e o que aparenta ser anomalia em um gráfico de dispersão (do tipo *stripPlot*), pode não o ser em um gráfico *boxplot*. Além disso, ainda é possível que existam outras anomalias de difícil identificação, como em outras variáveis (por ex. em contas contábeis) ou em despesas de valores menores. Análise de *boxplots*, neste caso, não são suficientes.

Um outro detalhe importante é que determinadas despesas poderão não ser anomalias de fato, se for levado em consideração o contexto de um determinado município – por exemplo, um gasto de 50 milhões poderá ser normal para o município de São Paulo (um município com muitas escolas e grande número de alunos), e não ser normal para municípios menores.

Em virtude dessas questões, o presente trabalho se propôs, no item 5.4, ao uso de algoritmos de clusterização nos dados de municípios (exceto dados de despesas) - com o objetivo principal de se criar grupos de municípios que sejam semelhantes entre si. Em seguida, no item 5.5, se procedeu à aplicação de bibliotecas específicas de detecção de anomalias nos dados de despesas de cada grupo. Naturalmente, caso haja municípios com gastos discrepantes, estes gastos serão as despesas anômalas naquele grupo, e não despesas anômalas em todo o conjunto de dados.

5.3 ANÁLISES EXPLORATÓRIAS – *DATAFRAME* DE MUNICÍPIOS

5.3.1 Resumo do *dataframe* de municípios

Em virtude da grande quantidade de atributos, alguns procedimentos adicionais de limpeza foram realizados – a retirada de colunas desnecessárias, de subfunções com muitos valores zerados[42], de contas contábeis de preenchimento por apenas até 10 municípios e de variáveis com grande dispersão de valores.

Figura 38 – Resumo do *dataframe* de municípios

```
Total de Municipios:  4989 Municípios
Total de colunas do dataframe apos exclusao:  132 colunas

Colunas do dataframe de despesas:

12 Variáveis categóricas:  ['CodUF' 'NomeUF' 'SigUF' 'CodMun' 'CodIBGE' 'CodIBGE_Completo'
 'NomeMunicipio' 'Regiao' 'MesoRegiao' 'NomeMesoRegiao' 'MicroRegiao'
 'NomeMicroRegiao']

4 Variáveis numéricas inteiras:  ['Pop_estimada' 'QtdEscolas' 'QtdDocentes' 'NUM_MATR_361']

116 Algumas Variáveis numéricas float:  ['IDHM' 'IDHM_E' 'IDHM_L' 'IDHM_R' 'IDEB_AI' 'IDEB_AF' 'TxEvasao_EF'
 'CustoAluno' 'DespesaProf' 'DespFUNDEB' 'DespVinc' 'DespProp' 'tgRemun'
 'tgFormacao' 'tgDidatico' 'tgAlim' 'tgTransp' 'tgManut' 'tgInvest'
 'tgConv' 'tgOutros' 'Ação Judicial FUNDEF - Precatórios'
 'Outras Transf Recursos do FNDE' 'Outros Recursos Destinados à Educação'
 'PDDE' 'PNAE' 'PNATE' 'Transf Convênios - Educação'
 'Vincul a Contrib Social do Salário-Educação' 'AdmFinanc' 'AdmGeral'
 'MerEscolar' 'ComunSocial' 'DespCusteadasRecRoyPetrIndeniz'
 'EducEspecial' 'EducJA' 'EnsFund' 'EnsFund_exc' 'FormRH' 'NormatFisc2'
 'OutrosEE' 'PlanOrc' 'ProtBenefTrab' 'ServDivInt' 'TI' 'TranspEsc'
 '3.31.90.01.00.00 - Aposentadorias' '3.31.90.03.00.00 - Pensões'
 '3.31.90.05.00.00 - Outros Benefícios Previdenciários'
 '3.31.90.08.00.00 - Outros Benefícios Assistenciais'] ... e demais contas contábeis.
```

Fonte: Elaborada pelo autor (2020).

5.3.2 Estatísticas e distribuição dos valores dos indicadores do IBGE, INEP e PNUD

A Figura 39 apresenta as estatísticas, histogramas e gráficos *boxplots* para alguns indicadores do IBGE (População estimada) e INEP (quantidade de escolas, quantidade de docentes e quantidade de alunos[43]). A descrição dos campos pode ser revista no item 4.2.6.

As principais constatações são resumidas abaixo:

- há uma certa dispersão dos dados devido à diferença entre média (valor alto) e mediana (valor baixo) e aos desvios padrão elevados. Essa dispersão é normal, pois há diferenças dos muitos municípios menores (menor densidade demográfica, menor número de escolas) com as poucas grandes capitais (maior densidade demográfica, maior número de escolas);

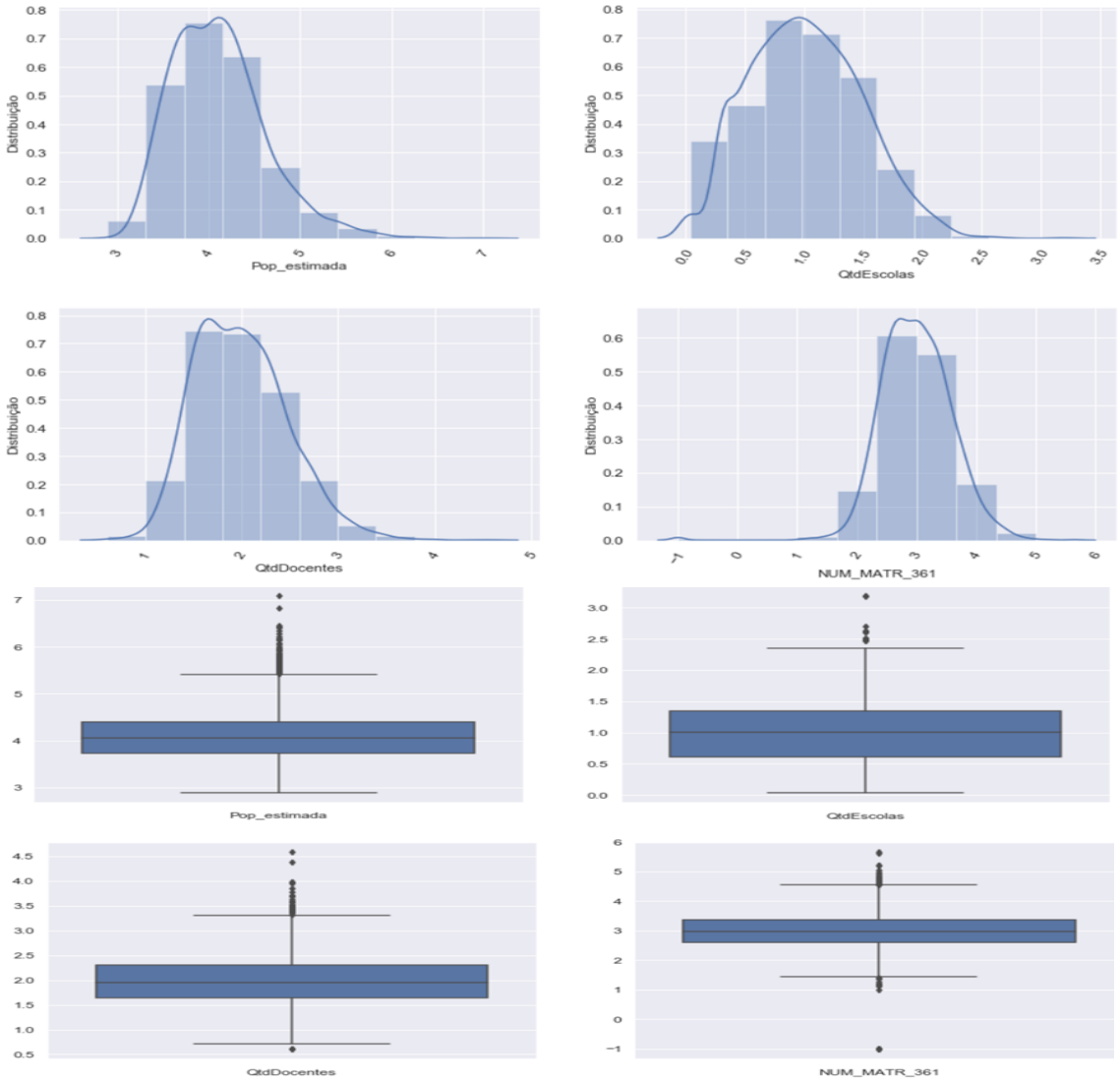
[42] É importante esclarecer que a retirada de colunas de subfunções ou de contas contábeis com muitos valores zerados não implica em perda de informação, pois o total de todas as despesas de um município ainda pode ser consultada somando-se as colunas dos grupos de despesas.

[43] O campo NUM_MATR_361 se refere à quantidade de alunos no Ensino Fundamental (modalidade 361), conforme especificado no item 4.2.6 e no APENDICE E.

- ao se normalizar os valores em notação logarítmica, é possível a geração de histogramas com curvas que se aproximam de uma distribuição normal, embora constem alguns *outliers* (histogramas de caudas muito compridas);
- nos *boxplots* de valores em notação logarítmica, 50% dos valores respeitam um determinado intervalo.

Figura 39 – Estatísticas e distribuição: população, escolas, docentes e alunos.

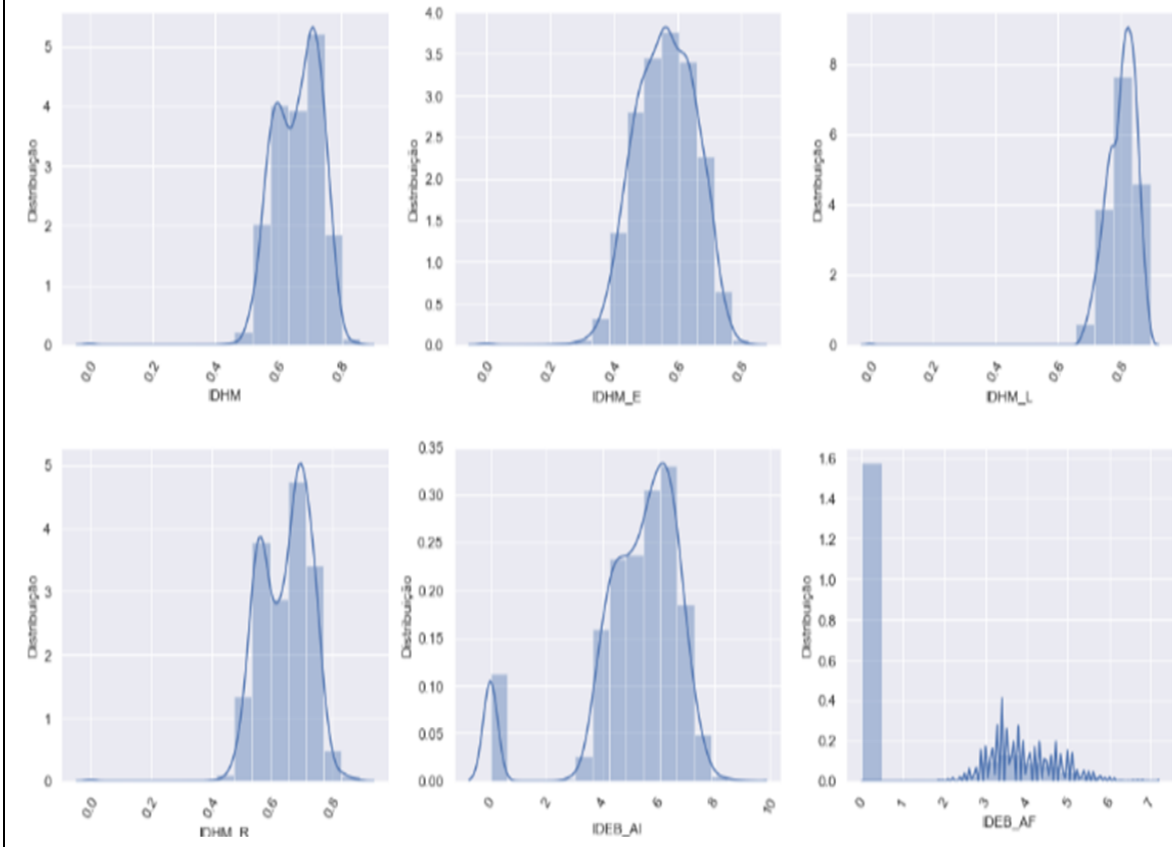
	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361
count	4989.00	4989.00	4989.00	4989.00
mean	37606.57	19.39	213.71	2693.72
std	227740.90	41.25	768.87	10708.04
min	781.00	1.00	4.00	0.00
25%	5369.00	4.00	43.00	396.00
50%	11531.00	10.00	90.00	957.00
75%	25354.00	22.00	202.00	2394.00
max	12252023.00	1539.00	38406.00	458634.00



Fonte: Elaborada pelo autor (2020).

A Figura 40 apresenta as estatísticas, histogramas e gráficos *boxplots* para outros indicadores, a saber: IDHM, IDHM Educação, IDHM Longevidade e IDHM Renda; Nota IDEB no Ensino Fundamental, sendo uma nota para Anos Iniciais (IDEB_AI) e outra para Anos Finais (IDEB_AF); e Taxa de Evasão Escolar. A descrição dos campos pode ser revista no item 4.2.6. Em uma breve análise, há pouca dispersão nos dados: alguns histogramas se assemelham a uma curva normal, com inclinação para esquerda ou direita, e com a presença de poucos *outliers*. A exceção ocorre com o indicador IDEB_AF, para o qual as estatísticas indicam mais de 75% de valores zerados porque poucos municípios estão com essa nota contabilizada.

Figura 40 – Estatísticas e distribuição: demais indicadores INEP e PNUD



	IDHM	IDHM_E	IDHM_L	IDHM_R	IDEB_AI	IDEB_AF	TxEvasao_EF
count	4989.00	4989.00	4989.00	4989.00	4989.00	4989.00	4989.00
mean	0.66	0.56	0.80	0.65	5.20	0.96	3.41
std	0.07	0.09	0.05	0.08	1.72	1.74	2.04
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.60	0.49	0.77	0.57	4.50	0.00	2.00
50%	0.67	0.56	0.81	0.66	5.60	0.00	3.00
75%	0.72	0.63	0.84	0.71	6.30	0.00	4.40
max	0.86	0.82	0.89	0.89	9.10	7.20	21.50

Fonte: Elaborada pelo autor (2020).

5.3.3 As principais constatações sobre dados de despesas

As demais variáveis (grupo de despesa, programas, tipo de gasto e subfunções) tiveram estatísticas, histogramas e *boxplots* gerados em cadernos *jupyter*. Pode-se resumir as seguintes conclusões:

- todos os histogramas das despesas apresentam grande dispersão dos dados, com alta concentração de valores baixos e pouca concentração de valores elevados;
- ao se normalizar os valores, é possível a geração de alguns histogramas que se aproximam de uma curva normal, embora com presença de *outliers*;
- de acordo com os gráficos *boxplot* normalizados, quase todas as variáveis, consideradas isoladamente, apresentam outliers;
- há algumas variáveis com predominância de valores zero (mais de 50% ou 75%) – como por exemplo: subfunções de Administração Financeira, de Comunicação Social e de Formação de RH. As exceções são as subfunções Merenda Escolar e Transporte Escolar – com o primeiro quartil (25%) diferente de zero;
- as variáveis com mais de 75% de valores zerados são aquelas de preenchimento por poucos municípios. Neste caso, estas colunas foram retiradas do *dataframe*, pois trata-se de despesas preenchidas por poucos municípios (eventos raros), não representando, desta forma, a realidade da maioria dos municípios;
- As subfunções principais - que são as modalidades de ensino (Educação Fundamental, Educação Especial e Educação de Jovens e Adultos) - possuem mais de 50% dos valores preenchidos. A maioria dos municípios, portanto, se concentra em declarar valores nas modalidades de ensino, e menos nas funções de apoio.

5.3.4 As principais constatações sobre dados das contas contábeis

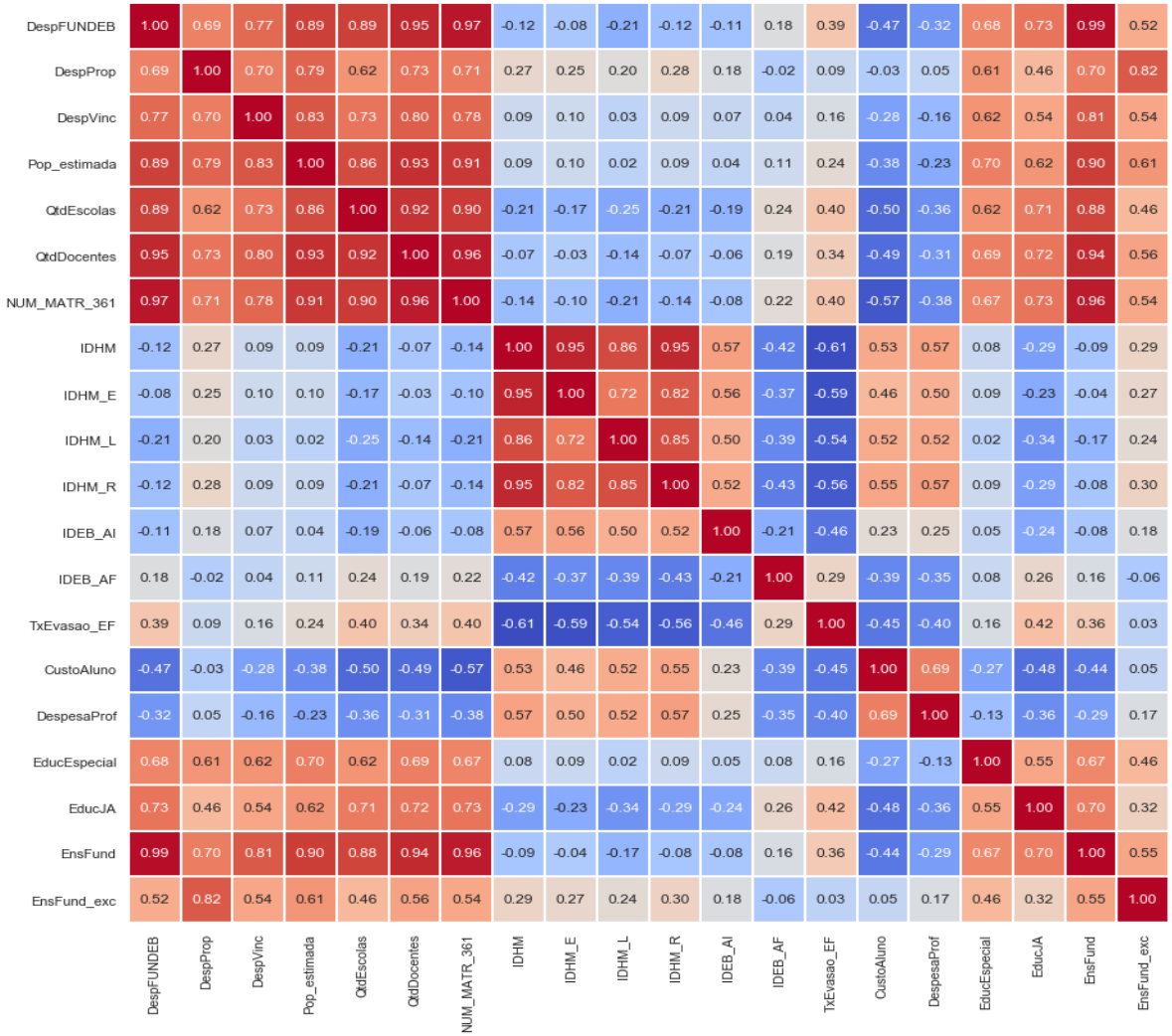
Finalmente, as estatísticas, histogramas e *boxplots* das 93 contas contábeis também se encontram gerados em cadernos *jupyter*. Foram detectadas 85 contas com mais de 50% de valores zerados; destas 85, são 78 contas com mais de 75% de valores zerados. Ao se normalizar os valores em log, ainda são poucas as contas em que os histogramas se aproximam de uma curva normal, mesmo com a presença de *outliers*. Não sendo possível a exclusão de tantas contas contábeis somente com o critério de 50 ou 75% de valores zeros, decidiu-se excluir as contas que tenham sido preenchidas por até dez municípios (25 contas).

5.3.5 Investigação de correlações

Um coeficiente de correlação mede o grau pelo qual duas variáveis quantitativas estão associadas ou relacionadas entre si. Embora a correlação não implique em causalidade, pode ser interessante quantificar a relação entre as variáveis através de inúmeras medidas, como o coeficiente de *Pearson* ou o coeficiente de *Spearman*.

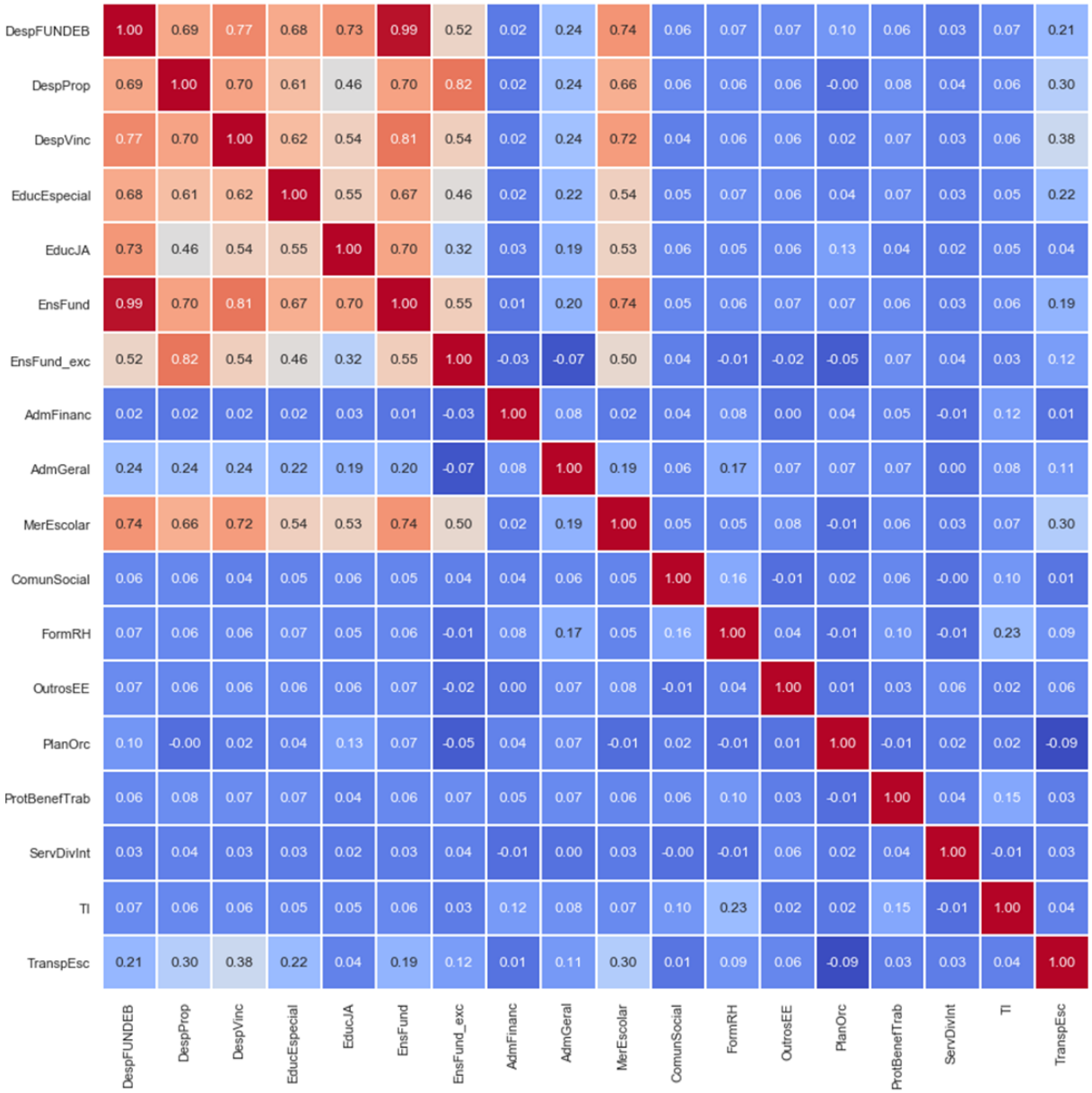
Desta forma, as figuras 41 a 43 apresentam os coeficientes de *Spearman* calculados para relações entre algumas variáveis do *dataframe* de municípios. Correlações positivas indicam que duas variáveis tendem a mudar juntas (se o valor de uma variável aumenta/diminui, o valor da outra variável tende a aumentar/diminuir também); correlações negativas indicam que duas variáveis seguem em direções opostas (se o valor de uma variável aumenta, o valor da outra variável tende a diminuir). A correlação de *Spearman* varia entre os valores -1 a 1, e a relação entre variáveis é mais intensa quanto mais o valor se aproxima de 1 (cores mais avermelhadas) ou -1 (cores em azul mais escuro).

Figura 41 – Cálculo das correlações entre variáveis pelo Método *Spearman* (1)



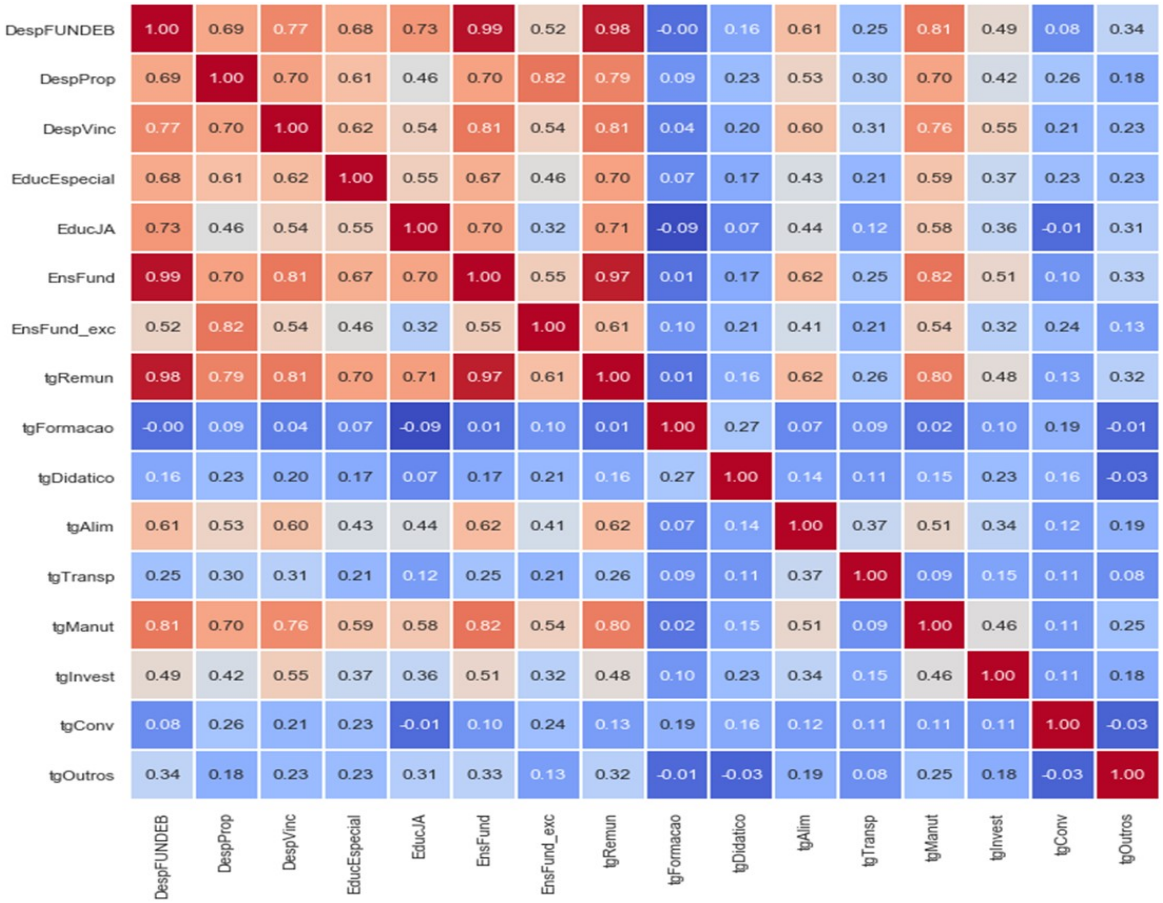
Fonte: Elaborada pelo autor (2020).

Figura 42 – Cálculo das correlações entre variáveis pelo Método *Spearman* (2)



Fonte: Elaborada pelo autor (2020).

Figura 43 – Cálculo das correlações entre variáveis pelo Método *Spearman* (3)

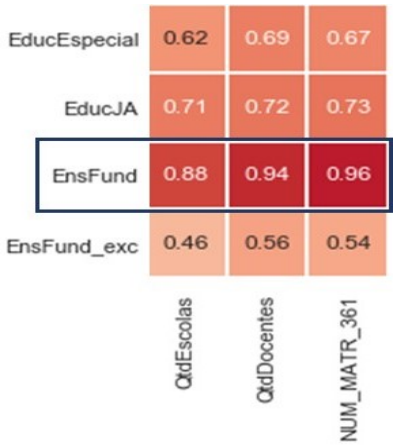


Fonte: Elaborada pelo autor (2020).

A partir dessas figuras com os cálculos de correlação nas páginas anteriores, a Tabela 9, apresentada a seguir, é um resumo das principais correlações percebidas entre algumas variáveis relevantes. A primeira linha da tabela, por exemplo, indica, na primeira coluna, que há relações fortes entre os Grupos de Despesa (próprias, FUNDEB e vinculadas), sendo ainda apontada, na segunda coluna da tabela, a correlação *Spearman* de maior valor (0,77) que ocorre entre as despesas FUNDEB e as despesas vinculadas. Em alguns caso, optou-se por indicar não só a maior, mas todas as relações (como é o caso da linha "População com quantitativos do INEP").

Outro exemplo pode ser citado (e está inserido na tabela abaixo): ao verificar correlações entre os quantitativos do INEP – ou seja, a quantidade de escolas, de alunos do Ensino Fundamental (representado pelo campo NUM_MATR_361) e de docentes – com as modalidades de ensino (Educação Especial, Educação de Jovens e Adultos, Ensino Fundamental e Ensino Fundamental - exceto FUNDEB), a segunda coluna descreve que as correlações de maior intensidade ocorrem entre esses quantitativos e o Ensino Fundamental.

Tabela 9 – Resumo das principais correlações

Correlação	Valor da maior correlação (Correlação de *Spearman*)
Figura 41	
Grupo de Despesa com Grupo de Despesa	despesas FUNDEB com despesas vinculadas (0,77)
Grupo de Despesa com População	despesas FUNDEB com população (0,89)
Grupo de Despesa com quantitativos do INEP[44]	despesas FUNDEB com quantidades de alunos (0,97) despesas FUNDEB com quantidades de professores (0,95) despesas FUNDEB com quantidades de escolas (0,89)
Grupo de Despesa com Modalidade de Ensino[45]	despesas FUNDEB com Ensino Fundamental (0,99) despesas próprias com Ens. Fundam. (exc FUNDEB)[46] (0,82)
População com quantitativos do INEP	população com a quantidade de professores (0,93) população com a quantidade de alunos (0,91) população com a quantidade de escolas (0,86)
População com Modalidade de Ensino	população com o Ensino Fundamental (0,90)
Quantitativos do INEP	quantidade de professores e quantidade de alunos (0,96)
Quantitativos do INEP com Modalidade de Ensino	quantidade de escolas com Ensino Fundamental (0,88) quantidade de professores com Ensino Fundamental (0,94) quantidade de alunos com Ensino Fundamental (0,96) 
Figura 42	
Merenda com Grupo de Despesa	merenda escolar com Despesas FUNDEB (0,74) merenda escolar com Despesas vinculadas (0,72)
Merenda com Modalidade de Ensino	merenda escolar com Ensino Fundamental (0,74)
Figura 43	
Tipos de Gastos com Grupo de Despesa	tgRemun com despesas FUNDEB (0,98) tgRemun com despesas vinculadas (0,81) tgManu com despesas FUNDEB (0,81) tgManu com despesas vinculadas (0,76)
Tipos de Gastos com Modalidade de Ensino	tgRemun com Ensino Fundamental (0,97) tgManu com Ensino Fundamental (0,82)

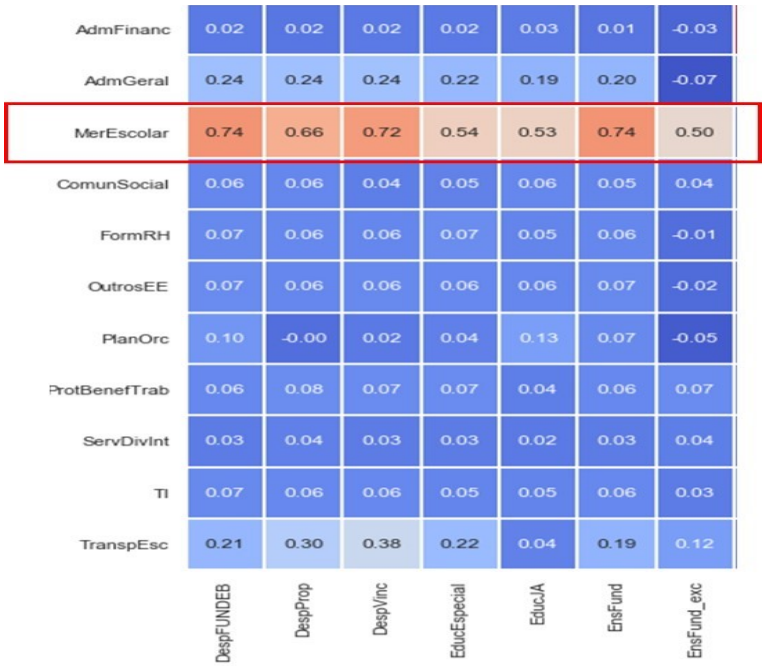
[44] Quantitativos do INEP são: quantidade de escolas, de alunos do Ensino Fundamental (representado pelo campo NUM_MATR_361) e de docentes em determinado município.
[45] Modalidade de Ensino são: Educação Especial, Educação de Jovens e Adultos, Ensino Fundamental e Ensino Fundamental (exceto FUNDEB).
[46] Ens. Fundam. (exc FUNDEB) é um termo encontrado na base de dados, e são as despesas gastas com Ensino Fundamental cujos recursos são próprios, ou seja, não são provenientes do FUNDEB.

Outras correlações detectadas[47]	
Despesas Próprias (DPr) com programas	DPr com "Vincul a Contrib Social do Salário-Educação" (0,92) DPr com PNAE (0,89)
Despesas FUNDEB (DF) com programas	DF com "Vincul a Contrib Social do Salário-Educação" (0,87) DF com PNAE (0,93)
Despesas Próprias com CC (as 5 mais)	3.31.90.11.00.00 – Venc. e Vant. Fixas - Pessoal Civil (0,97) 3.33.90.39.00.00 - Serviços de Terceiros – PJ (0,90) 3.31.91.13.00.00 – Obrig. Patr - Op. Intra-Orç. (0,86) 3.31.90.16.00.00 - Outras Desp. Variáveis - Pessoal Civil (0,67) 3.33.91.39.00.00 - Outros Srv. Terc. - PJ - Op. Intra-Orç. (0,67)
Despesas FUNDEB com CC (as 5 mais)	3.31.90.11.00.00 - Venc. e Vant. Fixas - Pessoal Civil (0,98) 3.33.90.39.00.00 - Serviços de Terceiros – PJ (0,93) 3.31.91.13.00.00 - Obrig. Patr - Op. Intra-Orç. (0,75) 3.33.90.36.00.00 - Outros Serv. de Terceiros – PF (0,70) 3.33.90.30.00.00 - Material de Consumo (0,63)
Índices IDHM com IDEB	IDHM com IDEB_AI (Anos Iniciais) - 0,57

Fonte: Elaborada pelo autor (2020).

Conforme a Figura 44, muitas funções de apoio ao ensino (como Administração Financeira, Administração Geral, Comunicação Social, Formação de RH, TI, etc.) não possuem correlações relevantes (ou seja, a intensidade das relações é de valor baixo, próximo de zero) com outras variáveis. A única função de apoio que demonstrou alguma correlação notável com determinadas variáveis (grupos de despesas e modalidades de ensino) foi Merenda Escolar (em destaque na figura abaixo).

Figura 44 – Correlações Grupos de Despesa/ Modalidades de ensino com Merenda Escolar



Fonte: Elaborada pelo autor (2020).

[47] Gráficos das outras correlações podem ser encontradas no caderno *jupyter* "AnaliseExploratoria_MDE_EF", item "Investigação de Correlações".

5.4 CLUSTERIZAÇÃO DE MUNICÍPIOS SEMELHANTES

5.4.1 Objetivos da clusterização de municípios

Algumas correlações entre os dados - como a população com os quantitativos do INEP (acima de 0,86) e os índices IDHM com IDEB_AI (acima de 0,50) - motivaram a clusterização de municípios semelhantes[48], sem considerar os dados de despesas ou de contas contábeis. Parte-se do pressuposto que municípios semelhantes (com populações similares, indicadores próximos e quantidades parecidas de escolas, professores e alunos matriculados) devem ter despesas educacionais também semelhantes, ao menos em mesma ordem de grandeza.

Foram testados quatro algoritmos de clusterização[49] – este capítulo detalha os dois algoritmos com os melhores resultados: *k-Means* e *Agglomerative Clustering*. Em determinados momentos, resultados obtidos com o algoritmo DBSCAN são também demonstrados.

Após a criação dos agrupamentos dos municípios, alguns gráficos foram criados para facilitar a visualização da coerência destes conjuntos - ou seja, como forma de validar se determinado algoritmo foi capaz de separar bem os dados.

5.4.2 Decisão sobre o escalonamento dos dados

O escalonamento de variáveis é um método utilizado para padronizar um intervalo de variáveis independentes, sendo também denominado de normalização de dados (SRIVASTAVA, 2019). Deve-se, portanto, uniformizar os dados de características dos municípios (IBGE, INEP e PNUD) em uma mesma escala, antes de se utilizar quaisquer algoritmos de clusterização - principalmente quando a similaridade se baseia no cálculo de distâncias entre os pontos.

Várias são as técnicas de escalonamento de variáveis – no presente trabalho, foram testadas as técnicas de escalonamento definidas na tabela abaixo para cada algoritmo de clusterização utilizado. Gráficos[50] de população (em valores escalonados) com quantidade de alunos matriculados (em valores escalonados) foram criados para fundamentar a escolha pelo *scaler* mais apropriado.

---

[48] Variáveis utilizadas: Pop_estimada, IDHM, IDHM_E, IDHM_L, IDHM_R, QtdEscolas, QtdDocentes, NUM_MATR_361, IDEB_AI, IDEB_AF e TxEvasao_EF.

[49] Foram testados os seguintes algoritmos de clusterização: *k-Means*, DBSCAN, *MeanShift* e *Agglomerative Clustering*, e estes podem ser visualizados no caderno *jupyter* "ClusterizacaoMunicipios".

[50] O caderno *jupyter* "ClusterizacaoMunicipios" utiliza três gráficos comparativos (população x quantidade de alunos matriculados, população x IDHM e quantidade de escolas x quantidade de docentes). No presente trabalho, é apresentado apenas um gráfico comparativo para cada algoritmo de clusterização testado. Parte do código está disponível no APÊNDICE I.

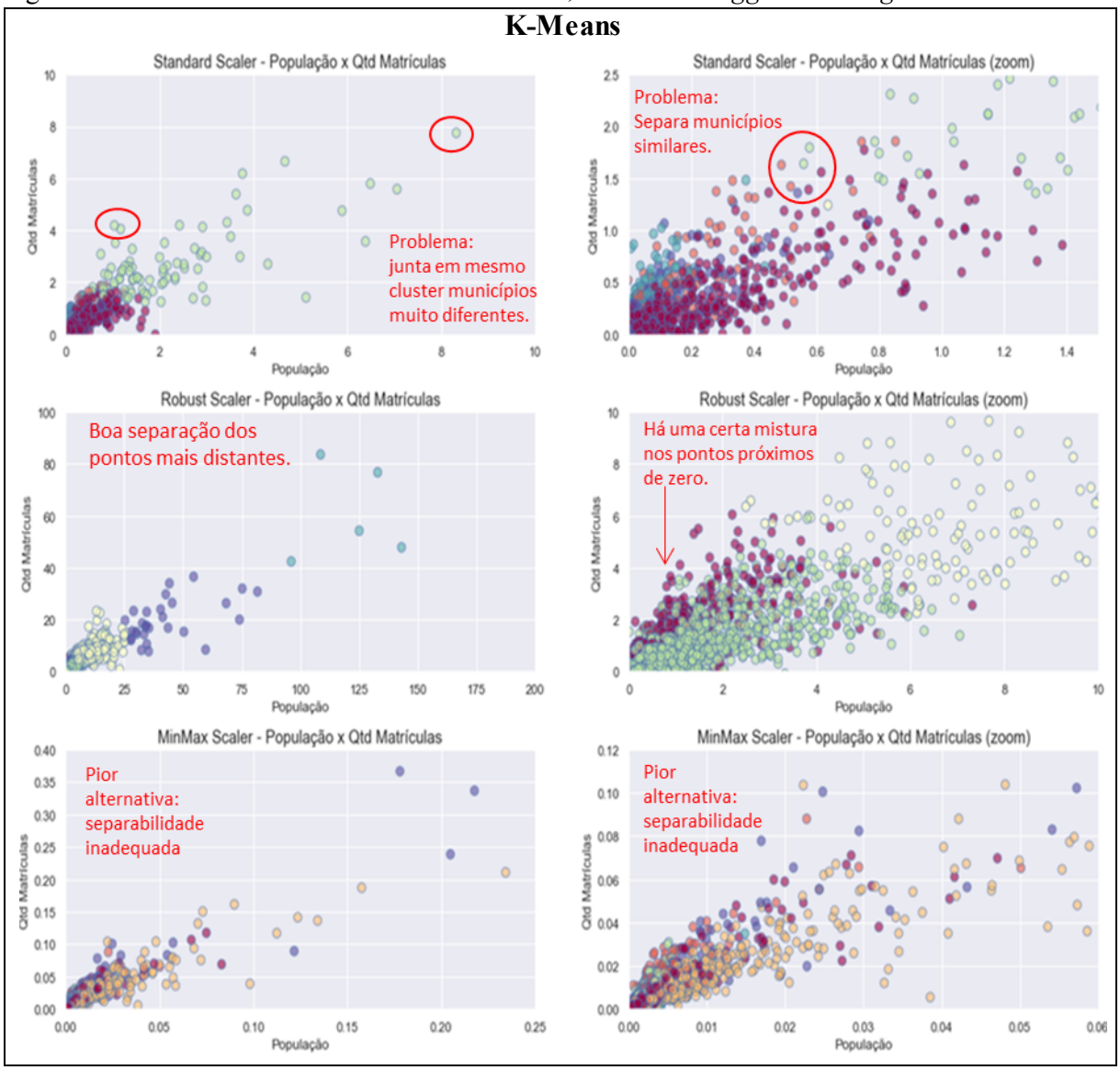
Tabela 10 – Técnicas utilizadas para a transformação de variáveis

Standard Scaler	Assume que os dados são normalmente distribuídos em cada variável e, por isso, dimensiona para que a distribuição tenha média de valor zero com desvio padrão de valor 1. Se dados não são normalizados, não é uma boa alternativa de escalonamento.
Robust Scaler	Utiliza abordagem semelhante ao escalonamento MinMax, mas usa o intervalo inter-quartil ao invés do intervalo entre 0 a 1. É adequada para dados com presença de *outliers*.
MinMax Scaler	Reduz os dados para que assumam o intervalo de valores entre 0 e 1, ou -1 a 1 se houver valores negativos. Essa técnica é adequada para distribuições que não sejam gaussianas ou quando o desvio padrão das variáveis é de valor reduzido. É sensível a *outliers*.

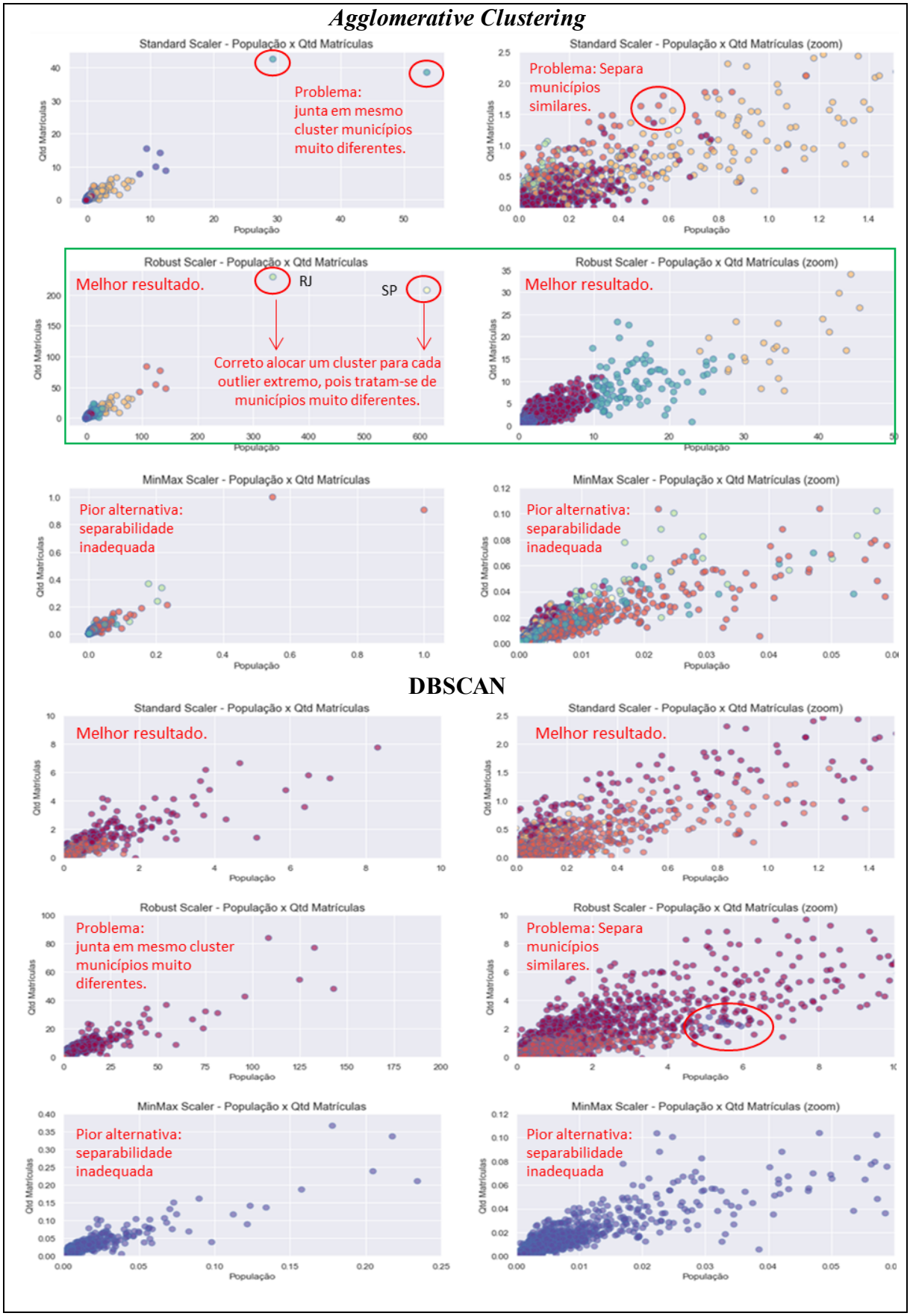
Fonte: Elaborada pelo autor (2020), adaptado de (SRIVASTAVA, 2019).

Nos gráficos da Figura 45[51], o lado esquerdo exibe todos os pontos (para visualizar os grupos com os pontos mais distantes), enquanto o lado direito realiza uma espécie de *zoom*, de forma a apresentar os grupos que são formados com os pontos mais concentrados.

Figura 45 – Escalonamento de dados com *K-Means,* DBSCAN e *Aggl. Clustering*



[51] Parâmetros utilizados para a geração dos gráficos: o número de clusters escolhido foi sete (7), e os parâmetros para DBSCAN foram: eps=0.9 e min_samples=5.



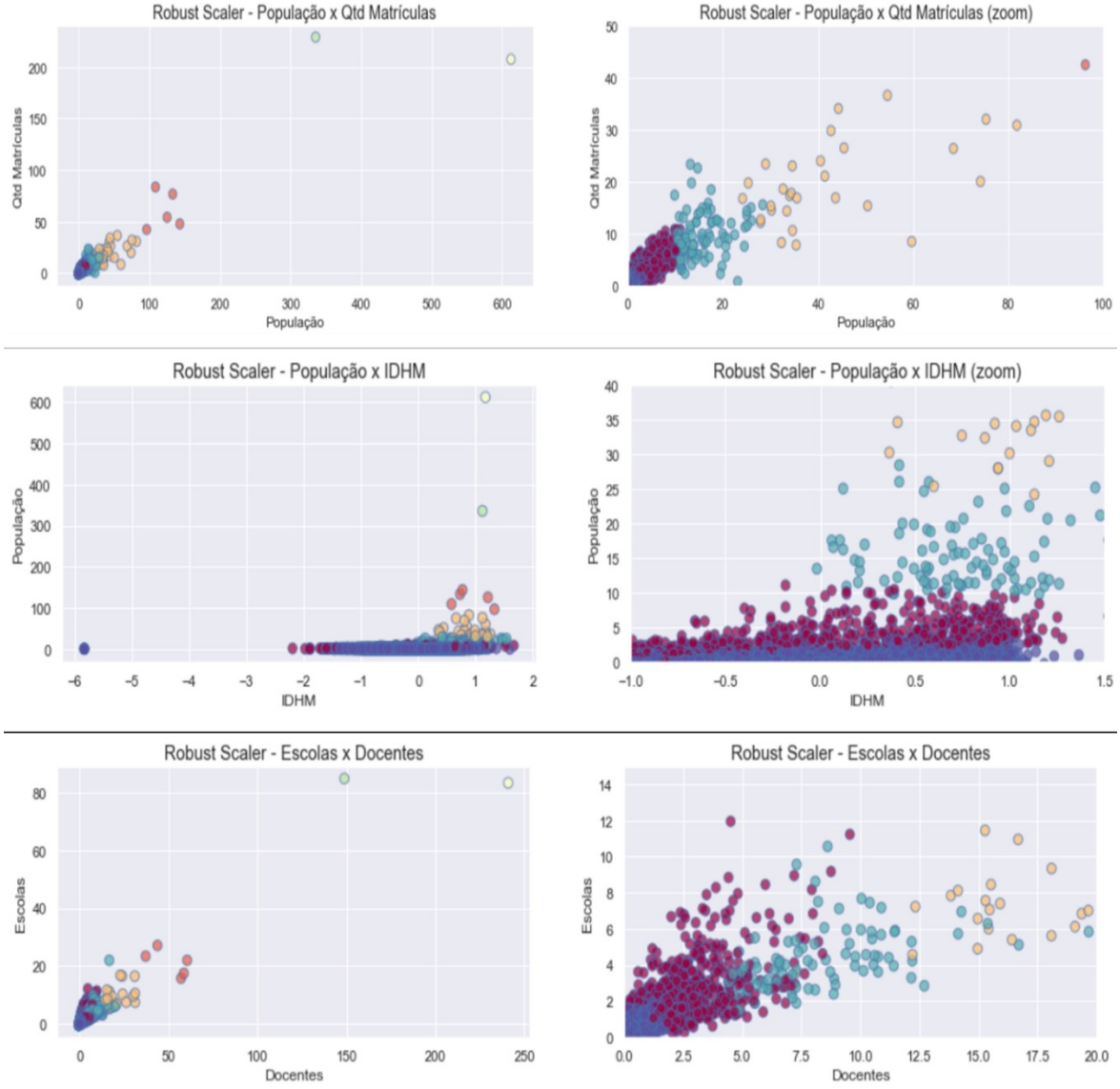
Fonte: Elaborada pelo autor (2020)

Os resultados apresentados comprovam:

- *K-Means:* o *RobustScaler* apresentou os melhores resultados – os dados são bem separados em ambos os lados do gráfico, embora haja uma pequena mistura dos pontos mais próximos de zero;
- *Agglomerative Clustering*: nota-se que o *RobustScaler* é a melhor alternativa entre todos os gráficos apresentados (conforme delineado na cor verde);
- DBSCAN: embora haja poucos grupos, o *StandardScaler* se mostrou o mais adequado para separar os dados desses grupos.

A figura abaixo mostram outros gráficos criados, com outras variáveis, com o uso do algoritmo *Agglomerative Clustering* e técnica de escalonamento *Robust Scaler*.

Figura 46 – Escalonamento de dados com *Aggl. Clustering* e *RobustScaler*

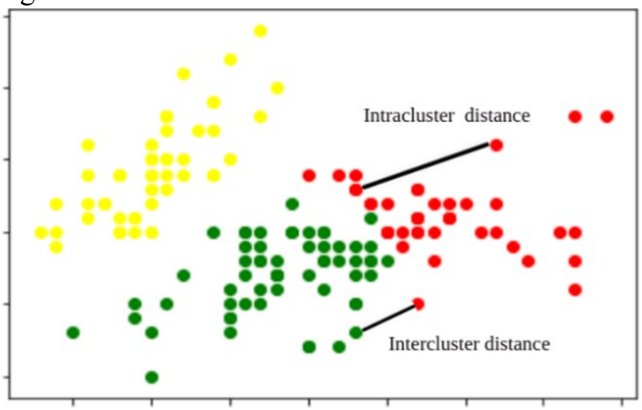


Fonte: Elaborada pelo autor (2020)

5.4.3 Clusterização *k-Means*

Dado o número k de clusters, *K-Means* particiona um conjunto de pontos em K grupos (ZAKI e MEIRA JR., 2014), de forma que cada ponto seja alocado ao grupo que lhe esteja mais próximo. Desta forma, os agrupamentos são criados com base nas distâncias mínimas entre os pontos pertencentes a cada cluster e seu centroide (ponto central do cluster). Um dos desafios em se utilizar o *K-Means* para agrupamentos é encontrar o número ideal de clusters – aquele que maximiza as diferenças entre clusters (inter-cluster) e minimiza as variações dentro de um clusters (intra-cluster), conforme se pode visualizar na figura abaixo.

Figura 47 – Distância inter-cluster e intra-cluster.

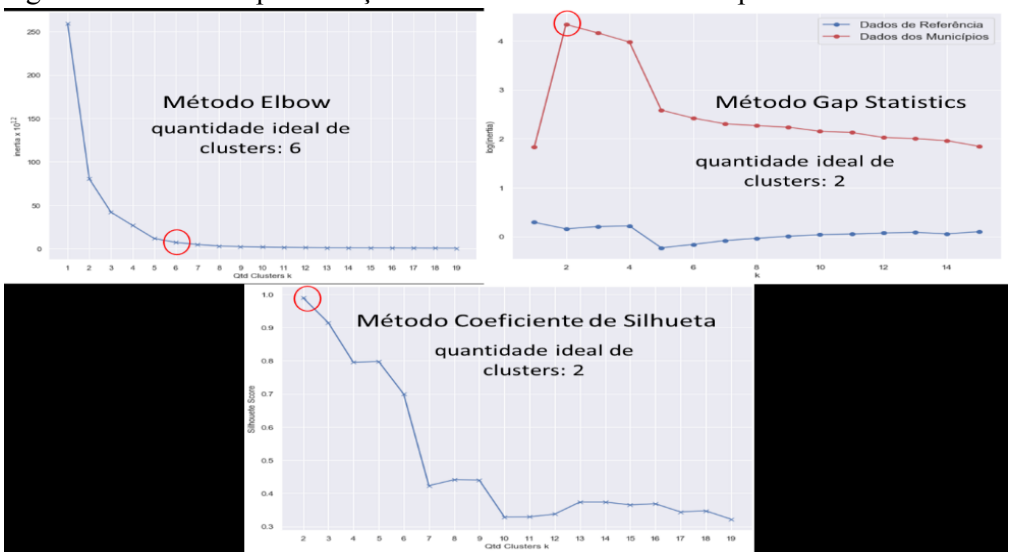


Fonte: SEN (2019)

Quando poucos clusters são formados, tem-se a maximização inter-cluster, mas as variações intra-cluster são prejudicadas, pois pontos muito distantes podem estar em um mesmo cluster. Por outro lado, ao se aumentar o número de clusters, as diferenças entre clusters se torna pequena, embora tem-se a vantagem de diminuir as variações em um cluster (intra-cluster). É preciso, portanto, achar o ponto ótimo, no qual os pontos de cada cluster sejam os mais mais homogêneos possíveis e que clusters formados sejam suficientemente diferentes um dos outros. Foram utilizados alguns métodos de seleção desse ponto ótimo com os dados dos municípios, a saber: *Elbow* (cálculo de *inertias*), *Gap Statistics* e *Coeficiente de Silhueta*. A

Figura 48 exibe os resultados encontrados em cada método.

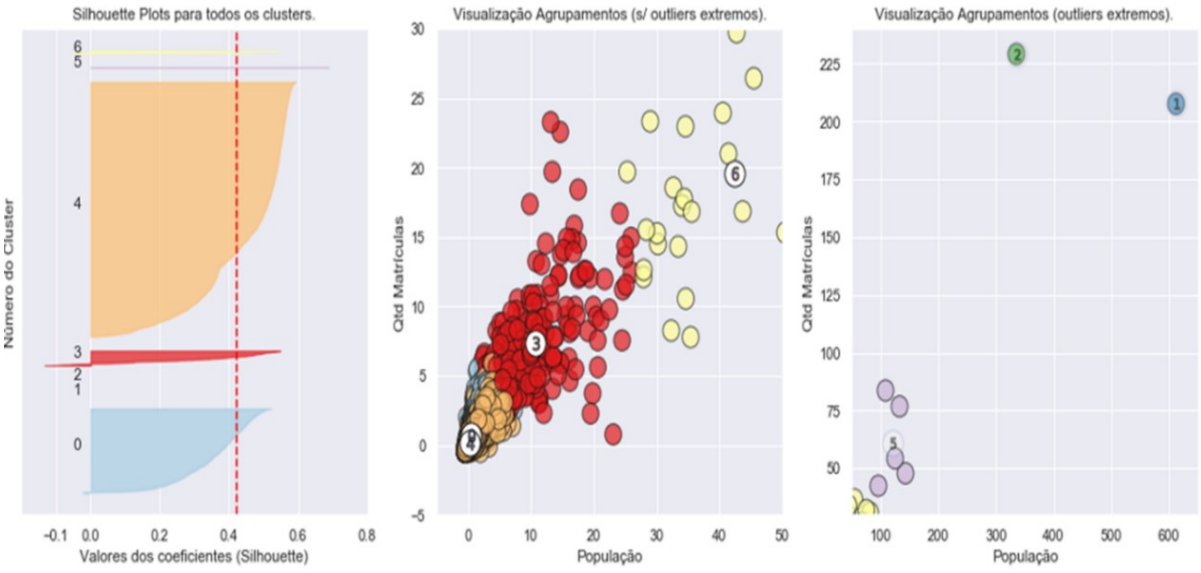
Figura 48 – Métodos para seleção do número ideal de clusters para *k-Means*.



Fonte: Elaborada pelo autor (2020)

Embora as sugestões variem, foi escolhido o valor de 7 clusters, pois 3 clusters já são alocados a poucos municípios anômalos. A Figura 49 mostra o gráfico de Análise de Silhueta (ALVES, 2019) para o valor de 7 clusters[52], utilizando-se população e número de alunos. Percebe-se que cada ponto extremo (lado direito) ficou em clusters diferentes (trata-se dos municípios de SP e RJ que são, de fato, distintos), enquanto pontos concentrados (lado esquerdo) foram adequadamente separados conforme faixas da população e número de alunos.

Figura 49 – Análise de Silhueta para clusterização k-Means com 7 clusters.



Fonte: Elaborada pelo autor (2020)

[52] O caderno *jupyter* "ClusterizacaoMunicipios" realiza a Análise de Silhueta para os valores de 2 a 9 clusters. Código disponível no APÊNDICE J.

Baseado na escolha do valor de 7 clusters, a Figura 50 apresenta os resultados da clusterização com o algoritmo *k-Means* e escalonamento *RobustScaler*, exibindo a quantidade de municípios em cada cluster, região e UF.

Figura 50 – Detalhes dos clusters com *k-Means* e escalonamento *RobustScaler*.

Cluster	Total	Total em %
4	3567	71.50
0	1180	23.65
3	206	4.13
6	29	0.58
5	5	0.10
2	1	0.02
1	1	0.02

Fonte: Elaborada pelo autor (2020)

Percebe-se algumas características da clusterização *k-Means*:

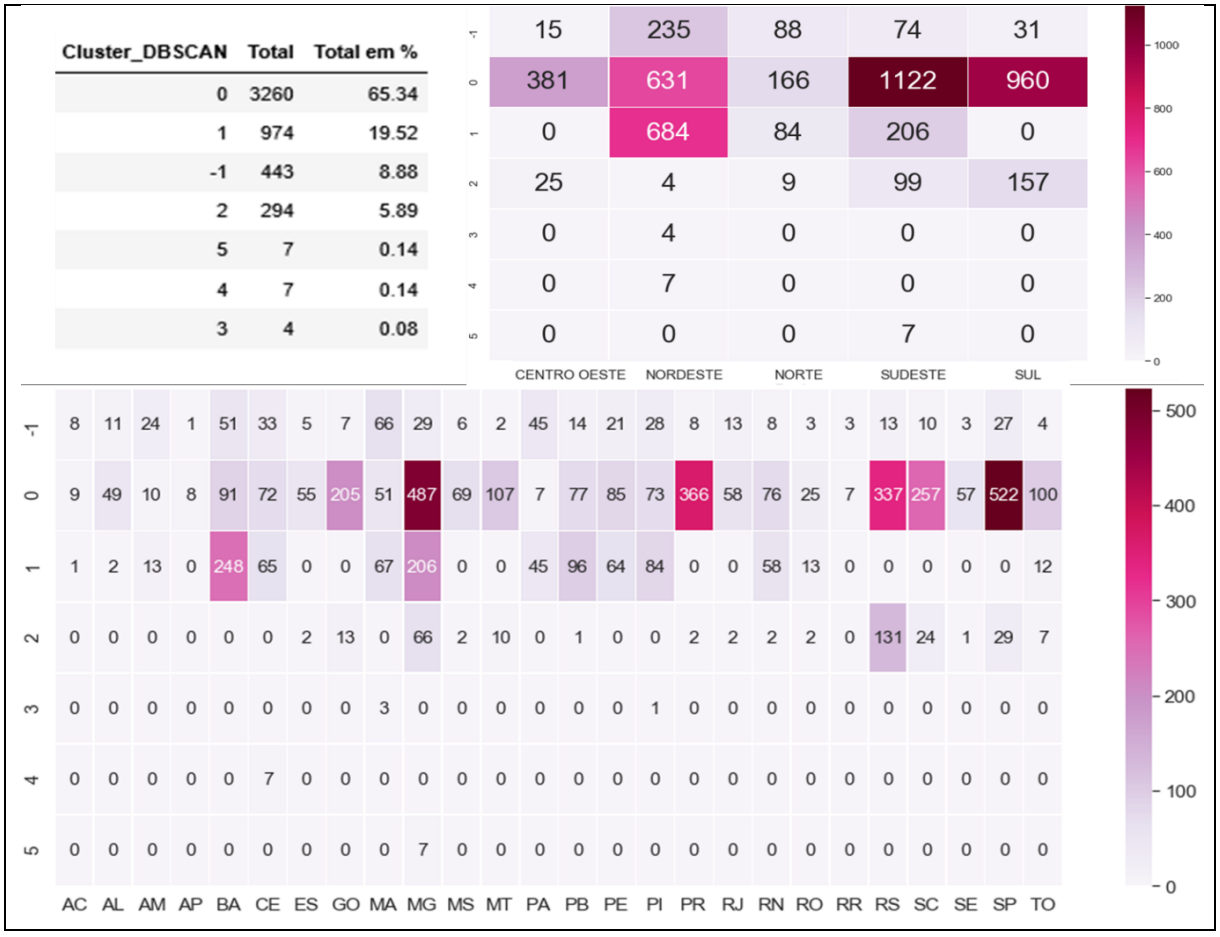
- são criados clusters específicos com municípios de valores extremos (municípios de população máxima, de maior quantidade de alunos ou de docentes);
- SP, de maior população, ficou em um cluster; enquanto RJ (metade da população de SP, mas maior quantidade de matrículas), ficou em outro cluster;
- ao se utilizar um valor K maior do que 10, são criados clusters específicos com municípios de valores mínimos (municípios com poucas escolas, por exemplo).
- geralmente ocorre a criação de clusters de poucos municípios (menos de 1% do total), que podem ser considerados anomalias de clusters.

5.4.4 Clusterização DBSCAN

Diferente de *k-Means*, que busca proximidade pela distância entre os pontos, DBSCAN usa a densidade local dos pontos – ou seja, utiliza o cálculo de áreas de maior e menor densidade de pontos – como critério de formação dos agrupamentos (ZAKI e MEIRA JR., 2014). Não requer o número de clusters e permite a definição de cluster irregulares.

No presente trabalho, convencionou-se por utilizar determinados parâmetros[53] de forma que não ocorresse muitos ruídos (pontos que não são alocados em nenhum cluster), mas que houvesse ao menos 6 clusters (exceto o cluster -1, composto de ruídos). Baseado nos parâmetros eps=0.9 e min_samples=5, a Figura 51 apresenta os resultados da clusterização com o algoritmo DBSCAN e escalonamento *StandardScaler*, exibindo a quantidade de municípios em cada cluster, região e UF.

Figura 51 – Detalhes dos clusters com DBSCAN e escalonamento *StandardScaler*.



Fonte: Elaborada pelo autor (2020)

[53] No caderno "Criacao_Tabela_DBSCAN", executou-se um código iterando entre os valores de eps (0.4 a 2.5) e min_samples (3 a 199), que foram salvas em tabela auxiliar, a fim de se obter a quantidade ideal de clusters e o número de elementos em cada cluster.

Percebe-se algumas características da clusterização DBSCAN:

- Os clusters não são bem separados como ocorre no *k-Means*, pois a separação de clusters considera faixas nas quais há grande densidade;
- os pontos extremos (municípios com população elevada, com poucos alunos ou poucos professores) não são alocados em algum cluster, mas inseridos no cluster *noise*. DBSCAN identifica os pontos *outliers*, ao contrário do *k-Means* (que os aloca em um cluster de tamanho pequeno);
- não é o algoritmo mais apropriado para a identificação de anomalias nas despesas, pois muitos pontos são alocados em poucos clusters (e o estudo requer alguns grupos semelhantes, e não grandes grupos de densidade similar).

5.4.5 Clusterização Hierárquica – *Agglomerative Clustering*

Algoritmos hierárquicos criam uma estrutura hierárquica de pontos aninhados conforme uma estratégia de agrupamento e um critério de ligação (métrica de dissimilaridade). As estratégias de agrupamento podem ser (ZAKI e MEIRA JR., 2014):

- aglomerativa (abordagem *bottow-up*): cada ponto é um cluster, e pares de clusters são mesclados sucessivamente à medida que se sobe na hierarquia;
- divisiva (abordagem *top-down*): todos os pontos estão em um cluster, e as divisões são executadas recursivamente à medida que se desce a hierarquia.

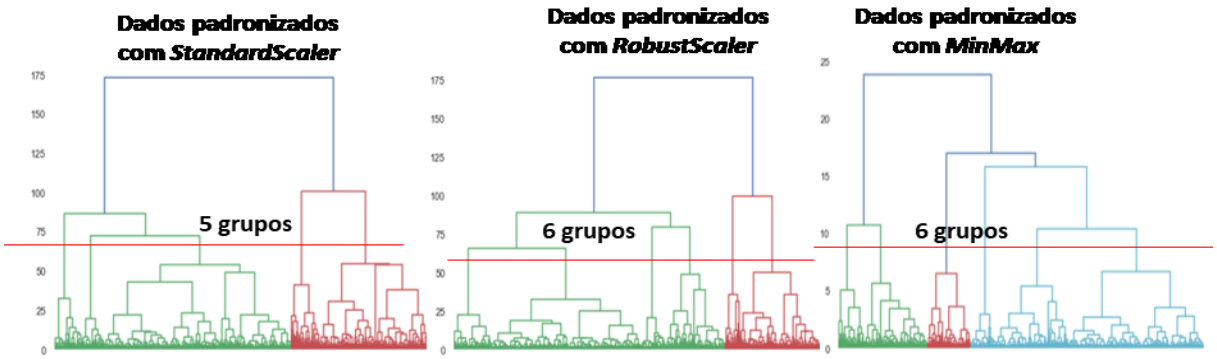
A dissimilaridade entre clusters é calculada conforme método escolhido no algoritmo:

- *Single*: a distância entre 2 clusters é dada pela menor distância entre dois pontos (distância mínima, vizinho mais próximo). É sensível a outliers;
- *Complete*: a distância entre 2 clusters é dada pela maior distância entre dois pontos (distância máxima, vizinho mais distante). Tende a gerar clusters muito grandes;
- *Average*: a distância entre 2 clusters é dada pela média das distâncias entre cada dois pontos (distância entre os centroides). É menos sensível a outliers, mas tende a gerar clusters grandes, globulares;
- *Ward*: Minimiza a soma das diferenças entre pontos nos clusters. Mais efetivo quando se tem outliers; tende a gerar clusters de tamanhos mais regulares.

Os resultados da clusterização hierárquica são normalmente mostrados como uma árvore de grupos ou dendogramas, a partir dos quais é possível obter o número de clusters. No presente trabalho, foi utilizado o algoritmo hierárquico aglomerativo. Conforme a Figura 52,

apenas para que fosse possível obter os dendogramas[54] abaixo, filtrou-se o *dataframe* para conter os registros de municípios com população de até 11 mil habitantes (50% dos dados) e com a utilização do método *Ward* nas situações especificadas na figura abaixo:

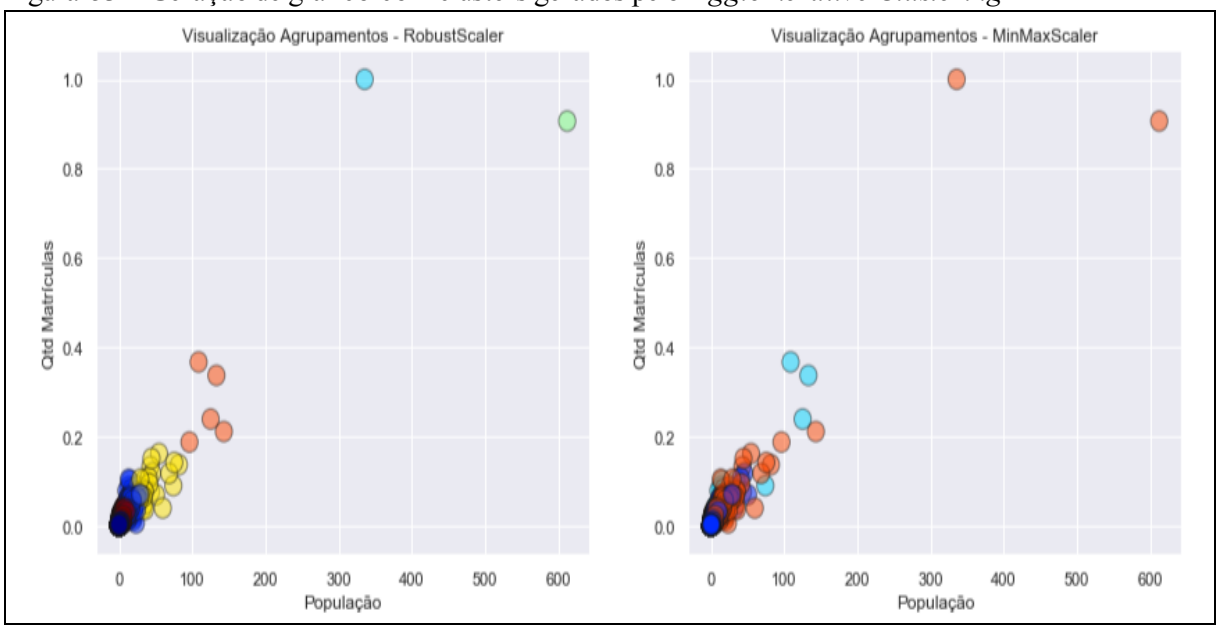
Figura 52 – Geração de dendogramas com método *Ward*.



Fonte: Elaborada pelo autor (2020)

Apesar das sugestões indicadas pelos dendogramas, seguiu-se a escolha de 7 clusters. A Figura 53 mostra o gráfico[55] resultante da aplicação do *Agglomerative Clustering* com 7 clusters, com as variáveis população e número de alunos. Claramente, a utilização do *RobustScaler* traz uma boa separação dos clusters se comparado com *MinMaxScaler*.

Figura 53 – Geração de gráfico com clusters gerados pelo *Agglomerative Clustering*



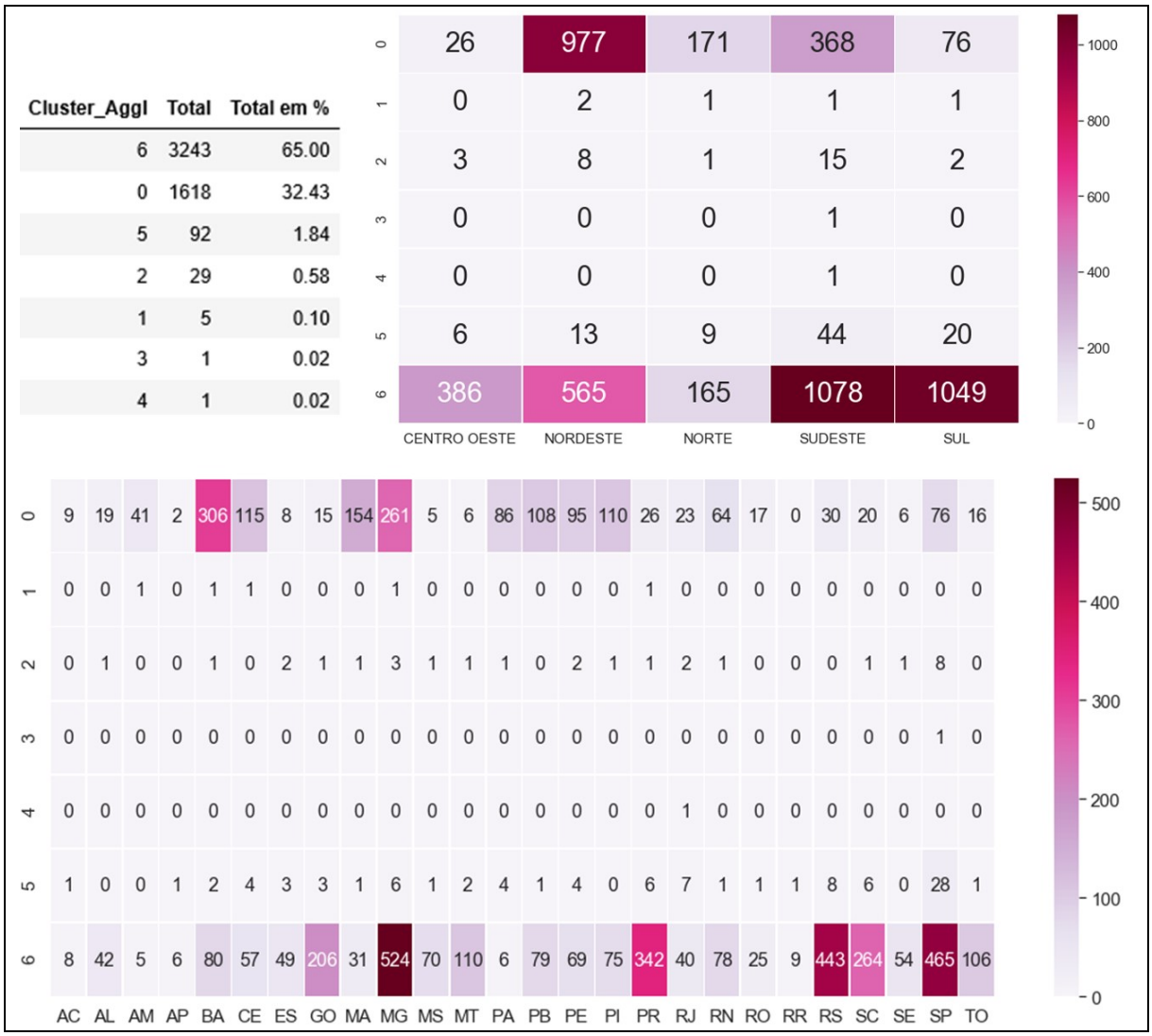
[54] O caderno *jupyter* "ClusterizacaoMunicipios" cria diversos dendrogramas, dependendo das amostras de dados, do método de escalonamento dos dados e da medida de dissimilaridade. No presente trabalho, é apresentado apenas um dendograma (em apenas uma amostra de 50% do total de municípios, que apresentampopulação de até 11 mil habitantes, escolhida aleatoriamente) para que fosse possível a visualização gráfica deste dendograma.

[55] O caderno *jupyter* "ClusterizacaoMunicipios" cria gráficos comparativos da aplicação do *Agglomerative Clustering* nos valores normalizados no *RobustScaler* e *MinMaxScaler*, para os valores de 2 a 7 clusters.

Fonte: Elaborada pelo autor (2020)

A Figura 54 apresenta os resultados da aplicação do *Agglomerative Clustering* com *RobustScaler*, exibindo a quantidade de municípios em cada cluster, região e UF.

Figura 54 – Detalhes dos clusters com *Aggl. Clustering* e escalonamento *RobustScaler*.



Cluster_Aggl	Total	Total em %
6	3243	65.00
0	1618	32.43
5	92	1.84
2	29	0.58
1	5	0.10
3	1	0.02
4	1	0.02

Fonte: Elaborada pelo autor (2020)

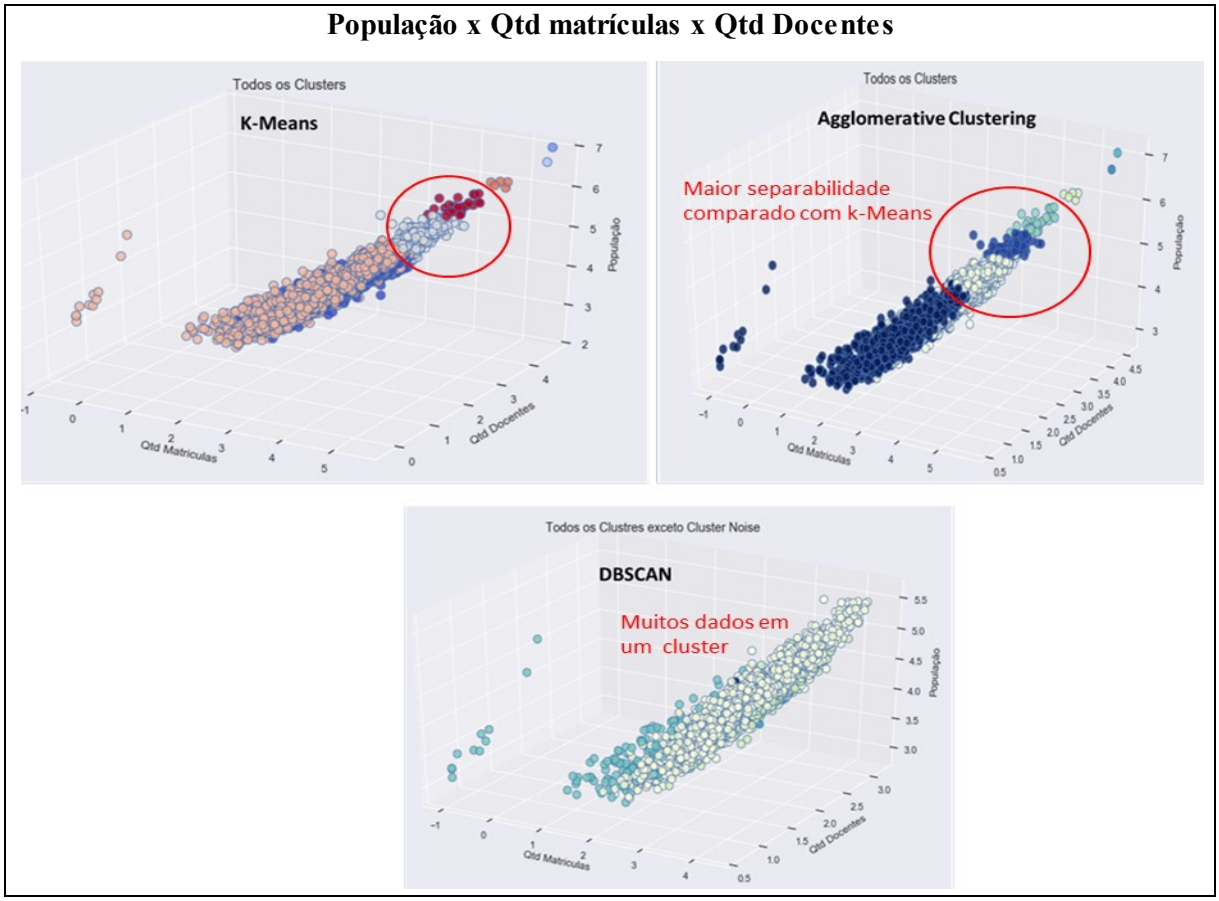
Embora algoritmos hierárquicos sejam simples e eficazes, com facilidade na visualização dos clusters em dendogramas e flexibilidade na escolha do corte para determinar o número de clusters, existem alguns dificultadores. Similar ao que ocorre no *k-Means*, a escolha do número de clusters não é trivial – deve-se decidir onde passar a linha no dendograma para se obter o número de clusters; e, além disso, métodos de dissimilaridade podem produzir resultados bem diferentes, e não há critério objetivo para avaliar qual o melhor método.

5.4.6 Validação dos algoritmos de clusterização

Alguns gráficos produzidos no caderno *jupyter*[56] (com variáveis normalizadas em log) foram escolhidos para serem apresentados no presente trabalho, com o objetivo de comparar os resultados de alguns algoritmos utilizados. Essa comparação não só procura demonstrar a coerência da separabilidade dos clusters, mas orientou a escolha do algoritmo para a tarefa de detecção de anomalias.

A Figura 55 mostra os clusters formados para os algoritmos *k-Means*, DBSCAN e *Agglomerative Clustering* (AgrC), considerando os dados de população, quantidade de alunos matriculados e quantidade de docentes. Os grupos formados pelo k-Means e AgrC são bem parecidos, mas o AgrC apresentou uma melhor separabilidade em áreas densas; o DBSCAN já não é muito adequado, pois agrupou muitos pontos em um mesmo grupo e inseriu os pontos extremos no cluster *noise* (que não aparecem no gráfico).

Figura 55 – Comparação dos resultados de alguns algoritmos de clusterização

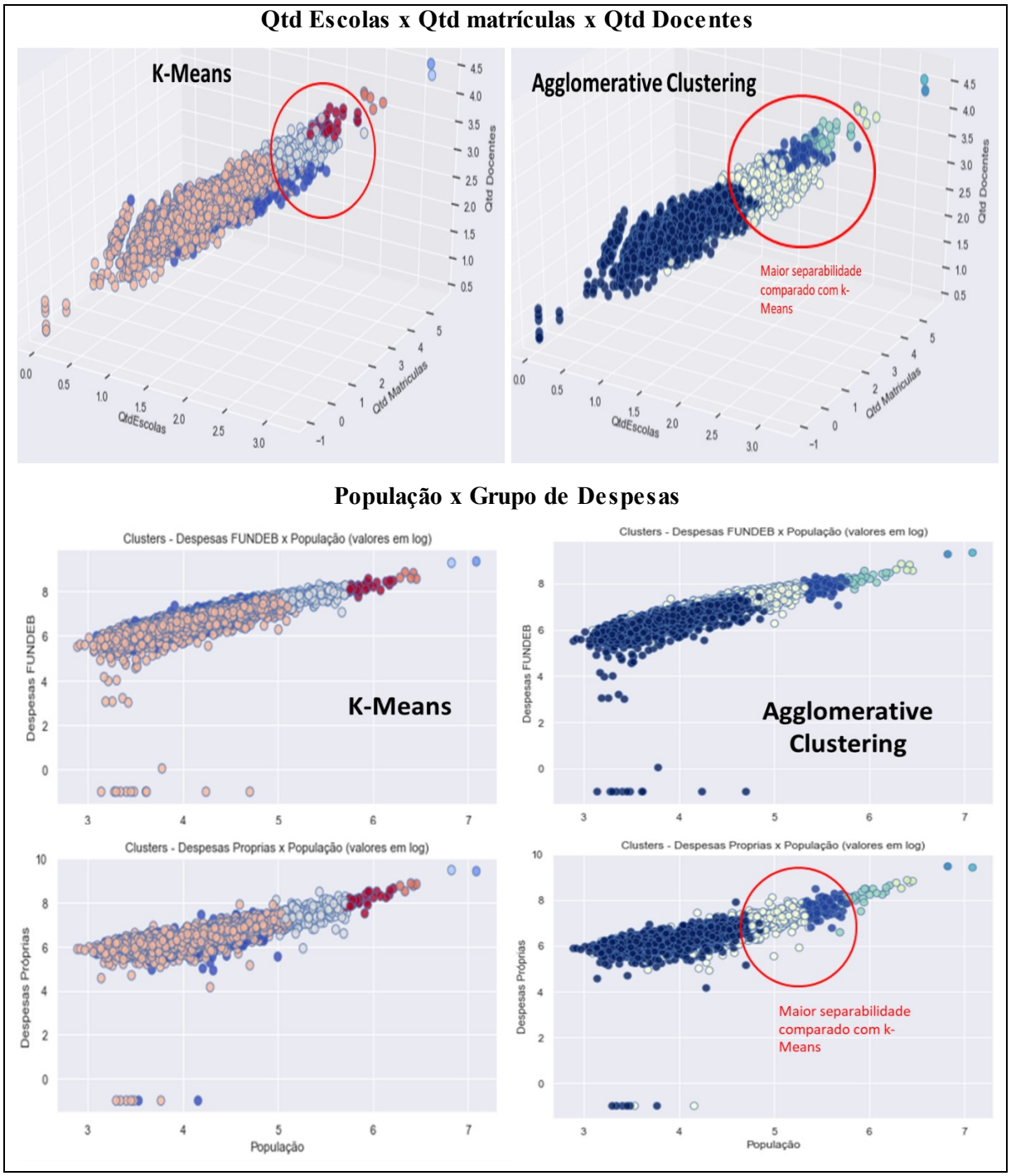


[56] O caderno *jupyter* "ClusterizacaoMunicipios" cria uma diversidade de gráficos 2D e 3D, com diferentes variáveis (população; indicadores IDHM e notas IDEB; quantidade de alunos, de escolas e de professores; e mesmo algumas despesas e contas contábeis), para a visualização dos resultados dos agrupamentos – para o presente trabalho, convencionou-se a selecionar apenas os atributos de população, quantidade de alunos e de professores.

Fonte: Elaborada pelo autor (2020)

As figuras seguintes comparam apenas os resultados dos algoritmos k-Means e *Agglomerative Clustering* (AgrC).

Figura 56 – Comparação dos resultados de alguns algoritmos de clusterização



Fonte: Elaborada pelo autor (2020)

De fato, o algoritmo com os melhores resultados, considerando os objetivos do presente trabalho e a geração dos gráficos com diversas variáveis (disponíveis nos cadernos *jupyter*), é o *Agglomerative Clustering* com dados escalonados com *RobustScaler*.

Em virtude da escolha pelo *Agglomerative Clustering*, foi utilizada, no caderno *jupyter*, uma medida da avaliação da qualidade da separação entre os clusters – o índice *Davies-Bouldin,* que compara a distância entre os clusters com o tamanho dos próprios clusters (DAVIES e BOULDIN, 1979). É utilizada principalmente quando os dados não são rotulados. Um valor mais próximo de zero significa um modelo com melhor separação entre os clusters.

Os melhores índices foram alcançados com 2 e 5 clusters (abaixo de 0,4); e o índice não foi satisfatório para 7 clusters (0,69). Entretanto, manteve-se este valor, pois é de interesse para o presente trabalho um maior número de grupos para a detecção de anomalias locais. Se há poucos grupos, poucos pontos anômalos serão identificados.

5.5 DETECÇÃO DE ANOMALIAS

A detecção de *outliers* ou anomalias tem como objetivo a identificação de eventos ou itens inesperados no conjunto de dados, que diferem da norma (GOLDSTEIN e UCHIDA, 2016). De modo geral, tais valores discrepantes são removidos antes do uso de algoritmos de mineração de dados. Para o presente trabalho, porém, o objetivo é, de fato, a detecção de despesas discrepantes em um dado grupo de município similares, que podem indicar falhas de preenchimento de registros, possíveis irregularidades sendo praticadas ou eventos atípicos que devem ser justificados (uma eventual obra em uma escola, por exemplo).

Neste sentido, o foco é o uso de algoritmos de detecção de anomalias não supervisionados (quando rótulos são desconhecidos), que usam apenas as informações intrínsecas dos dados para detectar os pontos que se desviam dos demais. Procurou-se utilizar algoritmos baseados em diferentes critérios de detecção - distância, similaridade e densidade.

5.5.1 Delimitação das estratégias para detecção de outliers

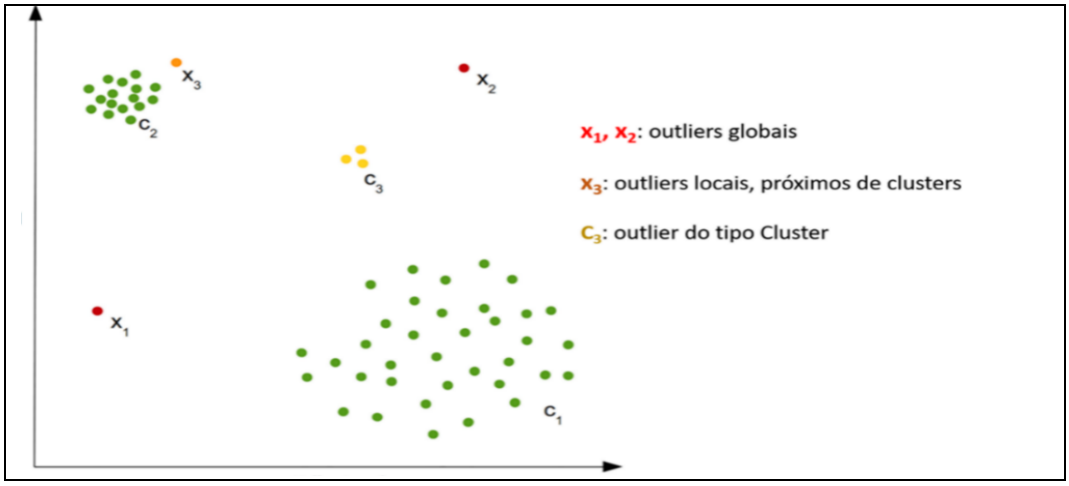
A AED permitiu a identificação de *outliers* globais, de valores atípicos quando comparados com o conjunto de dados. Entretanto, há o interesse de se detectar outros tipos de *outliers*, definidos na Tabela 11 e representados na Figura 57.

Tabela 11 – Os tipos de anomalias

Globais	Valores extremos (de baixo ou alto valor) muito diferentes dos demais pontos, facilmente detectáveis por técnicas estatísticas, histogramas e gráficos de dispersão.
Clusters	Poucos pontos juntos em um cluster, bem distantes de outros clusters mais densos, detectáveis pela aplicação de algoritmos de clusterização, como *k-Means*.
Locais	Poucos pontos próximos aos outros conjuntos densos, detectáveis por algoritmos específicos de detecção de *outliers*, como LOF. Esses pontos parecem normais, e são percebidos quando se analisa, isoladamente, um determinado cluster.

Fonte: Elaborada pelo autor (2020), adaptado de GOLDSTEIN e UCHIDA (2016).

Figura 57 – Representação gráfica dos tipos de anomalias



Fonte: Elaborada pelo autor (2020), adaptado de GOLDSTEIN e UCHIDA (2016).

Algoritmos de detecção de anomalias geram dois resultados: um rótulo indicando se uma instância é anômala ou não; e uma pontuação (*score*) que indica o grau de anormalidade (GOLDSTEIN e UCHIDA, 2016). Em algoritmos não supervisionados, as pontuações são mais comuns e permitem a classificação (*ranking*) dos pontos mais anômalos. Por meio de um limite (*threshold*), a classificação pode ser convertida em um rótulo. No presente trabalho, todos esses indicadores (rótulo, pontuação, limite e ranqueamento) foram criados e armazenados.

5.5.2 A biblioteca *Python Outlier Detection* (PyOD)

Para identificar as despesas discrepantes dos municípios, foi utilizada a biblioteca *Python Outlier Detection* (PyOD)[57], que conta com uma variedade de modelos para a detecção de anomalias em dados multivariados (ZHAO, NASRULLAH e LI, 2019). A Figura 58 apresenta gráficos de alguns algoritmos da PyOD, nos quais os pontos pretos representam *outliers*. A Tabela 12 lista os algoritmos escolhidos que foram aplicados no presente trabalho.

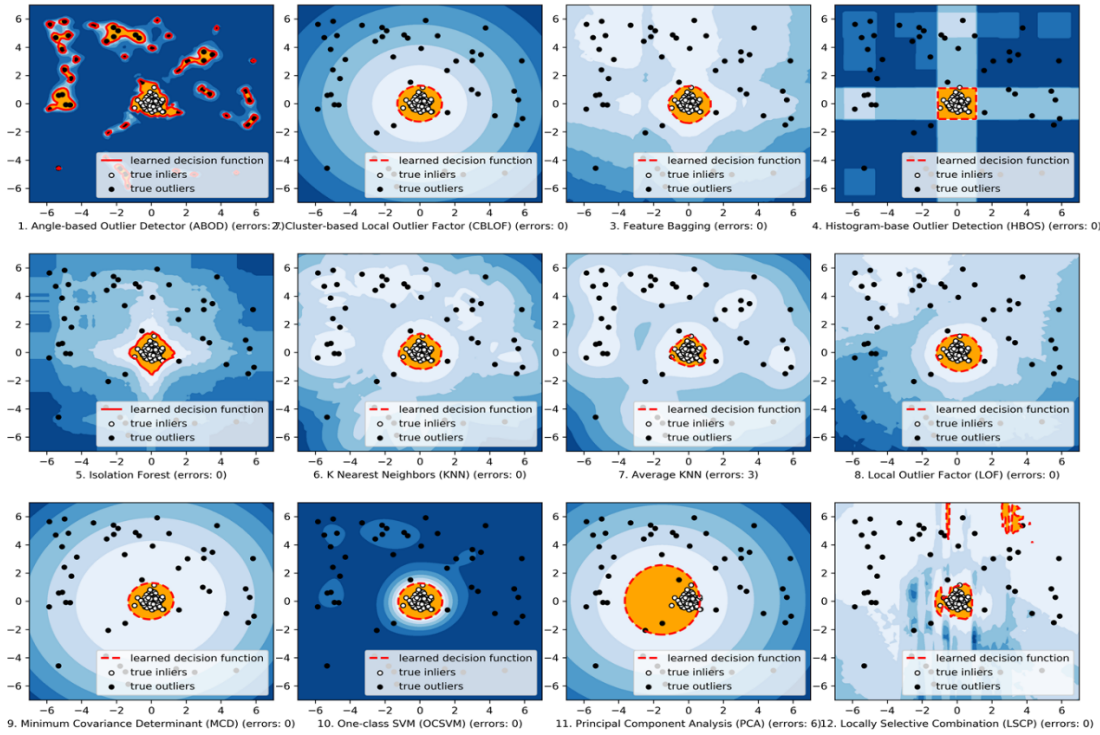
Tabela 12 – Algoritmos de detecção de anomalias utilizados nos dados

Angle-based Outlier Detection (ABOD)	Probabilístico
Local Outlier Factor (LOF)	Baseado em distância
Cluster-based Local Outlier Factor (CBLOF)	Baseado em distância/ densidade
Histogram-based Outlier Detection (HBOS)	Baseado em estatística
K Nearest Neighbors (KNN)	Baseado em distância
Average KNN (A_KNN)	Baseado em distância
Isolation Forest (IF)	*Ensemble* (agrupamento de recursos)
Feature Bagging (FB)	*Ensemble* (agrupamento de recursos)

Fonte: Elaborada pelo autor (2020), adaptado de ZHAO, NASRULLAH e LI (2019)

Figura 58 – Listagem de alguns modelos disponíveis na biblioteca PyOD

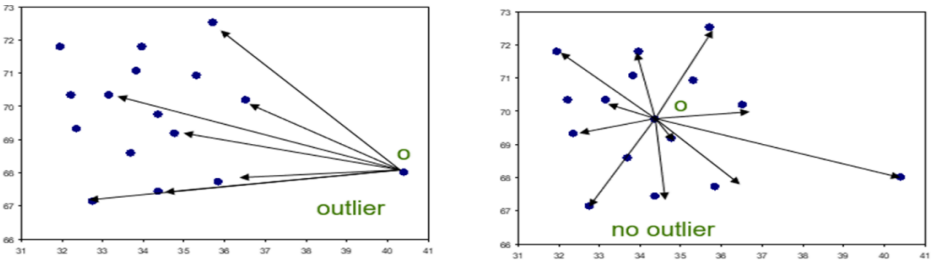
[57] Documentação sobre a biblioteca PyOD pode ser vista em https://pyod.readthedocs.io/en/latest/.



Fonte: ZHAO, NASRULLAH e LI (2019)

O algoritmo ABOD tem como diferencial o uso de medidas baseadas em ângulos. Em espaços multidimensionais, os ângulos são medidas mais estáveis do que as medidas de distância, e um bom exemplo são as medidas de similaridade baseadas em cosseno para análise de textos (KRIEGEL et al, 2008). Conforme a Figura 59, um ponto no espaço é considerado um *outlier* se a maioria dos outros pontos estiver localizada em direções semelhantes.

Figura 59 – Detecção de anomalias no algoritmo ABOD



Fonte: KRIEGEL et al (2008)

O algoritmo LOF foi o primeiro a introduzir a ideia de identificar as anomalias locais (GOLDSTEIN e UCHIDA, 2016). A pontuação da anomalia é baseada numa medida de densidade local e o desvio dessa medida entre um ponto e a dos seus vizinhos. As instâncias normais, com densidades similares de seus vizinhos, contêm uma pontuação próxima de 1; instâncias anômalas, com densidade local substancialmente inferior que a dos seus vizinhos, possuem pontuações maiores. Requer um número para k (quantidade de vizinhos).

O algoritmo CBLOF usa clusterização para determinar as áreas densas nos dados, criar clusters e calcular a sua estimativa de densidade (GOLDSTEIN e UCHIDA, 2016). A pontuação da anomalia de um ponto se baseia no tamanho do cluster ao qual ele pertence (parâmetro desativado por padrão) e na distância do maior cluster mais próximo. Consequentemente, todos os pontos em clusters pequenos, distantes dos clusters maiores, são anômalos (HE et al, 2003) – assim, localiza anomalias globais e de cluster, mas não as locais.

O algoritmo HBOS assume que variáveis são independentes, e calcula o grau de anomalia de uma instância de dado através de histogramas (GOLDSTEIN e UCHIDA, 2016). É considerado um algoritmo simples e de rápida execução, mas com menor precisão.

O algoritmo KNN calcula a pontuação de anomalia com base na distância de um ponto ao k-ésimo vizinho mais próximo, sendo três métodos possíveis de cálculo – a maior distância, a média (A-KNN) ou mediana das distâncias de todos os k-ésimos vizinhos próximos (ZHAO, NASRULLAH e LI, 2019). Detecta, portanto, as anomalias globais, não as locais.

O algoritmo *Isolation Forest* realiza o particionamento de dados usando uma estrutura de árvores. A pontuação de anomalia considera o quanto isolado um dado ponto está na estrutura, quando poucas partições são necessárias para seu isolamento.

O algoritmo *Feature Bagging* utiliza vários algoritmos de detecção (podendo ser LOF, kNN e ABOD) em diferentes amostras de um conjunto de dados multivariados, e utiliza medidas de média, ou outras métricas combinadas, para o cálculo da pontuação da anomalia (ZHAO, NASRULLAH e LI, 2019). É uma técnica que procura reduzir a variação entre os algoritmos, a fim de melhorar a eficácia e evitar sobreajustes (*overfitting*).

5.5.3 Escolha de Cluster para ser submetido aos algoritmos

Com base nos resultados do *Agglomerative Clustering*, escolheu-se o cluster 6, que foi submetido aos oito algoritmos de detecção escolhidos, em diferentes escopos de dados[58]: todos os atributos (grupos de despesas, tipos de gasto, programas, subfunções e contas contábeis); despesas Próprias e FUNDEB; e tipos de gasto de remuneração e manutenção. As pontuações de anormalidade e os rótulos de cada município, produzidos por cada algoritmo de detecção, em cada escopo de dados, foram armazenadas no *dataframe* de municípios[59].

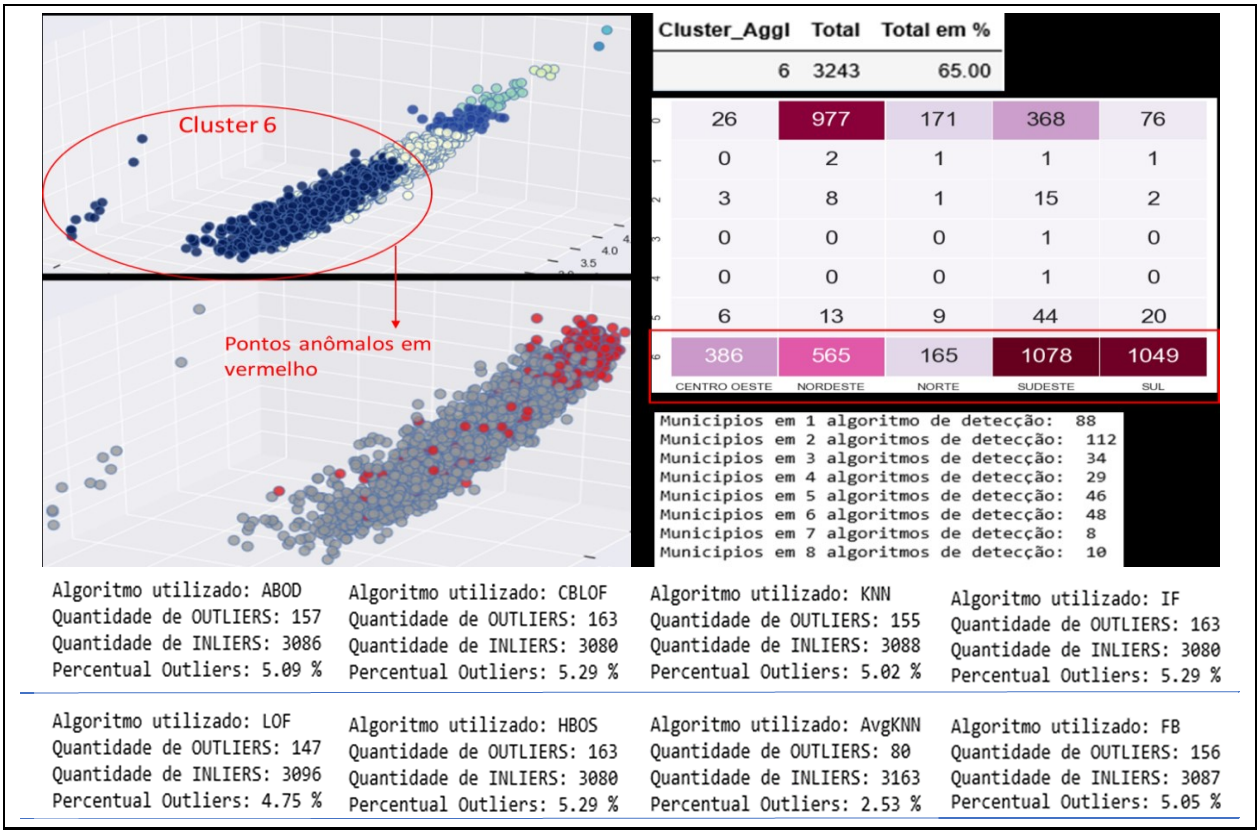
[58] Dados do IBGE, INEP e PNUD não foram considerados para a detecção, pois já foram utilizados para a clusterização de municípios semelhantes. Além disso, o foco é a detecção de anomalias em despesas com educação, não em características dos municípios.

[59] O caderno *jupyter* "DetecçãoOutliers_PyOD" disponibiliza todas as informações (rótulos e pontuações) referentes às execuções dos algoritmos de detecção, inclusive gráficos nos quais é possível visualizar as anomalias apontadas por cada algoritmo.

5.5.4 Resultados da detecção de anomalias

A Figura 60 exibe as características do cluster escolhido, o gráfico de dispersão com a indicação das anomalias e a quantidade de municípios anômalos em cada algoritmo, no escopo de todos os atributos.

Figura 60 – Indicação e quantidades dos municípios anômalos (escopo: todos atributos)



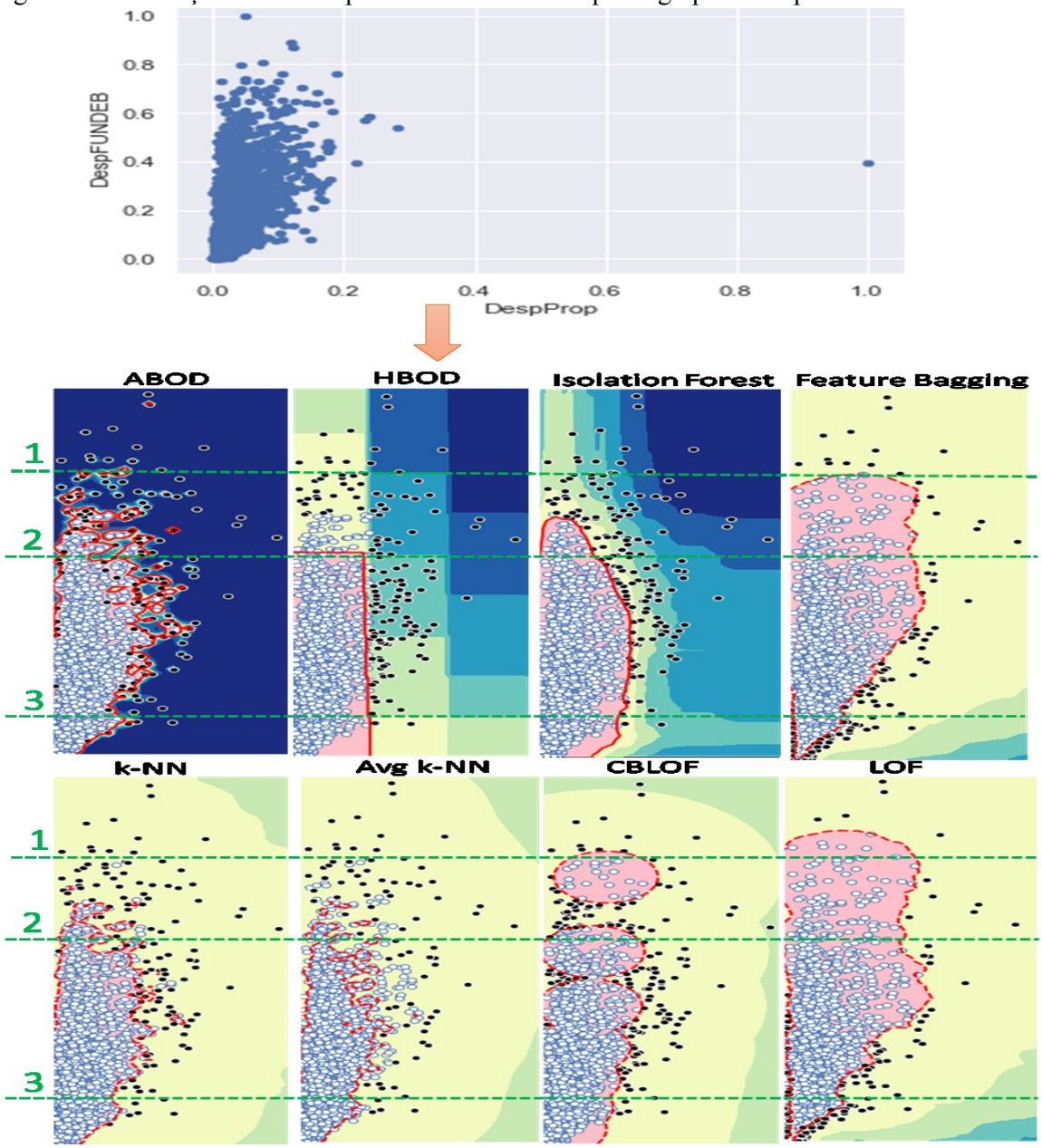
Fonte: Elaborada pelo autor (2020)

A Figura 61 apresenta o gráfico de dispersão das Despesas Próprias com as Despesas FUNDEB, e logo abaixo exibe os resultados dos algoritmos no escopo dos grupos de despesa (ou seja, a entrada para os algoritmos são apenas os dados das despesas Próprias e das despesas FUNDEB)[60]. Todos os algoritmos (exceto LOF) consideraram os pontos extremos (alto valor das despesas), acima da faixa pontilhada 1, como anômalos. Entre as faixas pontilhadas 1 e 2, poucos algoritmos (ABOD, CBLOF, kNN) detectaram como anômalos alguns pontos mais internos, localizados entre regiões normais. Abaixo da faixa pontilhada 3, os algoritmos LOF e FB detectaram como anômalos uma grande concentração de pontos bem próximos aos pontos normais[61].

[60] Código *python* disponível no APÊNDICE K.

[61] As faixas pontilhadas (em verde) na Figura 61 foram desenhadas no documento apenas para facilitar a visualização gráfica das constatações descritas a respeito dos pontos anômalos.

Figura 61 – Indicação dos municípios anômalos no escopo do grupo de despesas



Fonte: Elaborada pelo autor (2020), adaptado de ZHAO, NASRULLAH e LI (2019)

Algumas considerações se fazem necessárias:

- a linha vermelha representa o *threshold* (limite) da pontuação do grau de anomalia, estabelecido estatisticamente (*score* no percentil de 5%[62]) - todos os pontos acima da linha são classificados como anomalias;
- as seis camadas de cores representam faixas de valores da pontuação de anomalia, sendo que a primeira camada se inicia após o valor limite.

[62] A escolha do percentil de 5% como limite foi meramente aleatória, apenas para indicar, graficamente, que é possível ao usuário estabelecer este limite.

5.5.5 Validação dos modelos de detecção de anomalias

É necessário aferir a confiabilidade dos modelos gerados. Há diversas métricas conhecidas para a validação de algoritmos supervisionados, como acurácia, precisão, matriz de confusão e validação cruzada. No caso dos algoritmos de detecção de anomalias, os métodos tradicionais para avaliar a qualidade das pontuações de anormalidade (*scoring*) são a curva ROC (*Receiver Operating Characteristic*) e a curva PR (*Precision-Recall*) – mas apenas quando os rótulos (classes) estão disponíveis (GOIX, 2016).

No presente trabalho, porém, não há rótulos – ou seja, não há dados reais sobre municípios que apresentaram despesas discrepantes em educação. Desta forma, a validação dos modelos gerados, de forma não supervisionada, não é uma tarefa trivial. Uma alternativa viável foi a comparação das estatísticas e gráficos entre o conjunto de dados normais e os anômalos – para averiguar se tais anormalidades são consistentes.

A execução dos 8 algoritmos, em todo o conjunto de dados, apontou 375 municípios anômalos (citados ao menos em um algoritmo), ou 11,5% do total de municípios, sendo 10 municípios anômalos em comum (citados por todos os algoritmos), listados na Figura 62.

Realizou-se uma análise simplificada nesses dez municípios anômalos. Conforme estatísticas e histogramas de população e quantitativos de escolas, alunos e professores na Figura 63[63] – são municípios que se encontram acima da faixa do percentil de 75% quando comparados com o conjunto todo. Além disso, os demais histogramas de despesas, de algumas funções e de algumas contas contábeis, nas Figuras 64 e 65, apresentaram curvas mais deslocadas à direita – que podem ter influenciado na pontuação da anormalidade para certos municípios.

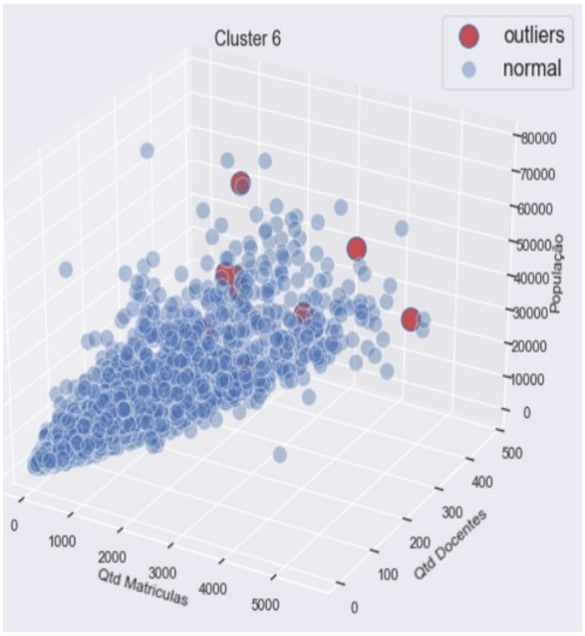
Em seguida, escolheu-se o município de Bombinhas para uma análise ainda mais detalhada (por ter a menor população), conforme a Figura 66. Comparando-se alguns histogramas de Bombinhas com todos os demais municípios, nota-se claramente que há despesas de valores atípicos quando comparadas com as mesmas despesas de seus semelhantes (os valores para Bombinhas estão representados pela linha vermelha vertical, localizadas ao final das curvas que representam os intervalos dos demais municípios). Tais valores devem ser apresentados aos órgãos de controle para as devidas investigações ou providências.

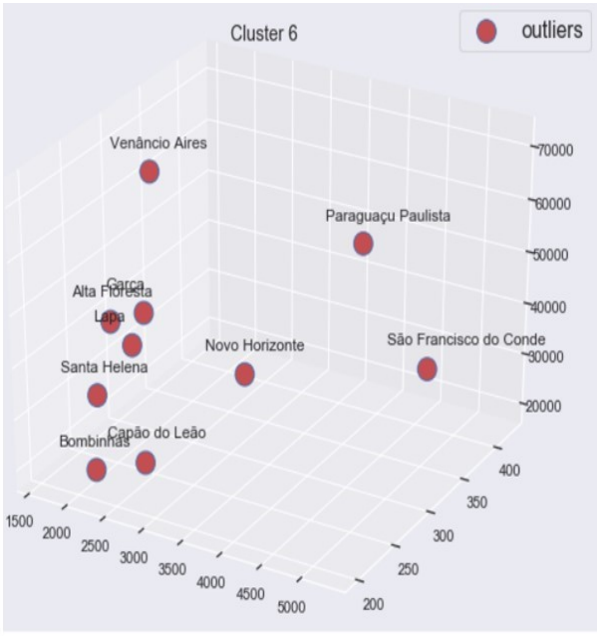
Pode-se dizer, desta forma, que os modelos gerados são válidos e permitem a identificação de fatos relevantes.

---

[63] Código *python* disponível no APÊNDICE L.

Figura 62 – Os 10 municípios anômalos em todos os algoritmos de detecção

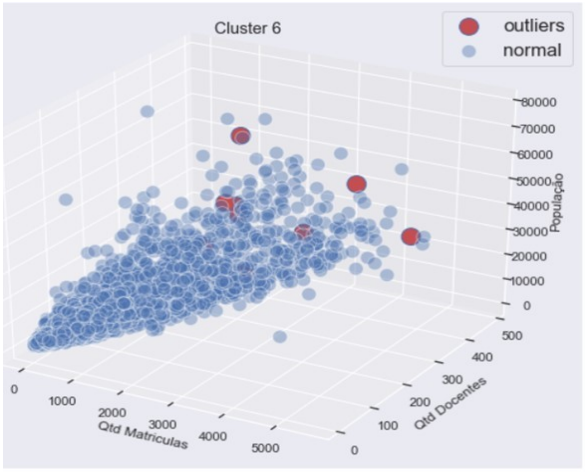


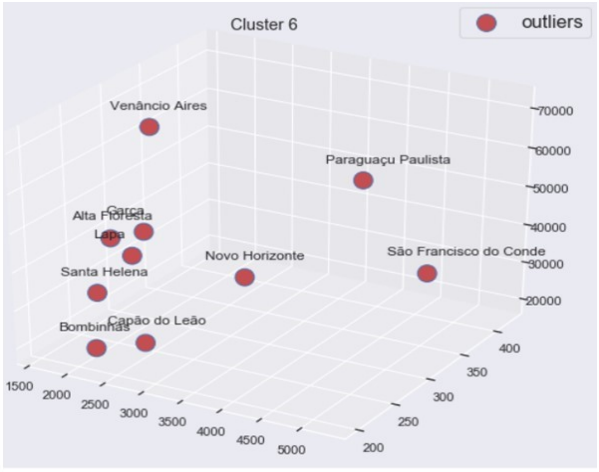


SigUF	NomeMunicipio	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361	DespFUNDEB	DespVinc	DespProp	tgRemun	tgManut
SC	Bombinhas	19769	11	213	2102	6964060.50	4684067.90	8958981.20	12854264.40	1857706.80
RS	Capão do Leão	25354	13	206	2825	17223486.90	2453994.70	3126054.70	17824240.80	2542848.40
PR	Santa Helena	26490	17	261	1628	8179828.60	10247410.10	4392123.20	13582136.70	7906225.20
BA	São Francisco do Conde	39802	50	323	5288	11535926.70	881183.50	79260447.90	80184416.20	9554227.50
SP	Novo Horizonte	41052	21	248	3691	15192903.70	5539053.60	6177325.00	16685934.40	3971600.70
SP	Garça	44390	25	267	2211	11789378.10	3930066.50	7062476.10	16775069.50	2577109.40
SP	Paraguaçu Paulista	45703	24	423	3587	15898405.40	3986312.90	7336933.90	17539570.30	5233980.20
PR	Lapa	48163	31	201	2738	12753182.10	3036825.20	9656951.70	17016054.40	5073990.60
MT	Alta Floresta	51782	19	199	2491	10673361.60	4575425.50	4919555.80	13954297.20	3654192.30
RS	Venâncio Aires	71554	37	266	2334	16587104.50	4433957.50	4664847.80	19133679.30	1055143.60

Fonte: Elaborada pelo autor (2020)

Figura 63 – Os 10 municípios anômalos – Comparação dos dados IBGE e INEP



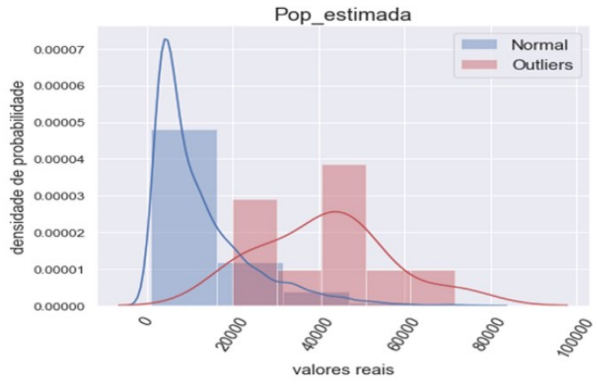


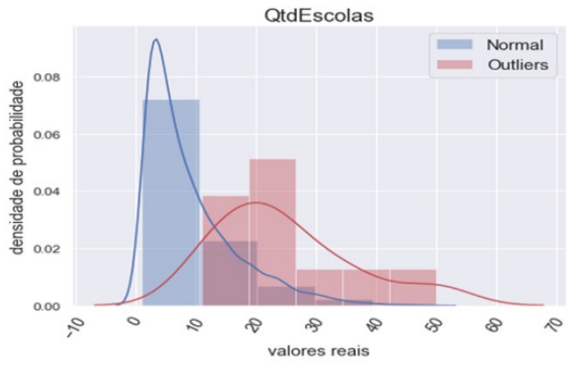
Estatísticas dos municípios não anômalos

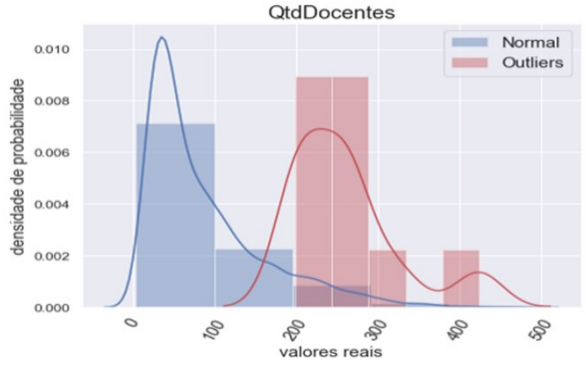
	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361
count	3233.00	3233.00	3233.00	3233.00
mean	12363.98	9.01	87.68	936.48
std	11164.00	7.53	72.89	908.83
min	781.00	1.00	4.00	0.00
25%	4377.00	3.00	35.00	298.00
50%	8347.00	6.00	61.00	587.00
75%	16684.00	12.00	118.00	1242.00
max	78013.00	49.00	482.00	5622.00

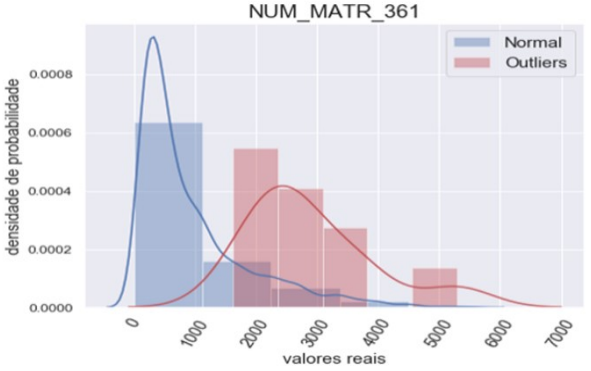
Estatísticas dos municípios anômalos

	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361
10.00	10.00	10.00	10.00	
41405.90	24.80	260.70	2889.50	
15057.67	11.84	69.19	1056.62	
19769.00	11.00	199.00	1628.00	
29818.00	17.50	207.75	2241.75	
42721.00	22.50	254.50	2614.50	
47548.00	29.50	266.75	3396.50	
71554.00	50.00	423.00	5288.00	



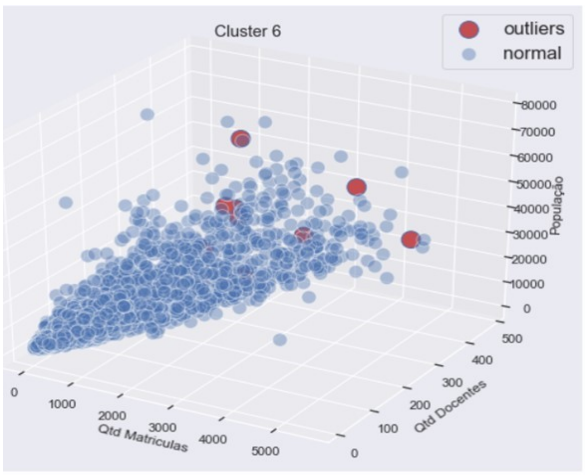


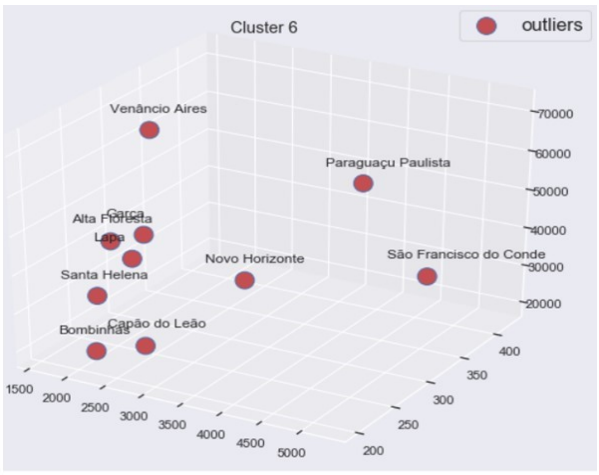




Fonte: Elaborada pelo autor (2020)

Figura 64 – Os 10 municípios anômalos – Comparação dos dados de despesas

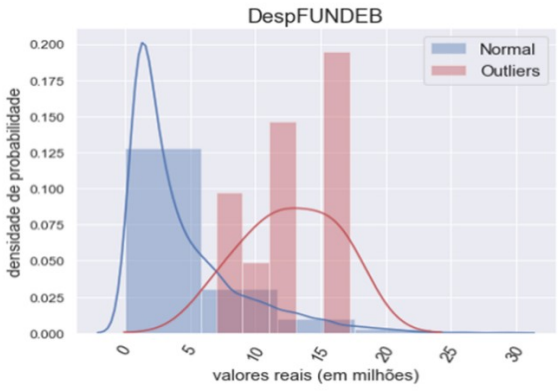


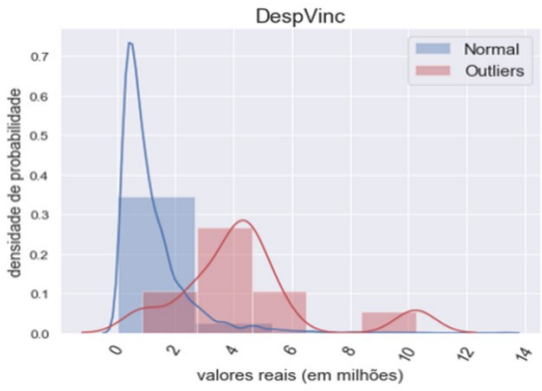


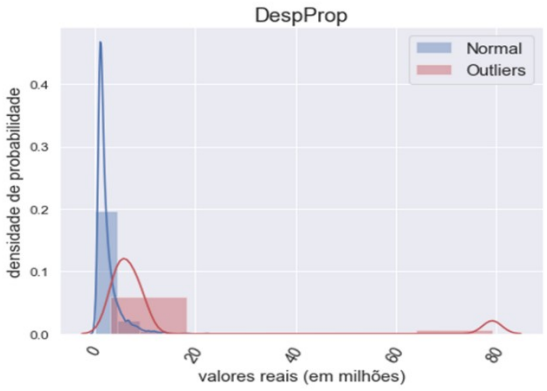
	DespFUNDEB	DespVinc	DespProp	DespVinc	tgRemun
count	3233.00	3233.00	3233.00	3233.00	3233.00
mean	4289156.39	1175863.79	2354977.16	1175863.79	5457233.62
std	4193131.01	1150610.20	2180574.85	1150610.20	4881259.90
min	0.00	2000.00	0.00	2000.00	0.00
25%	1334999.00	453094.20	1068547.60	453094.20	2069091.90
50%	2726222.20	822421.30	1594175.40	822421.30	3660412.80
75%	5878849.90	1486703.00	2788916.00	1486703.00	7243038.10
max	29267721.00	13313226.00	22283660.40	13313226.00	33520827.10

Estatísticas dos municípios anômalos

	DespFUNDEB	DespVinc	DespProp	DespVinc	tgRemun
count	10.00	10.00	10.00	10.00	10.00
mean	12679763.81	4376829.74	13555569.73	4376829.74	22554966.32
std	3521980.63	2449855.49	23178290.29	2449855.49	20351117.21
min	6964060.50	881183.50	3126054.70	881183.50	12854264.40
25%	10889002.88	3260135.53	4728524.80	3260135.53	14637206.50
50%	12271280.10	4210135.20	6619900.55	4210135.20	16895561.95
75%	15722029.98	4656907.30	8553469.38	4656907.30	17753073.18
max	17223486.90	10247410.10	79260447.90	10247410.10	80184416.20

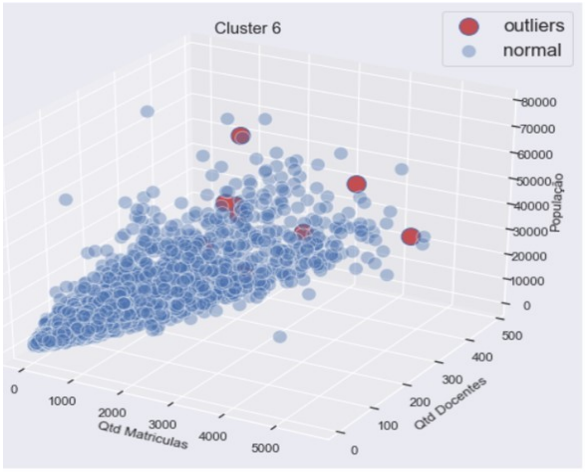


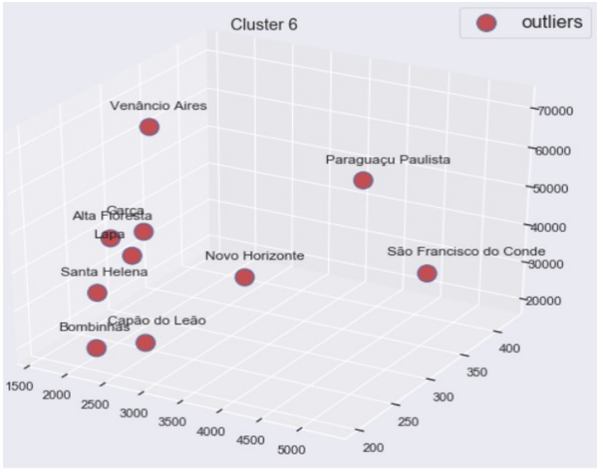




Fonte: Elaborada pelo autor (2020)

Figura 65 – Os 10 municípios anômalos – Comparação das subfunções e contas contábeis



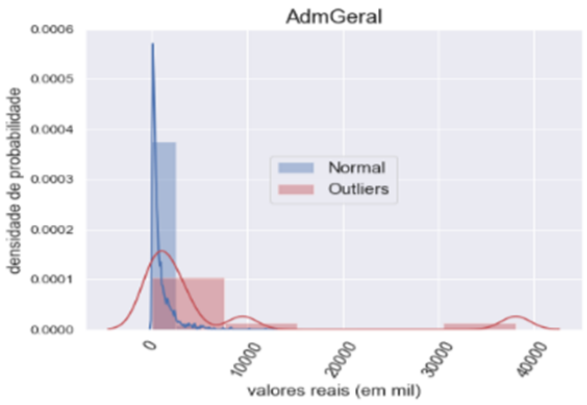


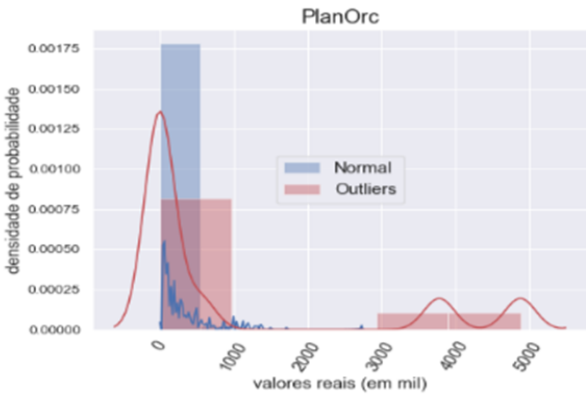
Estatísticas dos municípios não anômalos

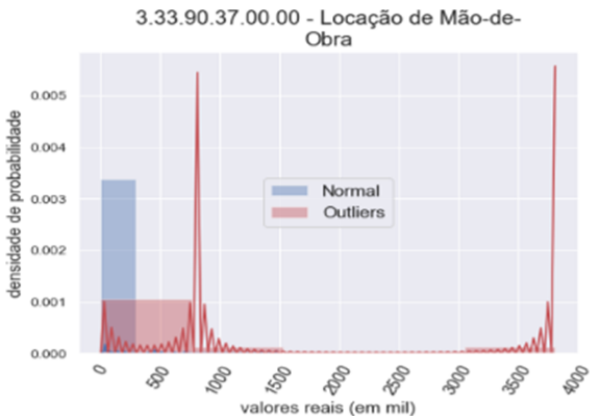
	AdmGeral	PlanOrc	Loc Mão de Obra	Inden e Rest Trab
count	3233.00	3233.00	3233.00	3233.00
mean	385891.06	35789.62	2891.97	204.33
std	916691.88	159118.00	46714.57	3661.90
min	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00
75%	357076.20	0.00	0.00	0.00
max	12942553.90	2756751.70	1475296.60	117179.20

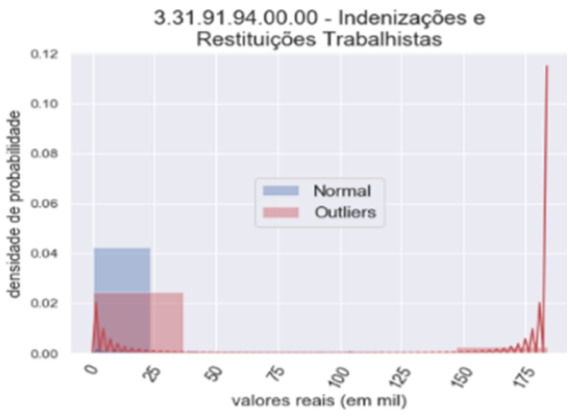
Estatísticas dos municípios anômalos

	AdmGeral	PlanOrc	Loc Mão de Obra	Inden e Rest Trab
10.00	10.00	10.00	10.00	
5825102.61	922495.94	462361.03	18390.74	
11675656.98	1826477.83	1204416.11	58156.63	
0.00	0.00	0.00	0.00	
311712.32	0.00	0.00	0.00	
1485411.80	0.00	0.00	0.00	
3361079.18	412500.00	0.00	0.00	
38060507.70	4887124.40	3812547.40	183907.40	









Fonte: Elaborada pelo autor (2020)

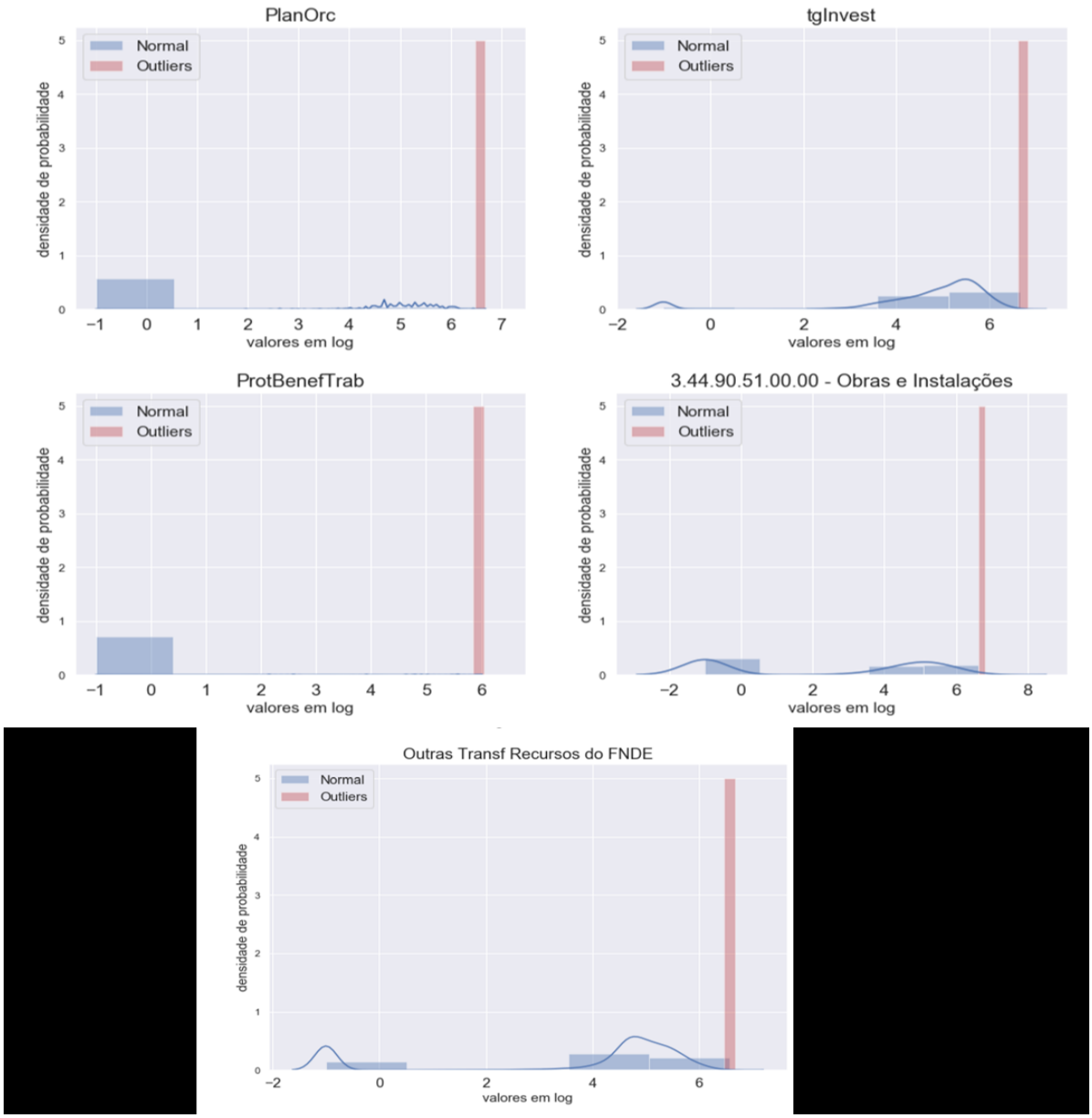
Figura 66 – Um exemplo de um município anômalo

Estatísticas do Cluster 6 (exceto Bombinhas)

	Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361	PlanOrc	tgInvest	ProtBenefTrab	Obras e Instalacoes	Outras Transf Rec FNDE
count	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00	3242.00
mean	12451.27	9.06	88.18	942.15	37367.36	284108.24	1305.80	133039.89	148096.19
std	11291.43	7.59	73.48	915.48	180532.20	423471.48	30665.23	281902.37	253380.70
min	781.00	1.00	4.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	4384.25	3.00	35.00	298.25	0.00	22180.70	0.00	0.00	7602.82
50%	8399.00	7.00	61.00	591.50	0.00	130872.25	0.00	10029.80	54170.90
75%	16814.25	12.00	119.00	1247.00	0.00	378673.40	0.00	138883.98	171412.65
max	78013.00	50.00	482.00	5622.00	4887124.40	4440136.20	1047799.40	4062826.00	3796703.70

Bombinhas (SC)

Pop_estimada	QtdEscolas	QtdDocentes	NUM_MATR_361	PlanOrc	tgInvest	ProtBenefTrab	Obras e Instalacoes	Outras Transf Rec FNDE
19769	11	213	2102	3787835.00	5273938.50	871000.00	5167917.60	3787835.00



Fonte: Elaborada pelo autor (2020)

6 FASE DE AVALIAÇÃO E IMPLANTAÇÃO

A fase de Avaliação examina se os resultados do modelo atendem aos objetivos de negócio e seus critérios de sucesso; e determina a decisão de o modelo ser implantado ou submetido às novas iterações de fases. A fase de Implantação resume-se em colocar o modelo obtido em produção, incluindo, quando aplicável, a confecção de relatório final.

Detectar anomalias em uma base de dados com centenas de atributos, de forma não supervisionada, mostrou-se uma tarefa complexa, pois há muitas questões desafiadoras, como:

- decidir qual o melhor algoritmo de escalonamento dos dados, bem como o algoritmo de clusterização mais adequado, dada a dificuldade de visualizar dados multidimensionais em apenas duas ou três dimensões;
- definir os atributos mais relevantes como entrada para os algoritmos (nas palavras da CGEBC, todas as despesas e contas contábeis são imprescindíveis e não devem ser descartadas durante a análise de dados[64]);
- estabelecer o limite (*threshold*) da pontuação de anormalidade de forma objetiva, visto que a fronteira entre o normal e o anômalo não é precisa;
- identificar os pontos anômalos mais internos, localizados em regiões com dois ou mais agrupamentos; e
- validar a acurácia dos resultados em uma situação sem dados históricos.

Além disso, frente à grande diversidade de algoritmos de detecção (de diferentes critérios – estatístico, distância, densidade, similaridade), é necessário escolher quais aplicar e otimizar os parâmetros de cada um. Alguns algoritmos ainda requerem a definição de critérios específicos, como o número de clusters (CBLOF) e o número de vizinhos próximos (kNN e *Average* kNN). Devido aos poucos recursos disponíveis (pessoal e tempo), todas estas questões não puderam ser trabalhadas em profundidade adequada.

Por outro lado, pode-se afirmar que o presente trabalho alcançou a finalidade principal de detectar anomalias em despesas dos municípios com o Ensino Fundamental (o *dataframe* de municípios é atualizado com os rótulos e as pontuações de cada algoritmo), tendo em vista que não havia nenhum trabalho prévio de análise estatística ou de dados no SIOPE pela CGEBC – e foi possível também atingir os objetivos definidos no item 3.6:

---

[64] Essa requisição, por parte da área de negócio, limitou o uso de PCA para redução da dimensionalidade, a fim de evitar a perda de informação.

- obter ao menos 1% de entes federativos com discrepâncias nos seus gastos educacionais declarados (quando se consideram todos os atributos) – conforme a Figura 60, o percentual de cada algoritmo variou entre 2 a 5%; e
- de fato, alguns entes foram identificados como anômalos em todos os algoritmos de detecção escolhidos (vide Figura 60, foram dez municípios identificados como anômalos pelos oito algoritmos aplicados).

Entretanto, percebeu-se que, para uma maior eficácia desta atividade de detecção, é necessário, ainda, estabelecer algumas estratégias mais individualizadas. Um exemplo seria determinar as despesas mais relevantes – caso fosse alimentação, pode-se cruzar os dados do SIOPE com bases de dados do SigPC (base de prestação de contas do FNDE que contém informações sobre o PNAE), a fim de detectar anomalias em gastos com merenda escolar. Desta forma, um caminho recomendado para o presente trabalho seria voltar à fase de entendimento do negócio para a reformulação de objetivos mais específicos, e apresentar os resultados consolidados em um painel gerencial.

Não obstante, o presente trabalho apresentou uma proposta, um caminho viável e produtivo para se chegar às anomalias nas despesas dos municípios com o Ensino Fundamental – através do uso concomitante de exploração de dados, de algoritmos de clusterização e de algoritmos focados na detecção de anomalias. É necessário reaplicar tais caminhos para a detecção de anomalias nas demais modalidades de ensino (Educação Infantil, Ensino Médio, etc.). O presente trabalho, portanto, não se encontra finalizado – na verdade, precisa ser submetido às novas iterações de fases, com reformulações de objetivos de negócio e de mineração de dados, e ser apresentado à área de auditoria da educação básica. Uma possibilidade seria a formulação de um painel consolidado, com a indicação dos municípios anômalos para cada modalidade de ensino, incluindo, ainda, para cada caso, os atributos envolvidos que influenciaram em uma pontuação maior das anomalias.

7 CONCLUSÃO

Entende-se que explorar os dados de despesas públicas, a fim de obter conhecimento estratégico e de valor agregado, representa um importante objetivo para o controle interno e também para a Administração Pública – pois possibilita, no caso do SIOPE, a identificação de despesas atípicas (anomalias) que podem indicar possíveis falhas ou irregularidades nos investimentos públicos em educação. Consequentemente, vislumbra-se a oportunidade de oferecer os resultados da detecção dessas despesas anômalas como insumos para uma série de atividades, a saber: planejamento das ações de controle (no caso, complementando a Matriz de Vulnerabilidade dos municípios); acompanhamento e adoção de providências cabíveis por parte das instâncias de controle; e criação de trilhas de auditoria voltadas para o monitoramento dos gastos públicos.

Desta forma, o presente trabalho teve como objetivo inicial uma extensa análise exploratória dos dados do SIOPE, cujos resultados delimitaram o escopo das despesas a serem investigadas (ou seja, as despesas municipais pagas no Ensino Fundamental, no ano de 2018) e determinaram as estratégias seguintes – o uso das técnicas de clusterização e detecção de anomalias.

Com base na ideia de que municípios semelhantes devem apresentar despesas educacionais também semelhantes, ao menos em ordem de grandeza – a clusterização buscou agrupar municípios semelhantes, ou seja, similares quanto aos dados de população; quantidades de alunos, docentes e escolas; e indicadores IDEB e IDHM. Várias execuções foram realizadas com diferentes algoritmos (k-*Means*, DBSCAN, *MeanShift* e *Agglomerative Clustering*), diferentes formas de escalonamento dos dados (*StandardScaler*, *RobustScaler* e *MinMaxScaler*) e diferentes parâmetros (número de clusters, *epsilon*, número de amostras, critério de ligação, etc.). Alguns métodos de cálculo do número ideal de clusters foram também utilizados (métodos *Elbow*, *Gap Statistics* e Coeficiente de Silhueta). Como medida da avaliação da qualidade da separação entre os clusters, foi utilizado o índice Davies-Bouldin, o qual foi interpretado em conjunto com gráficos que permitiram a verificação da coerência da separabilidade dos clusters. Apesar de sugestões (2 ou 6 clusters) pelos métodos mencionados, definiu-se como sete a quantidade ideal de clusters – pois já era esperada a criação de dois clusters com apenas um município (SP e RJ), e dois outros cluster com poucos municípios (menos de 1%). Concluiu-se que o *Agglomerative Clustering*, com dados escalonados com *RobustScaler*, apresentou os melhores resultados.

Finalmente, deu-se preferência ao cluster de maior número de municípios para submetê-lo a oito algoritmos de detecção de anomalias da biblioteca *PyOD*. Um parâmetro adicional foi a definição de quais atributos submeter aos algoritmos (escopo de despesa) – ou seja: todos os atributos de despesas; apenas grupos das despesas próprias e despesas FUNDEB; e tipos de gasto de remuneração e manutenção. Os resultados obtidos (classificação do município, se anômalo ou não; e pontuação da anormalidade, calculada por cada algoritmo em cada escopo de despesa) foram consolidados ao *dataframe* de municípios. Outras métricas também foram adicionadas, como a quantidade de vezes em que um município foi marcado como anômalo. No escopo de todos os atributos, foi possível identificar as despesas que influenciaram a pontuação da anormalidade (através de histogramas comparativos).

Pode-se afirmar que foi possível, por todo o trabalho, identificar as anomalias globais, locais e de cluster – ou seja, as despesas atípicas de cada município com relação aos seus municípios semelhantes.

O conjunto dos modelos gerados atendeu aos objetivos de negócio (indicar os municípios com discrepâncias nos seus gastos educacionais) e aos critérios de sucesso (obter ao menos 1% de municípios com discrepâncias nos seus gastos educacionais, sendo alguns apontados como anômalos em todos os algoritmos) definidos no item 3.6. Da mesma forma, pode-se dizer que todos os objetivos da mineração de dados (realizar tarefas de AED, de clusterização e de detecção de anomalias) foram também alcançados.

Como recomendações e sugestões para trabalhos futuros, há uma infinidade de possibilidades:

- detectar as anomalias nas despesas do Ensino Fundamental para os demais clusters de municípios, bem como reaplicar as técnicas utilizadas (AED, clusterização e detecção de anomalias em despesas) nas demais modalidades de ensino (Educação Infantil, Ensino Médio, Ensino Superior);
- realizar os mesmos estudos para os dados do SIOPE Estadual;
- realizar análises de dados considerando-se os valores de despesas *per capita* (dividindo-se as despesas pela população estimada ou pela quantidade de alunos matriculados);
- criar perfil normal de gastos com educação, para cada município ou grupo de municípios semelhantes (envolve analisar os dados do SIOPE dos anos anteriores a 2018) – para possibilitar a detecção mais imediata de anomalias a partir desse perfil

de gastos, a cada bimestre (por ser a periodicidade da transmissão de dados do SIOPE);

- propor ou buscar medidas de acurácia dos resultados dos algoritmos não supervisionados;
- consolidar as classificações e pontuações dos algoritmos de detecção de anomalia em uma plataforma de *business intelligence*, de forma a apresentar painéis gerenciais, com múltiplas visões dos dados, às equipes de auditoria;
- estudar a viabilidade de cruzamento do SIOPE com outras bases de dados dos entes federativos, como a base SIAFEM (trata-se de um sistema equivalente ao SIAFI, porém, no âmbito dos estados e municípios);
- verificar possibilidades de aplicar técnicas de *deep learning* para a detecção de anomalias, tais como *autoencoders* e redes generativas adversariais.

REFERÊNCIAS

AGUIAR, Gilson. 2012. **Pior que a corrupção é a má gestão**. 2012. Disponível em: <https://www.cbnmaringa.com.br/noticia/gilson-aguiar-pior-que-a-corrupcao-e-a-ma-gestao>. Acesso em: 20 fev 2020.

ALVES, Gisely. **Aprendizado não supervisionado com K-means**. Disponível em: <https://medium.com/neuronio-br/aprendizado-n%C3%A3o-supervisionado-com-k-means-f4272dee98a0>. Acesso em 20 dez 2019.

ANGÉLICO, Fabiano. **Má gestão + corrupção = nota baixa**. 2012. Disponível em <https://apublica.org/2012/07/ma-gestao-corrupcao-nota-baixa>. Acesso em: 20 fev 2020.

ARCOVERDE, Léo, TOLEDO, Luiz Fernando. CGU aponta uso irregular de quase R$ 51 milhões do Fundeb em todo o país. **G1 Portal de Notícias**, 2019. Disponível em: <https://g1.globo.com/educacao/noticia/2019/08/15/cgu-aponta-uso-irregular-de-quase-r-51-milhoes-do-fundeb-em-todo-o-pais.ghtml>. Acesso em 20 nov 2019.

Associação de Jornalistas de Educação (JEDUCA). **Financiamento da Educação Básica - Guia de Cobertura**. São Paulo: Editora Moderna, 2019. Disponível em: <http://jeduca.org.br/arquivos/Financiamento-da-Educacao-basica-121822.pdf>. Acesso em 02 fev 2020.

Controladoria lança ferramenta para avaliação preventiva e automatizada de editais de licitação. **Governo Federal**, 2015. Disponível em: <https://www.gov.br/cgu/pt-br/assuntos/noticias/2015/06/controladoria-lanca-ferramenta-para-avaliacao-preventiva-e-automatizada-de-editais-de-licitacao>. Acesso em 18 fev 2020

DAVIES, David L.; BOULDIN, Donald W. (1979). "A Cluster Separation Measure" IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227. doi:10.1109/TPAMI.1979.4766909.

PROVOST, Foster; FAWCETT, Tom. **Data Science para Negócios. O que você precisa saber sobre mineração de dados e pensamento analítico de dados**. 1.ed. Rio de Janeiro: Alta Books, 2016. ISBN: 9788576089728.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>. Acesso em: 19 nov 2019.

BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996**. Estabelece as diretrizes e bases da educação nacional. Brasília, 1996. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L9394compilado.htm>. Acesso em: 19 nov 2019.

BRASIL. **Lei nº 11.494, de 20 de junho de 2007**. Regulamenta o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação - FUNDEB. Brasília, 2007. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/lei/l11494.htm>. Acesso em: 19 nov 2019.

BRASIL, Controladoria-Geral da União. **Relatório de Fiscalização nº 201602219 - Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb)**. Disponível em <https://auditoria.cgu.gov.br/download/12682.pdf>. Acesso em: 19 nov 2019a.

BRASIL, Controladoria-Geral da União. **Relatório de Fiscalização nº 201602218 - Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb)**. Disponível em <https://auditoria.cgu.gov.br/download/12685.pdf>. Acesso em: 19 nov 2019b.

BRASIL, Controladoria-Geral da União. **Relatório de Auditoria Anual de Contas nº 201900673 - Fundo Nacional de Desenvolvimento da Educação - Exercício 2018**. Disponível em: <https://auditoria.cgu.gov.br/download/13670.pdf>. Acesso em 20 dez 2019c.

BRASIL, Controladoria-Geral da União. **Orientações para o acompanhamento das ações do Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB)**. Disponível em <https://www.cgu.gov.br/Publicacoes/controle-social/arquivos/fundeb2012.pdf>. Acesso em: 20 dez 2019d.

BRASIL, Controladoria-Geral da União. **Entenda os indicadores** (Matriz de Vulnerabilidade). Disponível em: <https://www.cgu.gov.br/assuntos/auditoria-e-fiscalizacao/programa-de-fiscalizacao-em-entes-federativos/1-ciclo/1o-ciclo/entenda-os-indicadores>. Acesso em 21 dez 2019e.

BRASIL, Controladoria-Geral da União. **Relatório de Fiscalização nº 201902570 - Mata Roma (MA) - Educação e Saúde**. Disponível em <https://auditoria.cgu.gov.br/download/13842.pdf>. Acesso em: 20 jan 2020a.

BRASIL, Controladoria-Geral da União. **Portaria nº 3.553, de 12 de novembro de 2019**. Aprova o Regimento Interno e o Quadro Demonstrativo de Cargos em Comissão e das Funções de Confiança da Controladoria-Geral da União - CGU e dá outras providências. Disponível em <http://www.in.gov.br/web/dou/-/portaria-n-3.553-de-12-de-novembro-de-2019-227654932>. Acesso em 02 fev 2020b.

BRASIL, Fundo Nacional de Desenvolvimento da Educação. **Portal SIOPE: Sistema de Informações sobre Orçamentos Públicos em Educação**. Disponível em: <https://www.fnde.gov.br/fnde_sistemas/siope>. Acesso em 06 ago 2019.

BRASIL, Instituto Brasileiro de Geografia e Estatística. **Estimativas da População**. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados>. Acesso em 08 out 2019.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB**. Disponível em: <http://portal.inep.gov.br/web/guest/educacao-basica/ideb/resultados>. Acesso em 08 out 2019a.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB**. Municípios – Ensino Fundamental Regular – Anos Iniciais. Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/divulgacao_anos_iniciais_municipios2017-atualizado-Jun_2019.xlsx >. Acesso em 08 out 2019b.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB**. Municípios – Ensino Fundamental Regular – Anos Finais. Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/divulgacao_anos_finais_municipios2017-atualizado-Jun_2019.xlsx >. Acesso em 08 out 2019c.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resultados do IDEB**. Municípios – Ensino Médio. Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/divulgacao_ensino_medio_municipios2017-atualizado-Jun_2019.xlsx >. Acesso em 08 out 2019d.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Instruções para utilização dos Microdados do Censo da Educação Básica 2018**. Disponível em: <http://download.inep.gov.br/microdados/microdados_educacao_basica_2018.zip>. Acesso em 12 out 2019e.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Taxas de Transição 2014/2015.** Disponível em: <http://portal.inep.gov.br/web/guest/indicadores-educacionais>. Acesso em: 12 out 2019f.

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Notas Estatísticas - Censo da Educação Básica 2019**. Disponível em: <http://portal.inep.gov.br/censo-escolar>. Acesso em 02 fev 2020.

BRASIL, Ministério da Educação. **SIOPE Municipal 2018: Manual de Orientações para o Usuário**. Disponível em: https://www.fnde.gov.br/index.php/centrais-de-conteudos/publicacoes/category/139-siope?download=11869:manual-siope-municipal. Acesso em 06 ago 2019.

BRASIL, Ministério do Planejamento, Desenvolvimento e Gestão. **Portaria nº 42, de 14 de abril de 1999**. Atualiza a discriminação da despesa por funções e dá outras providências. Disponível em <http://www.orcamentofederal.gov.br/orcamentos-anuais/orcamento-1999/Portaria_Ministerial_42_de_140499.pdf>. Acesso em 02 fev 2020.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão. Secretaria de Orçamento Federal. **Manual técnico de orçamento - MTO 2017**. Brasília, 2016. Disponível em: <http://www.orcamentofederal.gov.br/informacoes-orcamentarias/manual-tecnico/mto_2017-1a-edicao-versao-de-06-07-16.pdf >. Acesso em 02 fev 2020.

BRASIL. Programa das Nações Unidas para o Desenvolvimento. **O que é o IDHM**. Disponível em: <https://www.br.undp.org/content/brazil/pt/home/idh0/conceitos/o-que-e-o-idhm.html>. Acesso em 13 out 2019.

BRASIL. Tribunal de Contas da União. **Acórdão nº 618/2014**. Plenário. Relator: Ministro Valmir Campelo. Sessão de 19/3/2014. Disponível em: <https://pesquisa.apps.tcu.gov.br/#/redireciona/acordao-completo/%22ACORDAO-COMPLETO-1300927%22>. Acesso em: 02 fev 2020.

CHAPMAN, Pete et al. **CRISP-DM 1.0**: Step-by-step data mining guide. Tecnical report. The CRISP-DM consortium, 2000. Disponível em: <https://pdfs.semanticscholar.org/5406/1a4aa0cb241a726f54d0569efae1c13aab3a.pdf>. Acesso em: 29 jan 2020.

GOLDSTEIN, Markus e UCHIDA, Selichi. **A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data**. PLoS ONE, 11(4): e0152173, April, 2016. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173> Acesso em: 23 dez 2019.

GOIX, Nicolas. **How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?** 2016. Disponível em: <https://arxiv.org/abs/1607.01152> Acesso em: 25/02/2020.

HE, Zengyou; XIAOFEI, Xu e SHENGCHUN, Deng. **Discovering cluster-based local outliers**. Pattern Recognit. Lett., vol. 24, p. 1641-1650, 2003.

IBM. **IBM SPSS Modeler CRISP-DM Guide**. Disponível em: <https://www.ibm.com/support/knowledgecenter/SS3RA7_18.2.1/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html>, 2019. Acesso em 29 jan 2020.

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.

LEEK, Jeff. **The Elements of Data Analytic Style**. Leanpub, Victoria British Columbia, 2015.

PEDREGOSA, Fabian et all. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 2011. volume 12, p. 2825–2830.

PROJECT JUPYTER. **Project Jupyter®**. Disponível em: <https://jupyter.org>. Acesso em: 30 ago 2019.

PYTHON, Software Foundation. **Python**. Disponível em: <https://www.python.org/>. Acesso em: 30 ago 2019.

QUEIROZ, Christina. **Pesquisa FAPESP: Engrenagem Complexa. Alimentados por arrecadação tributária, regimes de financiamento à educação como o Fundeb, que expira em 2020, constituem desafio ao governo federal**. Disponível em: https://revistapesquisa.fapesp.br/2019/03/12/engrenagem-complexa>. Acesso em 02 fev 2020.

SEABORN. **SEABORN statistical data visualization**. Disponível em <https://seaborn.pydata.org/index.html>. Acesso em: 30 ago 2019.

SEN, Soumya. **Intercluster and Intracluster Distance**. Disponível em: <https://www.geeksforgeeks.org/ml-intercluster-and-intracluster-distance>. Acesso em: 20 dez 2019.

SRIVASTAVA, Shobhit. **Feature Scaling in Scikit-learn**. Disponível em: <https://medium.com/analytics-vidhya/feature-scaling-in-scikit-learn-b11209d949e7> Acesso em: 20 dez 2019.

THE PANDAS PROJECT. **Pandas - Python Data Analysis Library**. Disponível em: <https://pandas.pydata.org>. Acesso em: 30 ago 2019.

THESING , Ana Paula. **Analytics e Big Data são poderosas armas contra a corrupção**. 2019. Disponível em: <https://www.itforum365.com.br/analytics-e-big-data-sao-poderosas-armas-contra-a-corrupcao>. Acesso em 20 fev 2020.

ZAKI, Mohammed J. e MEIRA JR., Wagner. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. Cambridge University Press, 2014. Disponível em: <http://www.dataminingbook.info/pmwiki.php/Main/BookDownload>. Acesso em: 20 fev 2020.

ZHAO, Yue; NASRULLAH, Zain; LI, Zheng. **PyOD: A Python Toolbox for Scalable Outlier Detection**. Journal of Machine Learning Research, v. 20, n. 96, p. 1-7, 2019.

APÊNDICE A – *Script* para filtrar as Despesas Próprias

O apêndice abaixo apresenta o script em *SQL* utilizado para selecionar todos os registros do grupo de Despesas Próprias da base do SIOPE Municipal (de todos os municípios), para o ano de 2018.

```
SELECT COD_UF, NOM_UF, SIG_UF, COD_MUNI, NOM_MUNI,
Classif_Pasta,Codigo_subfuncao_Pasta, CodPasta_PAI, Past_Pai,
Cod_Exib_Pasta, Nome_Pasta, Nivel_Pasta, Ordem_Pasta,
cod_item, num_orde, num_nive, COD_ITEM_PAI,
Cod_CC, Cod_CC_f, Nome_Conta_Contabil, Descricao_CC,
[Dotação Atualizada], [Desp. Empenhadas], [Desp. Pagas], [Desp. Liquidadas]
 into DESPESAS_MUNIC_PROPRIAS_2018
 FROM (
 select t1.COD_UF, uf.NOM_UF, uf.SIG_UF, m.COD_MUNI, m.NOM_MUNI,
 case when t4.IDN_CLAS = 'P' then 'Desp proprias - EF'
 when t4.IDN_CLAS = 'I' then 'Desp proprias - não EF'
 when t4.IDN_CLAS = 'F' then 'Desp FUNDEF'
 when t4.IDN_CLAS = 'V' then 'Desp com Recursos Vinculados'
 when t4.IDN_CLAS = 'O' then 'Outras'
 else 'Outros' end as 'Classif_Pasta',
 t4.COD_SUBF as Codigo_subfuncao_Pasta,
case -- codigo pasta pai
 when t4.NUM_NIVE = 4 and t4.COD_EXIB not in (361, 362, 363, 364, 365) and
t4.COD_SUBF not in (361, 362, 363, 364, 365) then ''
 when t4.NUM_NIVE = 5 and t4.COD_EXIB not in (365) then t4.COD_SUBF
 when t4.NUM_NIVE = 5 and t4.COD_EXIB = 365 and t4.NUM_ORDE = 75 then
t4.COD_SUBF
 when t4.NUM_NIVE = 5 and t4.COD_EXIB = 365 and t4.NUM_ORDE = 93 then
t4.COD_SUBF
 else Null end as CodPasta_PAI,
case -- nome pasta pai
 when t4.NUM_NIVE = 4 and t4.COD_EXIB not in (361, 362, 363, 364, 365) and
t4.COD_SUBF not in (361, 362, 363, 364, 365)
 then 'Despesas Próprias Custeadas com Impostos e Transferências'
 when t4.NUM_NIVE = 5 and t4.COD_EXIB = 365 and t4.NUM_ORDE = 75
 then (select nom_past from mun_me_pasta where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='I'
 and cod_exib = t4.COD_SUBF
 and num_orde = 65 )
 when t4.NUM_NIVE = 5 and t4.COD_EXIB = 365 and t4.NUM_ORDE = 93
 then (select nom_past from mun_me_pasta where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='I'
 and cod_exib = t4.COD_SUBF
 and num_orde = 83 )
 when t4.NUM_NIVE = 5 and t4.COD_EXIB not in (365) and t4.COD_SUBF = 365
and t4.NUM_ORDE <=82
 then (select nom_past from mun_me_pasta where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='I'
 and cod_exib = t4.COD_SUBF
 and num_orde = 65 )
 when t4.NUM_NIVE = 5 and t4.COD_EXIB not in (365) and t4.COD_SUBF = 365
and t4.NUM_ORDE > 82
 then (select nom_past from mun_me_pasta where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='I'
 and cod_exib = t4.COD_SUBF
 and num_orde = 83 )
```

```
 when t4.NUM_NIVE = 5 and t4.COD_EXIB not in (365) and t4.COD_SUBF <> 365
 then (select nom_past from mun_me_pasta where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='I'
 and cod_exib = t4.COD_SUBF)
 else Null end as Past_Pai,
 t4.COD_EXIB as Cod_Exib_Pasta, t4.NOM_PAST as Nome_Pasta,
 t4.NUM_NIVE as Nivel_Pasta, t4.NUM_ORDE as Ordem_Pasta,
 t5.COD_ITEM, t5.NUM_ORDE, t5.NUM_NIVE, t5.COD_ITEM_PAI,
 t6.COD_EXIB as Cod_CC,
 substring(t6.COD_EXIB,1,1) + '.' + substring(t6.COD_EXIB,2,2) + '.' +
substring(t6.COD_EXIB,4,2) + '.' + substring(t6.COD_EXIB,6,2)
 + '.' + substring(t6.COD_EXIB,8,2) + '.' + substring(t6.COD_EXIB,10,2)as
Cod_CC_f,
 t6.NOM_ITEM as Nome_Conta_Contabil, t6.DES_ITEM as Descricao_CC,
 t10.NOM_COLU,
 cast(t1.val_decl as decimal(20,1)) as val_decl_f
 from [dbo].MUN_VALORES_DECLARADOS t1
 inner join MUN_INSTITUICAO t2 on t2.COD_INST = t1.COD_INST and t2.NUM_ANO
= t1.NUM_ANO and t2.NUM_PERI = t1.NUM_PERI and t2.COD_UF = t1.COD_UF and
t2.COD_MUNI = t1.COD_MUNI
 inner join [dbo].[UF] uf on uf.COD_UF = t1.COD_UF
 inner join dbo.MUN_MUNICIPIO m on m.COD_MUNI = t1.COD_MUNI
 inner join MUN_ME_PASTA_COLUNA t3 on t3.NUM_ANO = t1.NUM_ANO and
t3.NUM_PERI = t1.NUM_PERI and t3.COD_PAST = t1.COD_PAST and t3.COD_COLU =
t1.COD_COLU
 inner join MUN_ME_PASTA t4 on t4.NUM_ANO = t3.NUM_ANO and t4.NUM_PERI =
t3.NUM_PERI and t4.COD_PAST = t3.COD_PAST
 inner join MUN_ME_PASTA_ITEM t5 on t5.NUM_ANO = t4.NUM_ANO and t5.NUM_PERI
= t4.NUM_PERI and t5.COD_PAST = t4.COD_PAST and t5.NUM_ANO = t1.NUM_ANO
and t5.NUM_PERI = t1.NUM_PERI and t5.COD_PAST = t1.COD_PAST and t5.COD_ITEM
= t1.COD_ITEM
 inner join MUN_ME_ITEM t6 on t6.COD_ITEM = t5.COD_ITEM
 left join MUN_ME_ITEM t7 on t7.COD_ITEM = t5.COD_ITEM_PAI
 inner join MUN_ME_COLUNA t10 on t10.COD_COLU = t3.COD_COLU
 wheret1.num_ano = 2018 and t1.num_peri = 6 -- anual
and t4.idn_clas in ('I') -- DESPESAS PROPRIAS
and t4.IDN_DESP_RECE = 'D' -- despesas
 ) SQ
 PIVOT (sum(val_decl_f) for NOM_COLU in ([Dotação Atualizada], [Desp.
Empenhadas], [Desp. Pagas], [Desp. Liquidadas])) AS pt
 order by COD_UF, Ordem_Pasta, COD_CC
```

APÊNDICE B – *Script* para filtrar as Despesas FUNDEB

O apêndice abaixo apresenta o script em *SQL* utilizado para selecionar todos os registros do grupo de Despesas FUNDEB da base do SIOPE Municipal (de todos os municípios), para o ano de 2018.

```
SELECT COD_UF, NOM_UF, SIG_UF, COD_MUNI, NOM_MUNI,
 Classif_Pasta,Codigo_subfuncao_Pasta, CodPasta_PAI, Past_Pai,
Cod_Exib_Pasta, Nome_Pasta, Nivel_Pasta, Ordem_Pasta,
 cod_item, num_orde, num_nive, COD_ITEM_PAI,
 Cod_CC, Cod_CC_f, Nome_Conta_Contabil, Descricao_CC,
 [Dotação Atualizada], [Desp. Empenhadas], [Desp. Pagas], [Desp.
Liquidadas]
 into DESPESAS_MUNIC_FUNDEB_2018
 FROM (

 select t1.COD_UF, uf.NOM_UF, uf.SIG_UF, m.COD_MUNI, m.NOM_MUNI,
 case when t4.IDN_CLAS = 'P' then 'Desp proprias - EF'
 when t4.IDN_CLAS = 'I' then 'Desp proprias - não EF'
 when t4.IDN_CLAS = 'F' then 'Desp FUNDEF'
 when t4.IDN_CLAS = 'V' then 'Desp com Recursos Vinculados'
 when t4.IDN_CLAS = 'O' then 'Outras'
 else 'Outros' end as 'Classif_Pasta',
 t4.COD_SUBF as Codigo_subfuncao_Pasta,
 case -- codigo pasta pai
 when t4.NUM_NIVE = 5 and t4.COD_SUBF <> 365
 then (select COD_EXIB from MUN_ME_PASTA where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='F'
 and cod_exib = t4.COD_SUBF)
 when t4.NUM_NIVE = 5 and t4.COD_SUBF = 365
 then (select COD_EXIB from MUN_ME_PASTA where num_ano = 2018 and num_peri
= 6 and num_nive =4 and idn_clas='F' and num_orde =
 (select max(num_orde) from MUN_ME_PASTA where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='F'
 and cod_subf = t4.COD_SUBF and num_orde < t4.NUM_ORDE))
 else Null end as CodPasta_PAI,
 case -- nome pasta pai
 when t4.NUM_NIVE = 5 and t4.COD_SUBF <> 365
 then (select nom_past from MUN_ME_PASTA where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='F'
 and cod_exib = t4.COD_SUBF)
 when t4.NUM_NIVE = 5 and t4.COD_SUBF = 365
 then (select nom_past from MUN_ME_PASTA where num_ano = 2018 and num_peri
= 6 and num_nive =4 and idn_clas='F' and num_orde =
 (select max(num_orde) from MUN_ME_PASTA where num_ano = 2018 and num_peri
= 6 and num_nive = 4 and idn_clas='F'
 and cod_subf = t4.COD_SUBF and num_orde < t4.NUM_ORDE))
 else Null end as Past_Pai,
 t4.COD_EXIB as Cod_Exib_Pasta, t4.NOM_PAST as Nome_Pasta,
 t4.NUM_NIVE as Nivel_Pasta, t4.NUM_ORDE as Ordem_Pasta,
 t5.COD_ITEM, t5.NUM_ORDE, t5.NUM_NIVE, t5.COD_ITEM_PAI,
 t6.COD_EXIB as Cod_CC,
 substring(t6.COD_EXIB,1,1) + '.' + substring(t6.COD_EXIB,2,2) + '.' +
substring(t6.COD_EXIB,4,2) + '.' + substring(t6.COD_EXIB,6,2)
 + '.' + substring(t6.COD_EXIB,8,2) + '.' + substring(t6.COD_EXIB,10,2) as
Cod_CC_f,
 t6.NOM_ITEM as Nome_Conta_Contabil, t6.DES_ITEM as Descricao_CC,
```

```
 t10.NOM_COLU,
 cast(t1.val_decl as decimal(20,1)) as val_decl_f
 from [dbo].MUN_VALORES_DECLARADOS t1
 inner join MUN_INSTITUICAO t2 on t2.COD_INST = t1.COD_INST and t2.NUM_ANO
= t1.NUM_ANO and t2.NUM_PERI = t1.NUM_PERI and t2.COD_UF = t1.COD_UF and
t2.COD_MUNI = t1.COD_MUNI
 inner join [dbo].[UF] uf on uf.COD_UF = t1.COD_UF
 inner join dbo.MUN_MUNICIPIO m on m.COD_MUNI = t1.COD_MUNI
 inner join MUN_ME_PASTA_COLUNA t3 on t3.NUM_ANO = t1.NUM_ANO and
t3.NUM_PERI = t1.NUM_PERI and t3.COD_PAST = t1.COD_PAST and t3.COD_COLU =
t1.COD_COLU
 inner join MUN_ME_PASTA t4 on t4.NUM_ANO = t3.NUM_ANO and t4.NUM_PERI =
t3.NUM_PERI and t4.COD_PAST = t3.COD_PAST
 inner join MUN_ME_PASTA_ITEM t5 on t5.NUM_ANO = t4.NUM_ANO and t5.NUM_PERI
= t4.NUM_PERI and t5.COD_PAST = t4.COD_PAST and t5.NUM_ANO = t1.NUM_ANO
 and t5.NUM_PERI = t1.NUM_PERI and t5.COD_PAST = t1.COD_PAST and
t5.COD_ITEM = t1.COD_ITEM
 inner join MUN_ME_ITEM t6 on t6.COD_ITEM = t5.COD_ITEM
 left join MUN_ME_ITEM t7 on t7.COD_ITEM = t5.COD_ITEM_PAI
 inner join MUN_ME_COLUNA t10 on t10.COD_COLU = t3.COD_COLU
 where t1.num_ano = 2018 and t1.num_peri = 6 -- anual
 and t4.idn_clas in ('F') -- DESPESAS FUNDEB
 and t4.IDN_DESP_RECE = 'D' -- despesas
 ) SQ
 PIVOT (sum(val_decl_f) for NOM_COLU in ([Dotação Atualizada], [Desp.
Empenhadas], [Desp. Pagas], [Desp. Liquidadas])) AS pt
 order by COD_UF, Ordem_Pasta, COD_CC
```

APÊNDICE C – *Script* para filtrar as Despesas Vinculadas

O apêndice abaixo apresenta o script em *SQL* utilizado para selecionar todos os registros do grupo de Despesas Vinculadas da base do SIOPE Municipal (de todos os municípios), para o ano de 2018.

```
SELECT COD_UF, NOM_UF, SIG_UF, COD_MUNI, NOM_MUNI,
Classif_Pasta,Codigo_subfuncao_Pasta, Nome_Programa, CodPasta_PAI,
Past_Pai, Cod_Exib_Pasta, Nome_Pasta, Nivel_Pasta, Ordem_Pasta,
cod_item, num_orde, num_nive, COD_ITEM_PAI,
Cod_CC, Cod_CC_f, Nome_Conta_Contabil, Descricao_CC,
[Dotação Atualizada], [Desp. Empenhadas], [Desp. Pagas], [Desp. Liquidadas]
into DESPESAS_MUNIC_VINC_2018
FROM (
select t1.COD_UF, uf.NOM_UF, uf.SIG_UF, m.COD_MUNI, m.NOM_MUNI,
case when t4.IDN_CLAS = 'P' then 'Desp proprias - EF'
when t4.IDN_CLAS = 'I' then 'Desp proprias - não EF'
when t4.IDN_CLAS = 'F' then 'Desp FUNDEF'
when t4.IDN_CLAS = 'V' then 'Desp com Recursos Vinculados'
when t4.IDN_CLAS = 'O' then 'Outras'
else 'Outros' end as 'Classif_Pasta',
t4.COD_SUBF as Codigo_subfuncao_Pasta,
(select nom_past from MUN_ME_PASTA where num_ano = 2018 and num_peri = 6
and idn_clas='V' and num_nive = 4 and cod_subf = t4.COD_SUBF) as
Nome_Programa,
case -- codigo pasta pai
when t4.NUM_NIVE = 6
then (
select cod_exib from MUN_ME_PASTA where num_ano = 2018 and num_peri = 6 and
num_nive = 5 and idn_clas='V' and num_orde =
(select max(num_orde)
from MUN_ME_PASTA
where num_ano = 2018 and num_peri = 6 and num_nive = 5 and idn_clas='V'
and cod_subf = t4.COD_SUBF and num_orde < t4.NUM_ORDE))
else Null end as CodPasta_PAI,
case -- nome pasta pai
when t4.NUM_NIVE = 6
then (
select nom_past from MUN_ME_PASTA where num_ano = 2018 and num_peri = 6 and
num_nive = 5 and idn_clas='V' and num_orde =
(select max(num_orde)
from MUN_ME_PASTA
where num_ano = 2018 and num_peri = 6 and num_nive = 5 and idn_clas='V'
and cod_subf = t4.COD_SUBF and num_orde < t4.NUM_ORDE))
else Null end as Past_Pai,
t4.COD_EXIB as Cod_Exib_Pasta, t4.NOM_PAST as Nome_Pasta,
t4.NUM_NIVE as Nivel_Pasta, t4.NUM_ORDE as Ordem_Pasta,
t5.COD_ITEM, t5.NUM_ORDE, t5.NUM_NIVE, t5.COD_ITEM_PAI,
t6.COD_EXIB as Cod_CC,
substring(t6.COD_EXIB,1,1) + '.' + substring(t6.COD_EXIB,2,2) + '.' +
substring(t6.COD_EXIB,4,2) + '.' + substring(t6.COD_EXIB,6,2)
 + '.' + substring(t6.COD_EXIB,8,2) + '.' + substring(t6.COD_EXIB,10,2) as
Cod_CC_f,
t6.NOM_ITEM as Nome_Conta_Contabil, t6.DES_ITEM as Descricao_CC,
 t10.NOM_COLU,
 cast(t1.val_decl as decimal(20,1)) as val_decl_f
 from [dbo].MUN_VALORES_DECLARADOS t1
```

```
 inner join MUN_INSTITUICAO t2 on t2.COD_INST = t1.COD_INST and t2.NUM_ANO
= t1.NUM_ANO and t2.NUM_PERI = t1.NUM_PERI and t2.COD_UF = t1.COD_UF and
t2.COD_MUNI = t1.COD_MUNI
 inner join [dbo].[UF] uf on uf.COD_UF = t1.COD_UF
 inner join dbo.MUN_MUNICIPIO m on m.COD_MUNI = t1.COD_MUNI
 inner join MUN_ME_PASTA_COLUNA t3 on t3.NUM_ANO = t1.NUM_ANO and
t3.NUM_PERI = t1.NUM_PERI and t3.COD_PAST = t1.COD_PAST and t3.COD_COLU =
t1.COD_COLU
 inner join MUN_ME_PASTA t4 on t4.NUM_ANO = t3.NUM_ANO and t4.NUM_PERI =
t3.NUM_PERI and t4.COD_PAST = t3.COD_PAST
 inner join MUN_ME_PASTA_ITEM t5 on t5.NUM_ANO = t4.NUM_ANO and t5.NUM_PERI
= t4.NUM_PERI and t5.COD_PAST = t4.COD_PAST and t5.NUM_ANO = t1.NUM_ANO
 and t5.NUM_PERI = t1.NUM_PERI and t5.COD_PAST = t1.COD_PAST and
t5.COD_ITEM = t1.COD_ITEM
 inner join MUN_ME_ITEM t6 on t6.COD_ITEM = t5.COD_ITEM
 left join MUN_ME_ITEM t7 on t7.COD_ITEM = t5.COD_ITEM_PAI
 inner join MUN_ME_COLUNA t10 on t10.COD_COLU = t3.COD_COLU
 where t1.num_ano = 2018 and t1.num_peri = 6 -- anual
and t4.idn_clas in ('V') -- DESPESAS VINCULADAS
 and t4.IDN_DESP_RECE = 'D' -- despesas
 ) SQ
 PIVOT (sum(val_decl_f) for NOM_COLU in ([Dotação Atualizada], [Desp.
Empenhadas], [Desp. Pagas], [Desp. Liquidadas])) AS pt
 order by COD_UF, Ordem_Pasta, COD_CC
```

APÊNDICE D – *Scripts* para seleção de todas as despesas municipais

```
despesas_mun_2018 = '''
SELECT
    COD_UF as CodUF, NOM_UF as NomeUF, SIG_UF as SigUF,
    COD_MUNI as CodMunicipio,
    NOM_MUNI as NomeMunicipio,
    Classif_Pasta as GrupoDespesa,
    Codigo_subfuncao_Pasta as CodSubFuncao,
    'N/A' as NomePrograma,
    CodPasta_PAI as CodPasta_Pai, Past_Pai as NomePasta_Pai,
    Cod_Exib_Pasta as CodPasta, Nome_Pasta as NomePasta,
    Nivel_Pasta as NivelPasta, Ordem_Pasta as OrdPasta,
    cod_item as CodItem, num_orde as NumItem, num_nive as NivelItem,
    COD_ITEM_PAI as CodItem_Pai,
    Cod_CC as CodCC, Cod_CC_f as CodCC_f, Nome_Conta_Contabil as NomeCC,
    [Dotação Atualizada] as DA, [Desp. Empenhadas] as DE, [Desp. Liquidadas] as DL, [Desp. Pagas] as DP
FROM dbo.DESPESAS_MUNIC_proprias_2018
UNION ALL
SELECT  COD_UF as CodUF, NOM_UF as NomeUF, SIG_UF as SigUF,
        COD_MUNI as CodMunicipio,
        NOM_MUNI as NomeMunicipio,
        Classif_Pasta as GrupoDespesa,
        Codigo_subfuncao_Pasta as CodSubFuncao,
        'N/A' as NomePrograma,
        CodPasta_PAI as CodPasta_Pai, Past_Pai as NomePasta_Pai,
        Cod_Exib_Pasta as CodPasta, Nome_Pasta as NomePasta,
        Nivel_Pasta as NivelPasta, Ordem_Pasta as OrdPasta,
        cod_item as CodItem, num_orde as NumItem, num_nive as NivelItem, COD_ITEM_PAI as CodItem_Pai,
        Cod_CC as CodCC, Cod_CC_f as CodCC_f, Nome_Conta_Contabil as NomeCC,
        [Dotação Atualizada] as DA, [Desp. Empenhadas] as DE, [Desp. Liquidadas] as DL, [Desp. Pagas] as DP
FROM dbo.DESPESAS_MUNIC_FUNDEB_2018
UNION ALL
SELECT COD_UF as CodUF, NOM_UF as NomeUF, SIG_UF as SigUF,
        COD_MUNI as CodMunicipio,
        NOM_MUNI as NomeMunicipio,
        Classif_Pasta as GrupoDespesa,
        Codigo_subfuncao_Pasta as CodSubFuncao,
        Nome_Programa as NomePrograma,
        CodPasta_PAI as CodPasta_Pai, Past_Pai as NomePasta_Pai,
        Cod_Exib_Pasta as CodPasta, Nome_Pasta as NomePasta,
        Nivel_Pasta as NivelPasta, Ordem_Pasta as OrdPasta,
        cod_item as CodItem, num_orde as NumItem, num_nive as NivelItem, COD_ITEM_PAI as CodItem_Pai,
        Cod_CC as CodCC, Cod_CC_f as CodCC_f, Nome_Conta_Contabil as NomeCC,
        [Dotação Atualizada] as DA, [Desp. Empenhadas] as DE, [Desp. Liquidadas] as DL, [Desp. Pagas] as DP
FROM dbo.DESPESAS_MUNIC_VINC_2018
'''
```

APÊNDICE E – *Scripts* para seleção de dados da base do INEP

Os quantitativos de matrículas, por modalidade de ensino e para cada município, foram obtidos do próprio sistema SIOPE, fornecidos pelo INEP.

Tabela 13 – Descrição dos campos de quantitativo de matrículas

NUM_MATR_361	Número de alunos matriculados no Ensino Fundamental
NUM_MATR_362	Número de alunos matriculados no Ensino Medio
NUM_MATR_363	Número de alunos matriculados no Ensino Profissional
NUM_MATR_364	Número de alunos matriculados no Ensino Superior
NUM_MATR_365_1	Número de alunos matriculados na Educação Infantil (creche)
NUM_MATR_365_2	Número de alunos matriculados na Educação Infantil (pré-escola)
NUM_MATR_366	Número de alunos matriculados na Educação de Jovens e Adultos
NUM_MATR_367	Número de alunos matriculados na Educação Especial

Fonte: Elaborada pelo autor (2020).

Para a seleção dos quantitativos de escolas e professores, as condições de seleção foram obtidas do documento "Filtros da Educação Básica" do INEP, que trata de instruções para a utilização dos Microdados do Censo da Educação Básica - 2018 (BRASIL, INEP, 2019e).

Os registros de total de escolas por município foram buscados na base do INEP, considerando-se os seguintes filtros:

- Ano Censo: 2018;
- Apenas as Escolas municipais;
- Apenas as Escolas em funcionamento (em atividade); e
- escolas com pelo menos uma matrícula em turma de Escolarização.

Figura 67 – Critérios para seleção das escolas

```
select  i.UF, i.Nome_UF, e.co_municipio, i.Nome_Município,
        count(*) as qtd_escolas_ativas
  from [dbo].[ESCOLAS] e
  inner join db_ibge.[dbo].[LOCALIDADE_2018_Municipio] i
     on e.CO_MUNICIPIO = i.[Código Município Completo]
  where e.nu_ano_censo = 2018
     -- escola em atividade
     and e.TP_SITUACAO_FUNCIONAMENTO = 1
     -- escolas municipais
     and e.TP_DEPENDENCIA = 3
     -- matrícula em turma de Escolarização
     and (e.IN_REGULAR=1 OR e.IN_EJA=1 or e.IN_PROFISSIONALIZANTE=1)
  group by i.UF, i.Nome_UF, e.co_municipio, i.Nome_Município
```

```
-- filtro para escolas do Ensino Fundamental
select i.UF, i.Nome_UF, e.co_municipio, i.Nome_Município,
       count(distinct e.co_entidade) as qtd_escolas_ativas
from [dbo].[ESCOLAS] e
inner join db_ibge.[dbo].[LOCALIDADE_2018_Municipio] i
    on e.CO_MUNICIPIO = i.[Código Município Completo]
inner join [dbo].[MATRICULAS] m
    on m.CO_ENTIDADE = e.CO_ENTIDADE
where e.nu_ano_censo = 2018
    -- escola em atividade
    and e.TP_SITUACAO_FUNCIONAMENTO = 1
    -- escolas municipais
    and e.TP_DEPENDENCIA = 3
    -- matrícula em turma de Escolarização
    and (e.IN_REGULAR=1 OR e.IN_EJA=1 or e.IN_PROFISSIONALIZANTE=1)
    -- ensino fundamental + EJA (EF)
    and m.TP_ETAPA_ENSINO in (4,5,6,7,8,9,10,11,14,15,16,17,18,19,20,21,41,
                              65,67,69,70,73,74)
group by i.UF, i.Nome_UF, e.co_municipio, i.Nome_Município
```

Fonte: Elaborada pelo autor (2020).

Os registros de total de professores por município foram buscados na base do INEP, considerando-se os seguintes filtros:

- Ano Censo: 2018;
- Apenas as Escolas municipais;
- Não são contabilizados docentes em turmas de Atividade Complementar ou de Atendimento Educacional Especializado (AEE);
- Apenas funções de Docente/ Docente Titular coordenador de tutoria (de módulo ou disciplina) – EaD; e
- Os docentes podem atuar em várias escolas ou várias turmas de mesma escola, mas são contados apenas uma única vez em cada município.

Figura 68 – Critérios para seleção dos professores

```
select   i.UF, i.Nome_UF, d.co_municipio, i.Nome_Município,
         count(distinct id_docente) as qtd_docentes
    from [dbo].[DOCENTES] d
    inner join db_ibge.[dbo].[LOCALIDADE_2018_Municipio] i
         on d.CO_MUNICIPIO = i.[Código Município Completo]
    where    d.nu_ano_censo = 2018
       -- escolas municipais
        and  d.TP_DEPENDENCIA = 3
        -- não considerar turmas de Atividade Complementar e de AEE
        and  d.TP_TIPO_TURMA NOT IN (4,5)
         -- funções de Docente/ Docente Titular
        AND d.TP_TIPO_DOCENTE IN (1,5)
    group by  i.UF, i.Nome_UF, d.co_municipio, i.Nome_Município
```

Fonte: Elaborada pelo autor (2020).

APÊNDICE F – Código do gráfico de agrupamento de despesas pagas

O código abaixo exibe os comandos para a geração do gráfico de agrupamento de valores de despesas pagas por grupo de despesa, em cada UF. Os demais gráficos seguem códigos similares (presentes no caderno TCC_AnaliseExploratoria_MDE_EF).

```
count_gd_uf= df_mde.groupby(['SigUF','GrupoDespesa'])['DP'].count()
perc_gd_uf= df_mde.groupby(['SigUF','GrupoDespesa'])['DP'].count()/df_mde.groupby(['SigUF'])['DP'].count()*100
soma_gd_uf = df_mde.groupby(['SigUF','GrupoDespesa'])['DP'].sum()/1000000
soma_perc_gd_uf = df_mde.groupby(['SigUF','GrupoDespesa'])['DP'].sum()/ df_mde.groupby(['SigUF'])['DP'].sum()*100

soma_gd_uf = pd.DataFrame({ 'Total registros': count_gd_uf,
                            'Total em %': perc_gd_uf,
                            'Soma DP': soma_gd_uf,
                            'Soma em %':soma_perc_gd_uf}).reset_index()

print('Total de registros de despesas MDE: ', df_mde.shape[0])
print('Soma dos valores das despesas pagas em MDE (2018): ', round((df_mde['DP'].sum()/1000000),2), ' milhões')
print('\nSoma dos valores por Grupo de Despesa (em milhões) / UF: ')
print('Obs: Os percentuais de total e soma são agrupados por Estado.')

display(soma_gd_uf.sort_values(ascending=True, by='SigUF'))

print()

f, axes = plt.subplots(1, 1,  figsize=(20, 12))

# Configurações do gráfico
major_ticks = np.arange(0, 10000, 1000)
minor_ticks = np.arange(0, 10000, 250)
axes.set_yticks(major_ticks)
axes.set_yticks(minor_ticks, minor=True)
axes.grid(which='both', axis='y')
axes.set_title("Valores das Despesas Pagas por Grupo de Despesas em cada Estado (em milhões)", fontsize=15)
axes.set_xticklabels(axes.get_xticklabels(), rotation=40, ha="right", fontsize=12)

# Dados para a linha da média da Soma DP
x = major_ticks
y_mean = [soma_gd_uf['Soma DP'].mean()]*len(x)

# Incluir linha da média
sns.lineplot(x, y_mean, ax=axes, color='purple', legend='brief', label='Media da Soma DP')

# Incluir as categorias de despesas em barras
sns.catplot(x='SigUF', y='Soma DP', hue='GrupoDespesa',
            data=soma_gd_uf.sort_values(ascending=False, by='Soma DP'), kind='bar', ax=axes)

legend = axes.legend(loc='center', fontsize=15)

plt.close(2)
plt.show()

print('Media da Soma DP: ', round(soma_gd_uf['Soma DP'].mean(),1), ' milhões')
print('Mediana da Soma DP: ', round(soma_gd_uf['Soma DP'].median(),1), ' milhões')
print('Mínimo da Soma DP: ', round(soma_gd_uf['Soma DP'].min(),1), ' milhões')
print('Máximo da Soma DP: ', round(soma_gd_uf['Soma DP'].max(),1), ' milhões')

print()
print('Faixa de percentuais de cada grupo de despesa: ')
perc_min = soma_gd_uf.groupby('GrupoDespesa')['Soma em %'].min()
perc_max = soma_gd_uf.groupby('GrupoDespesa')['Soma em %'].max()
display(pd.DataFrame({'Mínimo': perc_min, 'Máximo' : perc_max  }).reset_index())
```

APÊNDICE G – Código do gráfico de dispersão de despesas pagas

Os códigos abaixo exibem os comandos para a geração dos gráficos de dispersão dos valores de despesas pagas por variáveis categóricas (UF, Região, Grupo de Despesa, Nome da Pasta e tipo de gasto). Os demais gráficos seguem códigos similares (presentes no caderno TCC_AnaliseExploratoria_MDE_EF).

```
features_categoricas = ['SigUF','Regiao', 'GrupoDespesa', 'NomePasta_Pai','NomePasta','Tipo de Gasto']
ix = 1

fig = plt.figure(figsize = (55,12))

# c = para cada coluna
for c in list(df_mde[features_categoricas]):
    if ix <= 3:
        ax2 = fig.add_subplot(2,3,ix+2)
        my_order = df_mde.groupby(by=c)['DP'].median().sort_values(ascending=False).iloc[::-1].index
        sns.stripplot(x=df_mde[c], y=(df_mde['DP']), ax=ax2, order=my_order, linewidth=0.5)
        ax2.set(ylabel="DP")

    locs, labels = plt.xticks()
    plt.setp(labels, rotation=90, fontsize=12)

    ix = ix +1
    if ix == 2:
        fig = plt.figure(figsize = (55,12))
        ix =1

plt.tight_layout()
```

```
df_selecao = df_mde[(df_mde['DP'] < (250000000))]

features_categoricas = ['SigUF','Regiao','GrupoDespesa','NomePasta_Pai','NomePasta','Tipo de Gasto']
ix = 1

fig = plt.figure(figsize = (55,12))

# c = para cada coluna
for c in list(df_selecao[features_categoricas]):
    if ix <= 3:
        ax2 = fig.add_subplot(2,3,ix+2)
        my_order = df_selecao.groupby(by=c)['DP'].median().sort_values(ascending=False).iloc[::-1].index
        sns.stripplot(x=df_selecao[c], y=(df_selecao['DP']), ax=ax2, order=my_order, linewidth=0.5)
        ax2.set(ylabel="DP")

    locs, labels = plt.xticks()
    plt.setp(labels, rotation=90, fontsize=12)

    ix = ix +1
    if ix == 2:
        fig = plt.figure(figsize = (55,12))
        ix =1

plt.tight_layout()
```

APÊNDICE H – Código para o cálculo do coeficiente de *Spearman*

Método *Spearman*: Grupos de Despesas com Indicadores e Modalidade de Ensino

```
fig, ax = plt.subplots(figsize=(15,15))

colunas_interesse = ['DespFUNDEB', 'DespProp', 'DespVinc',
                     'Pop_estimada', 'QtdEscolas', 'QtdDocentes', 'NUM_MATR_361',
                     'IDHM', 'IDHM_E', 'IDHM_L', 'IDHM_R', 'IDEB_AI', 'IDEB_AF',
                     'TxEvasao_EF', 'CustoAluno', 'DespesaProf',
                     'EducEspecial', 'EducJA', 'EnsFund', 'EnsFund_exc']

cm = df_consol[colunas_interesse].corr(method='spearman')

sns.set(font_scale=1)
sns.heatmap(cm,
            cmap='coolwarm',
            annot=True,
            annot_kws={'size': 12},
            square=True,
            fmt='.2f',
            linewidths=1,
            cbar=False,
            ax=ax);
```

Método *Spearman*: Grupos de Despesas com Subfunções e Modalidade de Ensino

```
fig, ax = plt.subplots(figsize=(15,15))

colunas_interesse = ['DespFUNDEB', 'DespProp', 'DespVinc',
                     'EducEspecial', 'EducJA', 'EnsFund', 'EnsFund_exc',
                     'AdmFinanc', 'AdmGeral', 'MerEscolar', 'ComunSocial',
                     'FormRH', 'OutrosEE', 'PlanOrc', 'ProtBenefTrab', 'ServDivInt', 'TI', 'TranspEsc']

cm = df_consol[colunas_interesse].corr(method='spearman')

sns.set(font_scale=1)
sns.heatmap(cm,
            cmap='coolwarm',
            annot=True,
            annot_kws={'size': 12},
            square=True,
            fmt='.2f',
            linewidths=1,
            cbar=False,
            ax=ax);
```

Método *Spearman*: Grupos de Despesas com Tipo de Gasto e Modalidade de Ensino

```
fig, ax = plt.subplots(figsize=(12,12))

colunas_interesse = ['DespFUNDEB', 'DespProp', 'DespVinc',
                     'EducEspecial', 'EducJA', 'EnsFund', 'EnsFund_exc',
                     'tgRemun', 'tgFormacao', 'tgDidatico', 'tgAlim', 'tgTransp', 'tgManut', 'tgInvest', 'tgConv', 'tgOutros']

cm = df_consol[colunas_interesse].corr(method='spearman')

sns.set(font_scale=1)
sns.heatmap(cm,
            cmap='coolwarm',
            annot=True,
            annot_kws={'size': 12},
            square=True,
            fmt='.2f',
            linewidths=1,
            cbar=False,
            ax=ax);
```

APÊNDICE I – Código para comparar diferentes *scalers*

Os códigos abaixo exibem os comandos para a geração dos gráficos de dispersão dos dados de População com Quantidade de matrículas, cada gráfico em diferentes escalonamento de variáveis (*Standard Scaler*, *Robust Scaler e MinMax Scaler*), com uso de *K-Means*.

```
# Dataframe de Municipios com dados de municipios, população e indicadores
df_mun_ind = pd.concat([df_consol.ix[:,2:8], df_consol.ix[:,12:23]], axis=1, sort=False)
lista_colunas = df_mun_ind.columns.values
lista_colunas_num = lista_colunas[6:]

# gerar dataframe com os dados a normalizar (Somente colunas numericas)
df_mun_ind_num = df_mun_ind[lista_colunas_num]

# Criar o modelo scaler e executar fit nos dados numéricos brutos
scaler_Standard = StandardScaler().fit(df_mun_ind_num)
scaler_Robust = RobustScaler().fit(df_mun_ind_num)
scaler_MinMax = MinMaxScaler().fit(df_mun_ind_num)

# Aplicar o modelo scaler e Gerar as colunas normalizadas
df_mun_ind_std = df_mun_ind.copy()
df_mun_ind_rbt = df_mun_ind.copy()
df_mun_ind_mm = df_mun_ind.copy()

colunas_std = scaler_Standard.transform(df_mun_ind_num)
colunas_rbt = scaler_Robust.transform(df_mun_ind_num)
colunas_mm = scaler_MinMax.transform(df_mun_ind_num)

df_mun_ind_std[lista_colunas_num] = colunas_std
df_mun_ind_rbt[lista_colunas_num] = colunas_rbt
df_mun_ind_mm[lista_colunas_num] = colunas_mm

# Criar modelo KMeans e aplicar aos dataframes numericos
model_kmeans = KMeans(n_clusters_global, random_state=random_state)

labels_std = model_kmeans.fit_predict(df_mun_ind_std.iloc[:,6:])
labels_rbt = model_kmeans.fit_predict(df_mun_ind_rbt.iloc[:,6:])
labels_mm = model_kmeans.fit_predict(df_mun_ind_mm.iloc[:,6:])
```

```
fontsize=14
a = 0.6
f, axes = plt.subplots(3, 2, figsize=(18, 14))

axes[0,0].scatter(x=colunas_std[:,0], y=colunas_std[:,7] , c=labels_std, cmap='Spectral', s=s_2D,
                  edgecolor=edgecolor, alpha=a)
axes[0,0].set_title("Standard Scaler - População x Qtd Matrículas", fontsize=fontsize)
axes[0,0].set(xlabel="População"), axes[0,0].set(ylabel="Qtd Matrículas")

axes[0,1].scatter(x=colunas_std[:,0], y=colunas_std[:,7] , c=labels_std, cmap='Spectral', s=s_2D,
                  edgecolor=edgecolor, alpha=a)
axes[0,1].set_title("Standard Scaler - População x Qtd Matrículas (zoom)", fontsize=fontsize)
axes[0,1].set(xlabel="População"), axes[0,1].set(ylabel="Qtd Matrículas")
axes[0,1].set_xlim(0,10), axes[0,1].set_ylim(0,10)

axes[1,0].scatter(x=colunas_rbt[:,0], y=colunas_rbt[:,7] , c=labels_rbt, cmap='Spectral', s=s_2D,
                  edgecolor=edgecolor, alpha=a)
axes[1,0].set_title("Robust Scaler - População x Qtd Matrículas", fontsize=fontsize)
axes[1,0].set(xlabel="População"), axes[1,0].set(ylabel="Qtd Matrículas")

axes[1,1].scatter(x=colunas_rbt[:,0], y=colunas_rbt[:,7] , c=labels_rbt, cmap='Spectral', s=s_2D,
                  edgecolor=edgecolor, alpha=a)
axes[1,1].set_title("Robust Scaler - População x Qtd Matrículas (zoom)", fontsize=fontsize)
axes[1,1].set(xlabel="População"), axes[1,1].set(ylabel="Qtd Matrículas")
axes[1,1].set_xlim(0,100), axes[1,1].set_ylim(0,50)

axes[2,0].scatter(x=colunas_mm[:,0], y=colunas_mm[:,7] , c=labels_mm, cmap='Spectral', s=s_2D,
                  edgecolor=edgecolor, alpha=a)
axes[2,0].set_title("MinMax Scaler - População x Qtd Matrículas", fontsize=fontsize)
axes[2,0].set(xlabel="População"), axes[2,0].set(ylabel="Qtd Matrículas")

axes[2,1].scatter(x=colunas_mm[:,0], y=colunas_mm[:,7] , c=labels_mm, cmap='Spectral', s=s_2D,
                  edgecolor=edgecolor, alpha=a)
axes[2,1].set_title("MinMax Scaler - População x Qtd Matrículas (zoom)", fontsize=fontsize)
axes[2,1].set(xlabel="População"), axes[2,1].set(ylabel="Qtd Matrículas")
axes[2,1].set_xlim(0, 0.2), axes[2,1].set_ylim(0, 0.2)

plt.subplots_adjust(left=None, bottom=None, right=None, top=None, wspace=None, hspace=0.3)
```

APÊNDICE J – Código para gráficos com Coeficiente de Silhueta

Os códigos abaixo exibem os comandos para a geração dos gráficos de dispersão dos dados de População com Quantidade de matrículas, com a informação do tamanho de cada cluster e o respectivo Coeficiente de Silhueta, considerando-se diferentes valores para o k (número de clusters).

```
X = df_mun_ind_rbt.iloc[:,6:].values     # dados numericos normalizados
range_n_clusters = [2, 3, 6, 7, 8, 9, 10, 12]
colors_range = cm.Paired

for n_clusters in range_n_clusters:

    fig, (ax1, ax2, ax3) = plt.subplots(1, 3)
    fig.set_size_inches(18, 6)

    # ------------------------------------------------------------
    # 1st Plot - Silhouette plot
    # ------------------------------------------------------------
    # Coeficiente silhouette possui intervalo de -1 a 1
    ax1.set_xlim([-0.2, 0.8])

    # Iniciar o modelo para cada valor de n_clusters
    clusterer = KMeans(n_clusters=n_clusters, random_state=random_state)
    cluster_labels = clusterer.fit_predict(X)

    # O silhouette_score_avg fornece o valor médio de todos os pontos,
    # proporcionando uma visão das densidades e separação dos grupos
    silhouette_avg = silhouette_score(X, cluster_labels)
    print("Para valor de n_clusters =", n_clusters,
          "A média do valor de score (silhouette_score) é :", silhouette_avg)

    # Calcular o silhouette score de cada ponto
    sample_silhouette_values = silhouette_samples(X, cluster_labels)

    y_lower = 200
    for i in range(n_clusters):
        # Agrega o valor de score para os pontos pertencendo ao cluster i e coloca em ordem
        ith_cluster_silhouette_values = sample_silhouette_values[cluster_labels == i]
        ith_cluster_silhouette_values.sort()

        size_cluster_i = ith_cluster_silhouette_values.shape[0]
        y_upper = y_lower + size_cluster_i

        color = colors_range(float(i) / n_clusters)
        ax1.fill_betweenx(np.arange(y_lower, y_upper),
                          0, ith_cluster_silhouette_values,
                          facecolor=color, edgecolor=color, alpha=0.7)
```

```
    # Insere no eixo y o numero do Cluster para cada Silhouette plot
    ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

    # Computa o novo valor de y_lower para o próximo gráfico (plot)
    y_lower = y_upper + 200  # 200 for the 0 samples

# Titulos
ax1.set_title("Silhouette Plots para todos os clusters.")
ax1.set_xlabel("Valores dos coeficientes (Silhouette)")
ax1.set_ylabel("Número do Cluster")

# A linha vertical em vermelho representa o average silhouette score
ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

ax1.set_yticks([])
ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8])

# ------------------------------------------------------------
# Configurações dos gráficos 2 e 3
# ------------------------------------------------------------
plt.suptitle(("Análise Silhouette para agrupamentos KMeans com n_clusters = %d" % n_clusters),
                 fontsize=16, fontweight='bold')

colors = colors_range(cluster_labels.astype(float) / n_clusters)

# ------------------------------------------------------------
# 2nd Plot: clusters para conjunto de dados sem outliers
# ------------------------------------------------------------
ax2.scatter(X[:, 0], X[:, 7] , c=colors, s=200, edgecolor='k', alpha=0.7)
ax2.set_xlim(-5,50)
ax2.set_ylim(-5,30)

# Rotular/ Marcar os centroides
centers = clusterer.cluster_centers_
# Desenhar circulos nos centroides
ax2.scatter(centers[:, 0], centers[:, 7], marker='o', c="white", alpha=1, s=300, edgecolor='k')

for i, c in enumerate(centers):
    ax2.scatter(c[0], c[7], marker='$%d$' % i, alpha=0.7, s=80, edgecolor='k')

# Titulos
ax2.set_title("Visualização Agrupamentos (s/ outliers extremos).")
ax2.set_xlabel("População")
ax2.set_ylabel("Qtd Matrículas")
```

```
# ------------------------------------------------------------
# 3rd Plot: clusters para outliers
# ------------------------------------------------------------
ax3.scatter(X[:, 0], X[:, 7] , c=colors, s=200, edgecolor='k', alpha=0.7)
ax3.set_xlim(50,650)
ax3.set_ylim(30,240)

# Rotular/ Marcar os centroides
centers = clusterer.cluster_centers_
# Desenhar circulos nos centroides
ax3.scatter(centers[:, 0], centers[:, 7], marker='o', c="white", alpha=0.2, s=300, edgecolor='b')
for i, c in enumerate(centers):
    ax3.scatter(c[0], c[7], marker='$%d$' % i, alpha=0.5, s=50, edgecolor='k')

# Titulos
ax3.set_title("Visualização Agrupamentos (outliers extremos).")
ax3.set_xlabel("População")
ax3.set_ylabel("Qtd Matrículas")
```

APÊNDICE K – Código para detecção de anomalias (grupo de despesa)

```
df_sample = df6_std.copy()
X1 = df_sample['DespProp'].values.reshape(-1,1)
X2 = df_sample['DespFUNDEB'].values.reshape(-1,1)
X = np.concatenate((X1,X2),axis=1)
col1 = 'DespProp'
col2 = 'DespFUNDEB'
outliers_fraction = 0.01 # 1% de outliers

xx , yy = np.meshgrid(np.linspace(0, 1, 100), np.linspace(0, 1, 100))

for i, (clf_name, clf) in enumerate(classifiers.items()):
    clf.fit(X)
    # cálculo do score de anomalia de cada Município
    scores_pred = clf.decision_function(X) * -1
    # determinar se é inlier (0) ou outlier (1)
    y_pred = clf.predict(X)
    # contagem de pontos normais e pontos outliers
    n_inliers = len(y_pred) - np.count_nonzero(y_pred)
    n_outliers = np.count_nonzero(y_pred == 1)

    # copy of dataframe
    df_out = df_sample
    df_out['outlier'] = y_pred.tolist()

    # lista dos pontos inlier
    IX1 =  np.array(df_out[col1][df_out['outlier'] == 0]).reshape(-1,1)
    IX2 =  np.array(df_out[col2][df_out['outlier'] == 0]).reshape(-1,1)
    # lista dos pontos outlier
    OX1 =  df_out[col1][df_out['outlier'] == 1].values.reshape(-1,1)
    OX2 =  df_out[col2][df_out['outlier'] == 1].values.reshape(-1,1)

    print('Algoritmo utilizado:', clf_name,
          '\nQuantidade de OUTLIERS:', n_outliers,
          '\nQuantidade de INLIERS:',n_inliers,
          '\nPercentual Outliers:', round((n_outliers/n_inliers)*100,2), '%','\n')

    plt.figure(figsize=(10, 10))

    # threshold: limite para converter o score em indicação de anômalo ou não
    threshold = stats.scoreatpercentile(scores_pred, 100 * outliers_fraction)

    # calculo do anomaly score para todos os pontos
    Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()]) * -1
    Z = Z.reshape(xx.shape)

    # fill colormap -from minimum anomaly score to threshold value
    plt.contourf(xx, yy, Z, levels=np.linspace(Z.min(), threshold,
7),cmap='YlGnBu_r')
    # draw red contour line where anomaly score is equal to thresold
    a = plt.contour(xx, yy, Z, levels=[threshold], linewidths=1, colors='red')
    # fill pink contour lines - range from threshold to max anomaly score
    plt.contourf(xx, yy, Z, levels=[threshold, Z.max()],colors='pink')

    b = plt.scatter(IX1,IX2, c='white',s=40, edgecolor='b')
    c = plt.scatter(OX1,OX2, c='black',s=40, edgecolor='w')
    plt.axis('tight')

    plt.legend(  [a.collections[0], b,c],
        ['learned decision function', 'inliers','outliers'],
        prop=matplotlib.font_manager.FontProperties(size=15), loc='upper right')

    plt.xlim((0, 1)), plt.ylim((0, 1)), plt.title(clf_name)
    plt.show()
```

APÊNDICE L – Código para histogramas e KDE de municípios anômalos

```
#=========================================
#  localização pontos anômalos
#=========================================
fig = plt.figure(figsize = (10,8))
ax = fig.add_subplot(111, projection='3d')

xs1 = df6_n_all['NUM_MATR_361']
ys1 = df6_n_all['QtdDocentes']
zs1 = df6_n_all['Pop_estimada']

xs2 = df6_o_all['NUM_MATR_361']
ys2 = df6_o_all['QtdDocentes']
zs2 = df6_o_all['Pop_estimada']

g = ax.scatter(xs2, ys2, zs2, s=250, c = 'r', edgecolor=edgecolor, alpha=1, label='outliers', zorder=10)
g = ax.scatter(xs1, ys1, zs1, s=150, c='b' , edgecolor='w', alpha=0.4, label='normal', zorder=2)
g = ax.legend(fontsize=16, loc='best')

ax.set_xlabel('Qtd Matriculas'), ax.set_ylabel('Qtd Docentes'), ax.set_zlabel('População')
ax.set_title('Cluster 6', fontsize=14)
plt.show()

#=========================================
#  localização somente pontos anômalos
#=========================================
fig = plt.figure(figsize = (10,8))
ax = fig.add_subplot(111, projection='3d')

xs2 = df6_o_all['NUM_MATR_361']
ys2 = df6_o_all['QtdDocentes']
zs2 = df6_o_all['Pop_estimada']

g = ax.scatter(xs2, ys2, zs2, s=250, c = 'r', edgecolor=edgecolor, alpha=1, label='outliers')
g = ax.legend(fontsize=16, loc='best')

ax.set_xlabel(''), ax.set_ylabel(''), ax.set_zlabel('')
ax.set_title('Cluster 6', fontsize=14)

for x, y, z, label in zip(xs2, ys2, zs2, df6_o_all['NomeMunicipio']):
    ax.text(x-500, y+0.5, z+2500, label)


plt.show()
```

```
# Gráficos de histogramas e KDE
# Azul: municípios normais; Vermelho: os 10 municípios anômalos

n_bins = 5; alpha = 0.4
list_colunas = ['Pop_estimada', 'QtdEscolas', 'QtdDocentes', 'NUM_MATR_361']

ix=1
fig = plt.figure(figsize = (25, 12))

for c in list_colunas:
    if ix <=2:
        ax2 = fig.add_subplot(2,3, ix+1)

        sns.distplot(df6_sample_n[c], bins=n_bins, color='b', kde=True, hist=True, label='Normal')
        sns.distplot(df6_sample_o[c], bins=n_bins, color='r', kde=True, hist=True, label='Outliers')

        ax2.set_xlabel("valores reais", size=15)
        ax2.set_ylabel("densidade de probabilidade", size=15)
        ax2.set_title('\n'.join(textwrap.wrap(c, 40)), size=18)
        ax2.legend(fontsize=15, loc='best')

        xlocs, xlabels = plt.xticks(); ylocs, ylabels = plt.yticks()
        plt.setp(xlabels, rotation=60, fontsize=14)
        plt.setp(ylabels, fontsize=12)

    ix = ix +1
    if ix == 3:
        fig = plt.figure(figsize = (25,12))
        ix =1
plt.tight_layout()
```