A implementation mathematical-computational method for the detection and treatment of financial outliers in higher education

Nathan C. Freitas, Roberta M. M. Gouveia, Gabriel A. de Albuquerque Júnior, Maria da Conceição M. Batista, Rodrigo Lins Rodrigues

Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brasil

{nathan.freitas, roberta.gouveia, gabriel.alves, maria.cmbatista, rodrigo.linsrodrigues}@ufrpe.br

Abstract. *The Higher Education Census occurs annually, collecting data from public and private Higher Educational Institutions (HEI) in Brazil. Different factors can lead to anomalies or outliers in some of these collected data. This work proposes a mathematical-computational method to detect and treat atypical HEI's financial values. Both univariate and bivariate analysis to that end. We analyzed the expenses and incomes of HEI in the census from 2016 to 2019. This analysis revealed that 204 out of 2,224 HEI, approximately 10%, reported some atypical data.*

Resumo. *O Censo da Educação Superior ocorre anualmente, coletando dados de Instituições de Ensino Superior (IES) no Brasil. Diferentes fatores podem levar a anomalias ou outliers em alguns destes dados coletados. Este trabalho propõe um método matemático-computacional para detectar e tratar valores financeiros atípicos das IES. Para tanto, adota-se as análises univariadas e bivariadas dos dados. Foram analisados dados de despesas e receitas das IES do Censo de 2016 a 2019. Esta análise revelou que 204 de 2.224 IES, aproximadamente 10%, reportaram algum dado atípico.*

1. Introdução

A capacidade de descoberta de conhecimento na área de ciência de dados permite que características inerentes ao mundo real sejam observadas por meio de técnicas advindas de várias áreas, tais como estatística, inteligência artificial e mineração de dados (Data Mining). A mineração de dados é um processo de explorar grandes volumes de dados com vistas a descobrir padrões e correlações úteis e consistentes, desde que tais bases de dados apresentem informações relevantes e suficientes para entendimento de um determinado contexto/cenário a ser analisado [Witten 2011].

No âmbito da educação, existe uma terminologia específica conhecida como mineração de dados educacionais (EDM, do inglês *Educational Data Mining*) para que se possam entender possíveis particularidades – como custos financeiros, fatores socioeconômicos, acessibilidade, entre outras – existentes na educação brasileira que podem fundamentar a aplicação de políticas públicas, além de contribuir para o trabalho de gestores, instituições, pesquisadores, especialistas e estudantes.

Dentre as várias análises e perspectivas exploradas em EDM, tem-se o estudo de fatores que podem influenciar o desempenho do aluno. Há trabalhos que utilizam de modelos preditivos para identificação de características relevantes através de modelos de regressão [Silva *et al.* 2019] e de classificação [Pinto *et al.* 2020]. Um outro cenário é em relação à obtenção de características que possam levar à evasão estudantil em uma instituição educacional. [Carrano *et al.* 2019].

No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), vinculado ao Ministério da Educação (MEC), é responsável pela extração e disponibilização de evidências educacionais. Uma das modalidades de avaliação e coleta de dados é o Censo da Educação Superior, que é realizado anualmente e serve como um repositório amplo para pesquisas sobre as Instituições de Educação Superior (IES) brasileiras. Informações persistidas no censo estão desde um aluno em relação a um curso, assim como a infraestrutura da própria instituição, como por exemplo, o valor de despesa e receitas das IES.

O estudo da integridade dos dados financeiros das instituições superiores pode servir de base a diversos estudos em relação ao aluno, principalmente quando refere-se ao seu desempenho acadêmico. O Brasil é um país continental, onde a necessidade de distribuição de investimento para todo território nacional é imprescindível diante do direito constitucional à educação com nível mais alto de ensino para qualquer cidadão, segundo o artigo 6 e 208, inciso V, da instituição . Perantes as instituições privadas esse direito deve ser prioridade de seus gestores, enquanto para públicas é obrigação o exercer a garantia desse direito de forma que a isonomia seja cumprida no país inteiro.

A Lei de Acesso à Informação (LAI) garante como direito o acesso à informação pública de forma que seja possível a participação popular para a busca de uma melhor gestão pública, isso inclui os dados presentes no Censo Superior. A transparência e integridade dos dados disponibilizados torna pesquisas relacionadas mais precisas e incentiva um processo contínuo de melhorias de fatores que possam auxiliar no desempenho do estudante que por consequência melhora a qualidade dos cursos, docentes e instituições. Visto que, uma vez identificado deficiência em um processo de ensino-aprendizagem é necessário analisar se o investimento relacionado é adequado.

Em relação às características que podem influenciar o percurso e desempenho do aluno, têm-se as variáveis financeiras das IES que são disponibilizadas pelo Censo da Educação Superior. Os valores de despesas e receitas podem ser relevantes para análises que envolvem a instituição, e consequentemente o desenvolvimento acadêmico do aluno. Dessa forma, se tais dados são devidamente utilizados com técnicas de mineração de dados, podem fornecer informações valiosas e indícios de como estes recursos financeiros devem ser aplicados. Para isso é necessário que os dados financeiros estejam concisos diante do contexto – como localização – em que estão inseridos, sendo necessário observar estruturas inconsistentes através de uma análise de anomalias, assim como diagnosticar e tratar (quando necessário) valores discrepantes/atípicos (outliers).

Em relação à detecção e tratamento de outliers no contexto financeiro de IES brasileiras, vale destacar que há casos em que os valores de receitas e/ou despesas da instituição são detectados como atípicos, porém não se faz necessário o respectivo tratamento. Nesses casos, os valores financeiros registrados pela IES não são retificados, visto que representam informações factíveis com a realidade e especificidade da IES, o que justifica os valores informados serem muito acima ou muito abaixo em relação às demais IES.

Na identificação de *outliers* um dos métodos mais comuns utilizados é utilizar o método de quartil que pode demonstrar bons resultados muito bem em relação a encontrar esses extremos [Borges e De Siqueira Gê 2020]. Mas um ponto a ser levado em consideração nessa abordagem é a distribuição assimétrica dos dados financeiros e não há tratamento das escalas dos dados, que por vezes é necessário para evidenciar melhor as instâncias financeiras. Outras abordagens de detecção podem ser utilizadas, pode-se utilizar técnicas de escalonamentos – transformação dos dados – junto com algoritmos de tais como aprendizado de máquina não supervisionado, por meio de agrupamentos/clusterização para criação de amostras mais coesas, por exemplo algoritmos de distâncias como agrupamento hierárquico, que aplicadas a algoritmos não supervisionados de detecção, como o *Local Outlier Factor*, podem auxiliar a identificar instâncias discrepantes [Machado 2022].

Diante do exposto, este trabalho tem como objetivo desenvolver um estudo analítico de investigação e diagnóstico de padrões em dados abertos da educação superior brasileira, em específico, aplicando técnicas de mineração de dados de forma a detectar e tratar variáveis atípicas para que possam ser identificadas possíveis instituições com características discrepantes. Utilizando, para isso, técnicas de detecção através da interseção de abordagens univariadas e bivariadas de dados. Além de detectar as IES que possuem valores financeiros discrepantes em relação às demais IES presentes no Censo da Educação Superior, o artigo apresenta um método desenvolvido pelos autores para o tratamento de outliers de variáveis financeiras baseado em três estratégias: **mediana**, **média** e **valor multiplicativo**.

Na seção 2 será descrito todo método utilizado para preparação dos dados desde a seleção, pré-processamento e transformação dos dados. Na seção 3 será apresentado os métodos de detecção e tratamento utilizados. Na seção 4 serão abordados os resultados do problema em si, com toda a construção dos agrupamentos para detecção e o tratamento das variáveis atípicas apresentando sua quantidade de instâncias detectadas e algumas características que elas apresentam.

2. Metodologia

A metodologia do *Knowledge Discovery in Databases* (KDD) é um processo no qual se extrai conhecimento que estava ocultado em uma base de dados. Suas etapas são divididas em Seleção, Pré-processamento, Transformação, Data Mining, Validação/Interpretação e Conhecimento [Fayyad 1996]. Neste trabalho foram aplicadas as três primeiras etapas do KDD que se trata de uma fase de preparação dos dados para realização da detecção e tratamento das variáveis financeiras.

A base de dados utilizada foi do Censo da Educação Superior dos anos de 2016 a 2019. Inicialmente foram extraídas as quatros bases existentes no censo: aluno, docente, curso e IES. Como objetivo do trabalho está sob a perspectiva das IES, as outras 3 bases foram utilizadas para obtenção das respectivas quantidades (em cada IES) de alunos, docentes e cursos, sendo essas informações incluídas na base de dados IES.

A etapa de pré-processamento é necessária para buscar a integridade dos dados nas bases, assim investigando possíveis inconsistências e dados nulos, a fim de tratá-los para que sejam utilizados em um método de detecção. Para o tratamento dos atributos nulos das instâncias foi utilizado cálculo de valores de substituição – utilização da média, mediana, moda ou regressão para definir os valores faltantes – ou por

substituição por valores conhecidos - quando os dados podem ser substituídos por uma carta marcada [Hair *et al.* 2009].

Ainda na etapa de pré-processamento, foram identificadas algumas inconsistências. Há instituições que mudam de categoria administrativa no decorrer dos 4 anos e como forma de padronização para criação das amostras foi escolhida a categoria mais frequente nos anos (moda). Outra característica presente nos dados do Censo da Educação Superior diz respeito à variável "referente financeiro" que representa se a instituição é mantida financeiramente pela sua mantenedora ou por ela mesma. Em poucas instâncias 0,19%, a referida variável foi registrada incorretamente pela IES. Tais IES são consideradas mantidas por ela mesma, mas as variáveis financeiras indicam que ela possui característica de ser mantida pela mantenedora vinculada.

Quanto às IES mantidas financeiramente pela mantenedora, as características presentes nessas instituições são os valores financeiros replicados (valores agregados) para essas IES, quando na verdade deveriam ter sido distribuídos adequadamente conforme as despesas e receitas reais da IES. Em relação a essas instituições, foi considerado apenas um tratamento pela quantidade de alunos, a ser discutido na Seção 4.

As instituições possuem 10 atributos financeiros que possuem as seguintes quantidades de valores zerados em cada coluna, conforme ilustra a Figura 1.

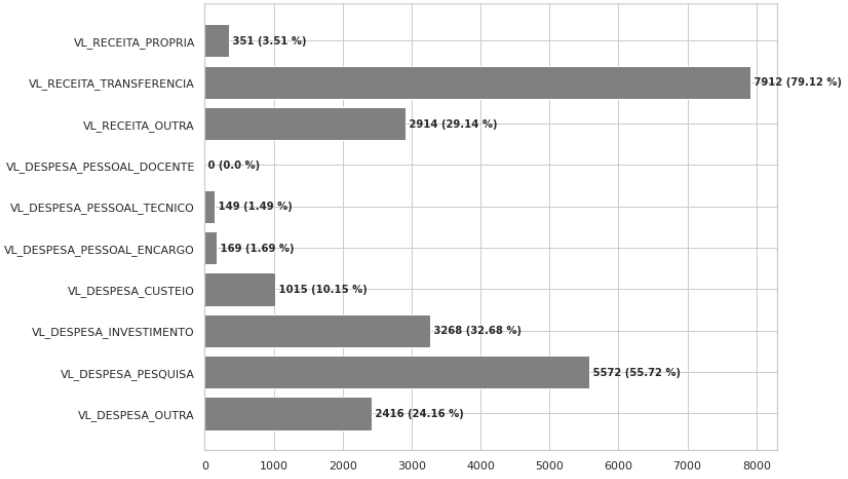

Figura 1. Quantidade de valores zerados em cada atributo financeiro.

Um dos motivos para esses atributos apresentarem valores zerados, refere-se a maneira que cada IES registra tais informações, não havendo uma padronização. Algumas IES registram os valores de receitas e despesas em apenas algumas das colunas indicadas na Figura 1, ou seja, não distribuem os valores nas 10 colunas. Por exemplo, uma determinada IES pode registrar suas receitas apenas no atributo "VL_RECEITA_PROPRIA", deixando as outras duas receitas zeradas. Na presença dessa característica, foi abordado a estratégia de juntar as variáveis financeiras em duas variáveis possíveis: receitas e despesas, para que não houvesse essa ambiguidade. A partir disso nenhuma das variáveis criadas diante da soma das receitas e despesas apresentam valores zerados.

Para aplicação da abordagem de identificar valores atípicos deve-se ter a premissa que os dados seguem uma distribuição normal. Para distribuições assimétricas, a aplicação de transformação logarítmica se torna uma boa alternativa para diminuição de assimetria quando ela está à esquerda. Neste trabalho, muitas variáveis apresentavam essas características, como as variáveis financeiras, quantidade de aluno, docente etc.

3. Detecção e tratamento de anomalias

Por vezes alguns atributos apresentam uma instância com uma dispersão muito grande comparado com a sua população. Essa dispersão pode ser benéfica, quando essa dispersão apresenta uma característica realista/factível dos dados apresentados que seja justificável, ou problemática, quando não representa a população. Primeiramente foi levado em consideração o tratamento de valores financeiros com possíveis erros de digitação. Logo após, foram criadas amostras nas instituições de acordo com a variáveis categórica, como categoria administrativa e região, para agrupar instituições com características semelhantes e nessas amostras aplicar o desenvolvido pelos autores do presente artigo.

Para detecção foram utilizadas duas abordagens: univariada e bivariada. Na análise univariada cada atributo é analisado isoladamente, neste trabalho foi utilizado o método Tukey que se baseia nos quartis [Tukey 1997]. Na análise bivariada, dois atributos são examinados para observar possíveis *outliers* através de um algoritmo de distância (Mahalanobis) que calcula a distância de cada instância a partir do centro médio das observações. Porém, ele pode ser enviesado, pois baseado na média que pode ser alterada com a presença de *outliers*. Para isso há métodos robustos desse algoritmo que buscam ter a sensibilidade de desconsiderar os valores outliers, o estimador Covariância de Determinante Mínimo (MCD) é um deles e foi utilizado neste trabalho.

Na abordagem univariada com o atributo quantidade de alunos foi utilizada a estratégia de discretização através de classes [Witten 2011] de forma que sejam criadas amostras por meio dessa nova variável categórica. Na abordagem bivariada foram levadas em consideração as variáveis contínuas para fazer a detecção. O objetivo de utilizar duas abordagens consiste em obter uma confiança maior nas instâncias detectadas como atípicas a partir da estratégia de considerar as que são identificadas em ambas as abordagens. Assim, ao tratar o problema como o conjunto de instâncias observadas como *outliers* no método univariado e bivariado, ao final, consideram-se como outliers aquelas instâncias que pertencem à interseção desses dois métodos.

Após instâncias detectadas como atípicas foi criada uma abordagem de tratamento que segue a Figura 2, a qual é dividida em três tipos:

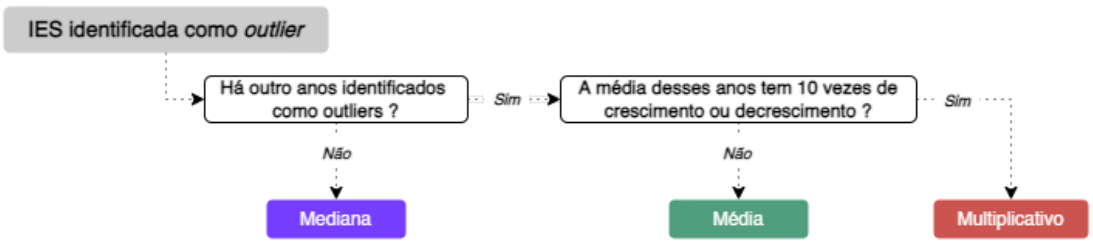

Figura 2. Fluxograma de tipos de tratamentos para instituições identificadas como outliers.

Aplicando mediana: É utilizado a mediana da amostra que a instituição pertence. Por exemplo, aplicar a mediana do grupo de categoria administrativa pública federal, região de todo Brasil e quantidade de alunos inferior a 500.

Aplicando a média: É utilizado a média dos anos não detectados como outliers da instituição. Nesse caso, por exemplo, se o ano de 2016 for considerado como outlier e os anos de 2017 e 2018 não forem outliers e tiverem a média de 50 milhões, esse valor será replicado ao ano de 2016.

Aplicando valor multiplicativo: Se o ano da instituição outlier tiver um crescimento superior a mais de 10 vezes em relação a média dos valores financeiros não outliers, então é aplicado uma redução de unidades do valor outlier aplicando logaritmo na base 10 da divisão desses dois valores.

Inicialmente é checado se as IES identificadas possuem seus outros anos identificados como outliers, caso negativo é aplicado a mediana do agrupamento. Se a média dos anos das IES foram 10 vezes maior ou menor foi utilizado o método multiplicativo que calcula a taxa de crescimento, caso contrário essa média é replicada para a instância discrepante.

4. Resultados e discussão

Na presença de instituições que possuíssem algum dos seus anos com erros de digitação, levou-se em consideração o crescimento ou decrescimento dos valores financeiros de um ano para outro. Caso houvesse uma diferença de 10 vezes – com margem de erro de 10% de tolerância – foi considerado como erro de digitação (adicionadas ou retiradas unidades). No total, 56 instituições foram constatadas com um crescimento anormal entre os anos, sendo que, destas, 31 foram determinadas que de fato possuíam erros de digitação entre seus anos, através de uma análise manual.

Foi feita uma análise de instituições que possuem erro de digitação na variável do “referente financeiro” e foram tratadas, pois aquelas que apresentam o valor financeiro mantido pela mantenedora precisaram ser aplicadas um tratamento nos seus valores financeiros de acordo com suas quantidades de alunos, pois instituições com essa características possuem valores agregados de todas instituições pertencentes à mantenedora. Para ter um valor mais próximo das receitas e despesas da própria instituição utilizou-se a distribuição dessa quantia total entre as instituições de acordo com sua quantidade de alunos. A Tabela 1 apresenta um exemplo da distribuição de dados financeiros de instituições mantidas financeiramente pela mantenedora de código 135 no ano de 2019.

Tabela 1. Distribuição das receitas pela quantidade de alunos

Código IES	Valor receita informado	Quantidade de alunos	Porcentual qt. alunos	Valor receita após distribuição
190	R$ 94.702.654,67	1197	48,44%	R$ 45.875.790,22
191		175	7,08%	R$ 6.706.986,87
192		613	24,80%	R$ 23.493.616,88
193		381	15,41%	R$ 14.602.068,56
194		105	4,24%	R$ 4.024.192,12

Fatores podem influenciar se uma IES possui o valor financeiro maior ou menor em relação a outra IES. Diante do contexto brasileiro, as diferenças entre as 5 (cinco) regiões geográficas (Norte, Nordeste, Sudeste, Sul e Centro-Oeste) se tornam bastante significativas – assim como a categoria administrativa –, pois podem existir diferentes infraestruturas e investimentos públicos e privados entre as regiões, o que por consequência pode causar uma diferença entre a quantidade de despesa e receita em relação ao aluno. Por exemplo, as IES da região Sudeste possuem o valor médio das suas receitas maior em 86,65% em relação à região Norte. Dessa maneira, concluiu-se que o motivo de suas diferenças está igualmente ligado à sua infraestrutura que pode agregar mais pessoas, empresas, investimento e entre outros. A Figura 3 demonstra a influência da categoria administrativa.

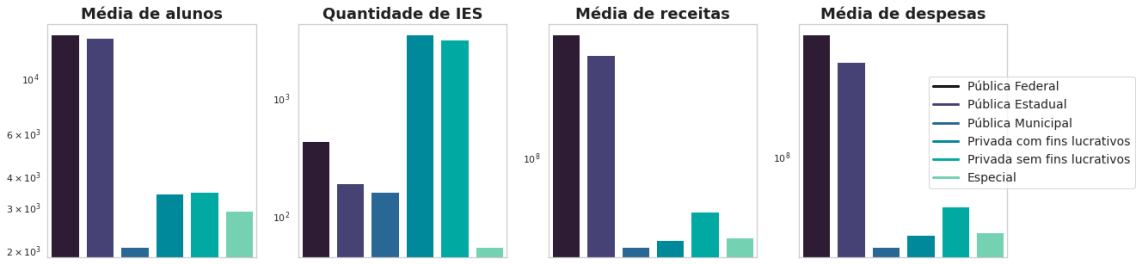

Figura 3. Média dos valores financeiros de IES mantidas por elas mesmas com quantidade de alunos em relação a categoria administrativa.

Fundamentado nisso, as categorias administrativas foram levadas em consideração para criação das amostras com IES com características financeiras semelhantes. Em relação às categorias administrativas privadas com e sem fins lucrativos, elas possuíam a quantidade de IES consideravelmente grande, o que foi necessário considerar a variável de região para criação de amostras menores. Com base nas distribuições das categorias administrativa houve inicialmente a suposição de junção de categorias administrativas em relação às variáveis financeiras, receitas e despesas, que poderia ter alguma semelhança em suas distribuições, como podemos ver na Figura 4, dessa forma foi realizado um teste de hipótese de Mann-Whitney que constatou que realmente poderia ser feita a junção das distribuições dos valores financeiros pelas categorias Especial e Municipal poderiam ser juntadas em uma única categoria. A partir da realização do teste, foi levado em consideração a transformação

logarítmica dos dados financeiros e da variável de quantidade de aluno que será utilizada na abordagem univariada e bivariada.

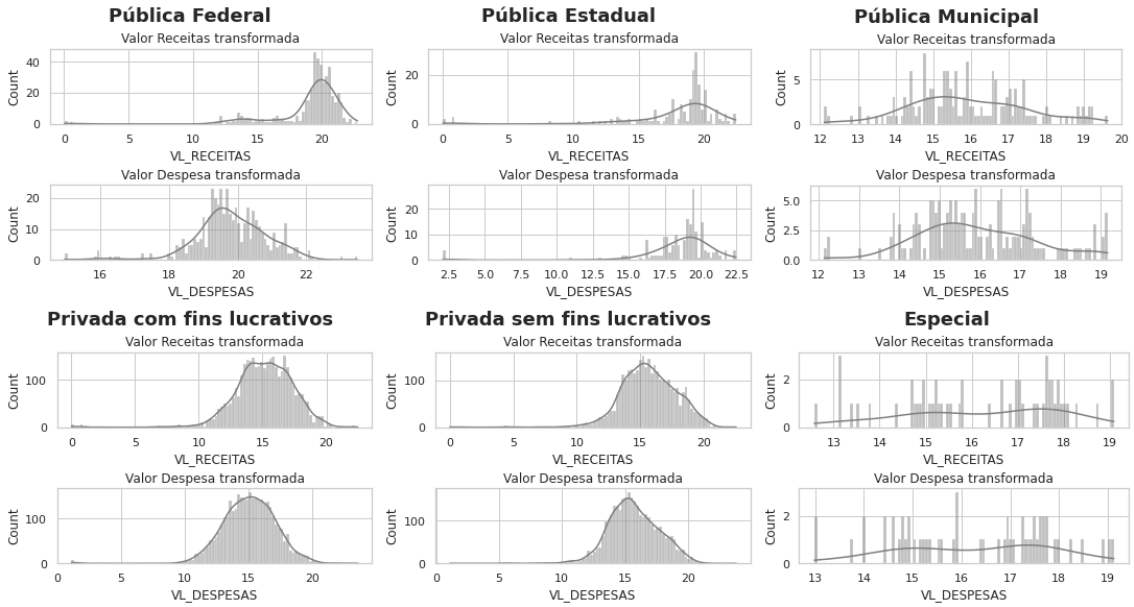

Figura 4. Distribuições de receitas e despesas transformadas por categoria administrativa.

Todas essas amostras foram submetidas a uma sub-amostragem por frequência por tercil pela quantidade de alunos de suas amostras na análise univariada de detecção de outliers. Quanto à abordagem bivariada, há instituições que possuem todos seus anos identificados como outliers, esse valor representa 48,28% das instâncias das receitas e 46,92% das instâncias das despesas do total identificado. Em relação ao algoritmo Mahalanobis em si, o seu método robusto com MCD consegue aumentar a percentagem de detecção, entretanto o número de falso positivo cresce junto, pois ele coloca como positivo às margens do centro médio das observações.

Na Tabela 2 temos o resultado da detecção que possui os seguintes valores: quantidades de instâncias *outliers* na univariada (conjunto U) e bivariada (conjunto B), instâncias detectadas apenas na abordagem univariada ($U - B$) e bivariada ($B - U$) e a interseção das abordagens ($U \cap B$). Após detecção de outliers por meio das duas abordagens (uni e bivariada), o presente trabalho considera as instâncias realmente discrepantes àquelas que foram identificadas em ambas as abordagens, ou seja, a interseção (U ∩ B) apresentada na Tabela 2. Dessa forma, tais instâncias representam os valores financeiros da IES de forma inconsistente, por isso precisam de tratamento para que reflitam de maneira mais próxima a realidade das IES.

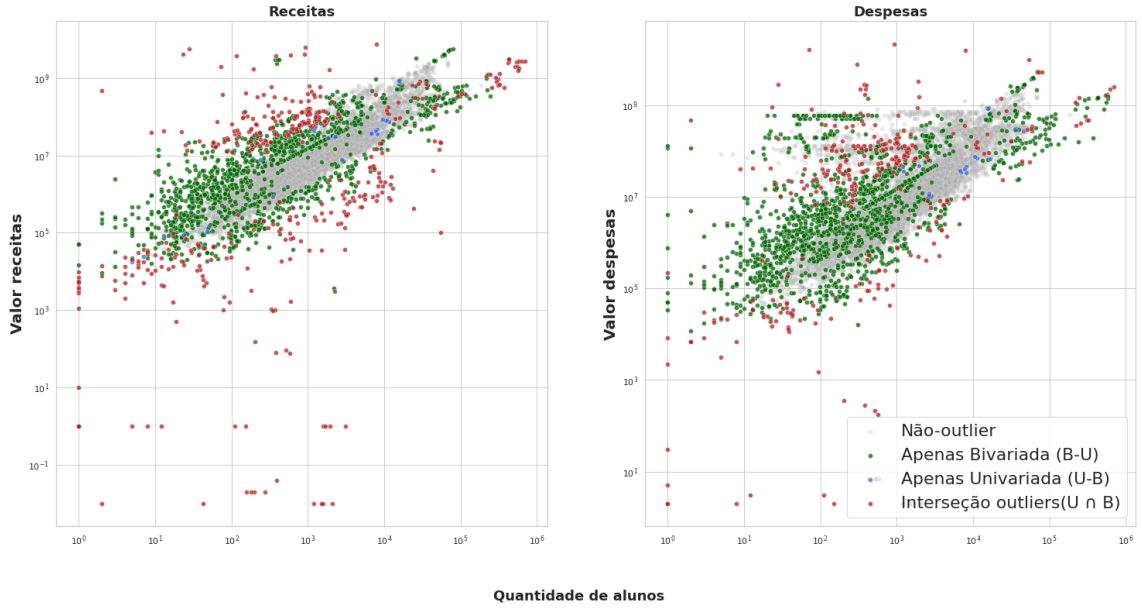

Figura 5. Outliers presentes nas abordagens nos valores financeiros

Tabela 2. Quantidade de instâncias e IES detectadas

Valor financeiro	U	B	$U - B$	$B - U$	$U \cap B$
Receitas	349 (176 IES)	1218 (582 IES)	23 (11 IES)	892 (471 IES)	326 (168 IES)
Despesas	268 (140 IES)	1118 (552 IES)	17 (9 IES)	867 (458 IES)	251 (132 IES)
Total (União)	419 (214 IES)	1414 (673 IES)	27 (15 IES)	1125 (581 IES)	393 (204 IES)

Ao total foram identificadas 393 instâncias que são referentes a 204 instituições (9,17% do total das IES) com os valores financeiros detectados como *outliers*, sendo 30 (14,71%) dessas instituições consideradas de categoria administrativa Pública ou Especial, as privadas com fins lucrativos com 101 (49,51%) e sem fins lucrativos com 73 (35,78%). As instâncias dessas IES foram aplicadas a seus respectivos tratamentos segundo o fluxograma da Figura 2. A Figura 6 representa a quantidade de instâncias em cada tipo de tratamento.

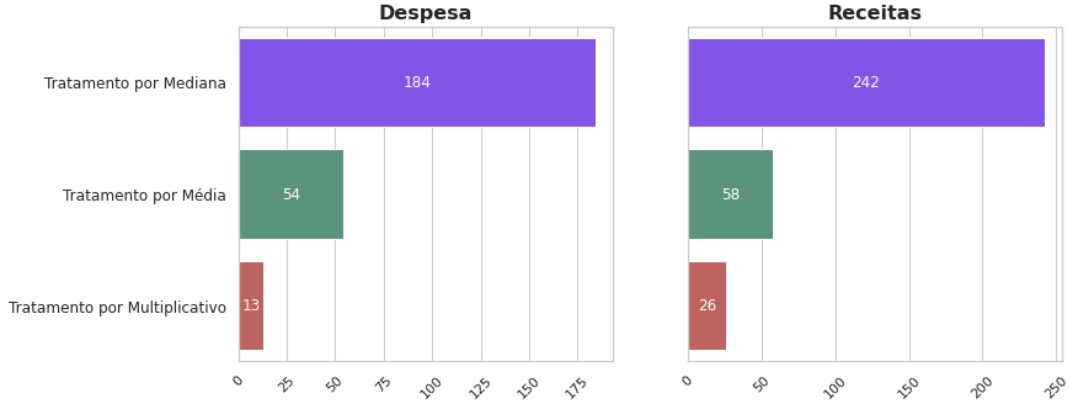

Figura 6. Quantidade de instâncias nos tipos de tratamento para as variáveis identificadas como outlier.

Outro ponto a se levar em consideração é que a abordagem conseguiu detectar muito bem os outliers inferiores. Para as instituições com valores inferiores a R$20.000,00 de despesa ou receita, o algoritmo conseguiu detectar que 95,12% das instâncias possuem em pelo menos um de suas variáveis financeiras um valor financeiro inconsistente. Valores acima desse valor a detecção se torna mais sutil, conseguindo identificar alguns casos. A Figura 7 demonstra a quantidade de instituições detectadas no intervalo de R$ 1.000,00 a R$100.000,00.

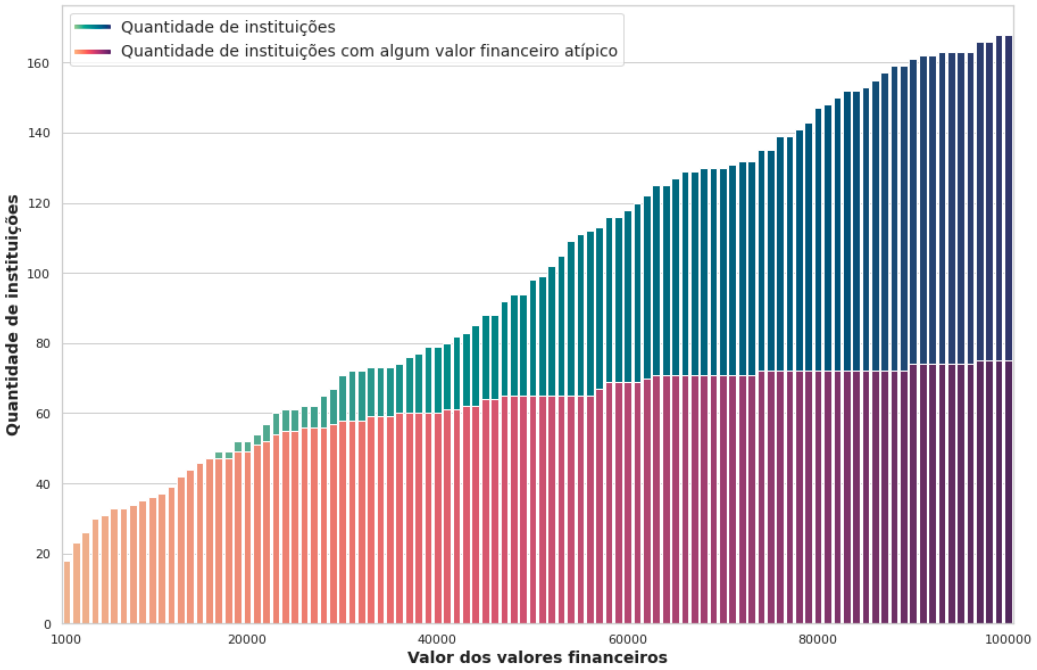

Figura 7. Quantidade total de com outliers em relação a quantidade total existente.

A abordagem de transformação logarítmica junto com abordagem de substituição do valor pela mediana, média ou multiplicativa se mostrou uma boa abordagem para que os valores financeiros pudessem ficar mais próximos da realidade de cada instituição em comparação a outras IES com características semelhantes. Na Figura 8 tem-se o resultado referente às receitas após a identificação e tratamento. Nota-se que o gráfico de caixa identifica outliers, o motivo disso é devido à distribuição não considerar as categorias administrativas e região e que certos outliers foram mantidos em vez de tratados a partir do pressuposto de existir falso positivo. Por exemplo, tem-se a instituição da UFRJ que foi detectada na abordagem bivariada, entretanto na univariada ela não foi detectada. O valor das despesas e receitas da UFRJ podem ser consideradas como outliers quando compara-se a sua quantidade de alunos, porém há fatores que possam influenciar os valores financeiros da instituição que não está presente no Censo, como a contratação de entidades sem fins lucrativos.

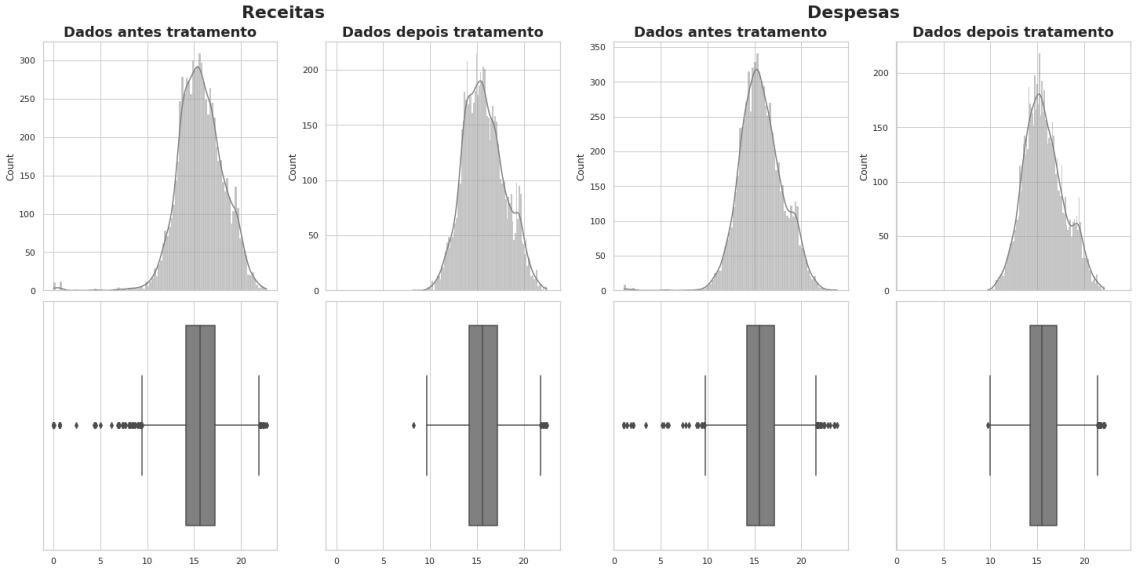

Figura 8. Distribuição gráfico de caixa antes e depois do tratamento em instituições mantidas financeiramente por ela mesma.

5. Considerações finais

As análises realizadas para os anos de 2016 a 2019 mostraram uma ocorrência significativa de valores atípicos nos dados financeiros reportados pelas IES. Foi observado que das 2.224 IES que reportaram valores neste período, 204 reportaram algum valor financeiro atípico, o que representa aproximadamente 9,2% das IES. Destas instituições que reportaram valores atípicos, aproximadamente 50% são instituições privadas e 36% são instituições sem fins lucrativos.

Considerando que estes valores são utilizados na definição de políticas educacionais e disponibilizados em portais de transparência, se faz necessária uma validação mais detalhada dos dados coletados, especialmente para estes grupos que apresentaram mais valores atípicos. Neste sentido, o método proposto demonstra a possibilidade de existir inconsistência na base de dados do Censo Superior que pode comprometer o desempenho de pesquisas que utilizam os dados mencionados, tornando-se necessário a compreensão de como esses dados foram preenchidos, principalmente em relação às instituições públicas.. Por outro lado, cria-se a possibilidade de aplicar o mesmo modelo de detecção e tratamento para aplicar diferentes atributos existentes dentro da mesma base, como por exemplo a quantidade de técnicos existentes na instituição.

Logo, é possível perceber que a realização de trabalhos com perspectivas semelhantes vai proporcionar melhor entendimento da natureza dos dados financeiros do censo superior. Este trabalho demonstra a possibilidade de existir instituições com valores financeiros discrepantes e uma abordagem de revelar e tratar dados financeiros inconsistentes. Entretanto é de conhecimento a presença de outras técnicas de detecção através de modelos de aprendizado de máquina, por exemplo, que podem ser realizadas em trabalhos futuros, junto com técnicas de tratamento dessas variáveis através de modelos de regressões.

Referências

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2022) Censo da Educação Superior. Acesso em 10 de outubro de 2022. Disponível em https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior

Fayyad, U. Piatetsky-Shapiro, G. e Smyth, P. (1996) From Data Mining to Knowledge Discovery in Databases. AI Magazine.

Silva, P. do Nascimento, R. L. S. Lima, M. Fagundes, R. & de Souza, F. D. F. (2019). Modelos de regressão aplicados a predição do desempenho escolar de estudantes do ensino fundamental. Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE), p. 1631.

Pinto, G. Freitas Júnior, O. & Costa, E. (2020). Mineração de Dados Educacionais: Um Modelo de Predição do Perfil do Aluno para Melhoria do IDEB. Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE), p. 1172.

Machado, R. G. (2022) Subsídio às Fiscalizações Públicas: Identificação dos Municípios com gastos discrepantes na Educação Básica, CADERNOS DE FINANÇAS PÚBLICAS. 22(01).

Carrano, D., Albergaria, E., Infante, C., & Rocha, L. (2019). Combinando Técnicas de Mineração de Dados para Melhorar a Detecção de Indicadores de Evasão Universitária. Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE), p. 1321.

de Siqueira Gê, E. A., & Borges, E. F. (2021). Identificação de outliers em processos de dispensas e inexigibilidades em licitações públicas: um estudo comparativo entre UFRN, IFRN e UFERSA nos anos de 2017 e 2018. Revista Inovar Contábil. 2(01).

Witten, I. H. Frank, Eibe; Hall, Mark A. (2011). Data mining: practical machine learning tools and techniques. 3rd ed. Burlington, MA:Elsevier/Morgan Kaufmann.

Hair, J. F, Anderson, R. E. Tatham, R. L., Black, W. C. (2009). Análise multivariada de dados. São Paulo: Bookman.

Tuley, JOHN W. (1977). Exploratory Data Analysis. Addison‑Wesley Publishing Company Reading, Massachusetts, 1st ed.