Análise de Sobrepreço em Itens de Licitações Públicas

Mariana O. Silva[1], Lucas L. Costa[1], Guilherme Bezerra[1], Larissa D. Gomide[1], Henrique R. Hott[1], Gabriel P. Oliveira[1], Michele A. Brandão[1,2], Anisio Lacerda[1], Gisele Pappa[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG, Brasil

[2]Instituto Federal de Minas Gerais (IFMG) – Ribeirão das Neves, MG, Brasil

mariana.santos@dcc.ufmg.br, lucas-lage@ufmg.br
{guilhermebezerra, larissa.gomide, henriquehott, gabrielpoliveira}@dcc.ufmg.br
michele.brandao@ifmg.edu.br, {anisio, glpappa}@dcc.ufmg.br

Abstract. *The analysis of overpricing in public bidding items can help government agencies identify indications of fraud in the acquisition of public goods or services. In this context, this article presents two main contributions: a methodology for treating and standardizing the description of bid items; and a statistical approach for detecting overpricing based on grouping performed with item descriptions. The results indicate that the proposed strategies are promising for identifying possible irregularities in public purchases.*

Resumo. *A análise de sobrepreços em itens de licitações públicas pode auxiliar órgãos governamentais a identificar indícios de fraude na aquisição de bens ou serviços públicos. Nesse cenário, este artigo apresenta duas contribuições principais, são elas: uma metodologia para tratamento e padronização da descrição de itens licitados; e uma abordagem estatística para detecção de sobrepreço baseada em agrupamentos realizados com a descrição dos itens. Os resultados indicam que as estratégias propostas são promissoras para a identificação de possíveis irregularidades nas compras públicas.*

1. Introdução

As licitações são procedimentos adotados pela administração pública para assegurar a imparcialidade e a escolha das melhores propostas na aquisição de bens ou serviços públicos. Durante o processo licitatório, o licitante estabelece um valor de referência que serve como base para determinar a modalidade de licitação e, dentre outras finalidades, averiguar se as propostas apresentadas são ou não vantajosas. No entanto, tal avaliação de preços pode ser um desafio, especialmente diante da diversidade de itens licitados e da variação nos preços de mercado.

Para solucionar tal problema, é comum que órgãos governamentais coletem e analisem dados históricos de licitações e os preços nelas praticados, a fim de identificar padrões e verificar se há indícios de sobrepreço ou irregularidades. Como exemplo, poderíamos citar o Sistema da Controladoria Geral da União (CGU)[1] e o Sistema do Tribunal de Contas do Estado de Minas Gerais (TCE/MG).[2] Entretanto, padronizar as entradas de texto e unidades de medidas dos itens comercializados, bem como a definição

[1] https://paineldeprecos.planejamento.gov.br/
[2] https://bancodepreco.tce.mg.gov.br/

de limiares para determinar se um preço está dentro dos conformes ou não, continuam sendo desafios a serem superados.

Nesse contexto, este trabalho tem como objetivo apresentar uma estratégia de detecção de sobrepreço em compras públicas de bens e serviços. Especificamente, os dados utilizados são tratados e padronizados para garantir a uniformidade das descrições dos itens licitados e facilitar a comparação entre os preços. Dessa forma, através de análises estatísticas, são definidos critérios objetivos para determinar se um preço está dentro dos padrões aceitáveis ou se há indícios de sobrepreço. Com a abordagem proposta, também são identificadas possíveis anomalias, que podem ser provenientes de erros de digitação ou indícios de fraudes nas licitações.

O trabalho está organizado da seguinte maneira. Na Seção 2, são apresentados trabalhos relacionados que abordam problemas semelhantes de padronização de texto e determinação de sobrepreço. Nas Seções 3 e 4, descrevemos o conjunto de dados utilizado e detalhamos a metodologia adotada para tratá-los e padronizá-los. Na Seção 5, apresentamos a abordagem utilizada para identificação de sobrepreço, os resultados obtidos e analisamos exemplos de sobrepreço e anomalia para ilustrar a aplicação prática da metodologia proposta e suas limitações. Finalmente, na Seção 6, concluímos o trabalho e apresentamos possíveis trabalhos futuros para aprimorar nossa abordagem.

2. Trabalhos Relacionados

O objetivo deste trabalho é a detecção de sobrepreço em compras públicas e, por conseguinte, a identificação de indícios de fraudes. Nesse contexto, os documentos de licitação apresentam uma estrutura definida, porém utilizam termos e formas distintas para se referir a um mesmo item, o que dificulta o processamento automático desses documentos [Silva et al. 2022, Oliveira et al. 2022]. Para solucionar esse problema, técnicas especializadas em Processamento de Linguagem Natural (PLN) são utilizadas para processar os dados textuais. Na literatura, duas frentes principais são exploradas: detecção de fraudes e processamento textual de documentos públicos.

Em relação à detecção de fraudes, [Oliveira et al. 2022] propõem uma abordagem de decisão hierárquica que envolve o pré-processamento dos dados para aumentar a estruturação e padronização textual. Os dados são então organizados em três categorias, baseado em taxas de adesão e ocorrência, permitindo a classificação dos itens comprados de acordo com a validade da transação. Já [Pereira et al. 2022] apresentam uma modelagem baseada em grafos e medidas de centralidade, seguida por algoritmos de classificação para rotular as empresas como fraudulentas ou não. Em [Luna and Figueiredo 2022], é utilizada uma abordagem similar, mas em vez de rodar um algoritmo de classificação, são calculadas métricas para observar possíveis indícios de fraudes. Por fim, [Gabardo and Lopes 2014] apresentam uma estratégia de identificar cartéis em empresas de construção civil através de dados de redes sociais.

Considerando o aspecto de pré-processamento de texto e organização de dados textuais, podemos citar os trabalhos de [Pereira et al. 2021] e [Constantino et al. 2022]. No primeiro, foi realizado um estudo sobre a diversidade de termos usados para se referir a um mesmo serviço em sites governamentais. Para resolver esse problema, os autores coletaram dados e organizaram os termos em uma taxonomia, permitindo uma melhor padronização e estruturação dos dados. Já em [Constantino et al. 2022], é apresentado

um processo que envolve a segmentação de termos após a coleta de documentos, seguida por uma arquitetura de aprendizado ativo e, por fim, uma etapa de classificação semântica, aprimorando a eficácia da análise de dados textuais.

A revisão da literatura revelou a importância de detectar possíveis indícios de fraudes em compras públicas, mas notou-se a escassez de estudos que abordam o desafio da padronização das entradas. Além disso, a falta de parâmetros estatísticos comparativos entre itens similares não foi explorada nos trabalhos analisados. Portanto, a principal contribuição deste trabalho é justamente apresentar uma abordagem que lida com esses desafios. É importante ressaltar que, embora em [Oliveira et al. 2022] seja realizado um pré-processamento nos dados textuais de licitações, esse procedimento é mais simples do que a nossa metodologia proposta, uma vez que não inclui a categorização de itens similares conforme discorrido nas Seções 4.2 e 4.3.

3. Conjunto de Dados

Esta seção apresenta o conjunto de dados considerado na metodologia de análise de sobrepreço em itens de licitações públicas. Em primeiro lugar, a Seção 3.1 apresenta uma breve descrição de informações relevantes sobre os dados, incluindo as principais fontes, período e escopo de abrangência. Em seguida, a Seção 3.2 contém uma análise exploratória inicial sobre os itens licitados, com foco nos atributos considerados nas análises deste trabalho, como a descrição do item e sua natureza.

3.1. Descrição dos dados

O conjunto de dados considerado neste trabalho contém informações relativas a itens de licitações públicas ocorridas nos 853 municípios do Estado de Minas Gerais. O termo *item* se refere tanto a bens (e.g., veículos) quanto a serviços (e.g., desenvolvimento de *sites*). Além dos itens de licitação, também são considerados itens dispensados de licitação, de acordo com a Lei Federal nº 14.133, de 1º de abril de 2021.[3] Um exemplo de item para o qual é dispensada a licitação é a aquisição de medicamentos destinados ao tratamento de doenças raras definidas pelo Ministério da Saúde (Art. 75).

As informações disponibilizadas são públicas e são provenientes de duas fontes principais. Para licitações municipais, os dados vêm do Sistema Informatizado de Contas dos Municípios (SICOM),[4] desenvolvido pelo Tribunal de Contas do Estado de Minas Gerais (TCE-MG) com informações dos portais da transparência de todos os municípios mineiros. Já para licitações estaduais, são considerados dados retirados do Portal da Transparência do Governo do Estado.[5] Apesar de os dados virem de duas fontes consolidadas, é necessária uma etapa adicional de junção e agregação das informações para obter os dados necessárias para o cálculo e análise de sobrepreço. O conjunto final de dados apresenta as seguintes informações para cada item licitado/dispensado: identificador da licitação/dispensa, ano exercício, descrição textual, unidade de medida, valor unitário homologado e natureza da despesa.

[3]Lei nº 14.133/21: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm

[4]SICOM: https://portalsicom1.tce.mg.gov.br/

[5]Portal da Transparência do Estado de Minas Gerais: https://www.transparencia.mg.gov.br/compras-e-patrimonio/compras-e-contratos

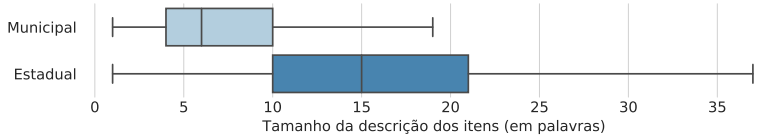

Figura 1. Distribuição do tamanho das descrições dos itens em palavras.

Figura 2. Nuvens de termos mais frequentes nos itens de licitação/dispensa municipais (esquerda) e estaduais (direita).

3.2. Análise exploratória

Esta seção apresenta uma breve caracterização e análise exploratória do conjunto de dados de itens licitados descrito na seção anterior. Tais análises são importantes para entender melhor os dados trabalhados e também para auxiliar na definição da metodologia para detecção de sobrepreço nos processos licitatórios. Nesse sentido, as análises aqui realizadas visam obter informações como a quantidade de itens licitados por ano, as principais naturezas das despesas, os principais termos que aparecem na descrição dos items e também a distribuição do tamanho dessas descrições.

O conjunto de dados considerado neste trabalho possui um total de 12.805.984 itens licitados/dispensados no âmbito municipal e 1.361.523 itens no âmbito estadual. Tais itens estão em processos licitatórios que ocorreram no período de 2014 a 2021 para os municípios e 2009 a 2021 para o estado. A natureza de despesa mais frequente dos itens de licitação/dispensa é "material de consumo", representando 41,7% dos itens municipais e 85,3% dos estaduais. Apesar de pequenas diferenças no nome da categoria, outros tipos de natureza incluem "material permanente" e "serviços".

Ao analisar as descrições dos itens, observam-se algumas diferenças ao comparar os itens municipais e estaduais. Considerando o tamanho da descrição em número de palavras (Figura 1), é possível notar uma diferença significativa entre os dois conjuntos. De forma geral, os itens municipais possuem descrições mais curtas que os itens estaduais (mediana de 6 palavras para municipais e 15 para estaduais). Além disso, ao analisar as nuvens de palavras para os dois conjuntos (Figura 2), percebe-se uma diferença nos termos mais frequentes. Enquanto para os itens municipais os termos mais frequentes envolvem materiais de consumo em geral (e.g., "caixa", "papel", "folha", "plástico"), os itens estaduais parecem estar mais ligados à área da saúde, contendo termos como "matéria prima", "dosagem", "farmacêutica", "composição", etc.

No entanto, existem termos frequentes nas descrições de ambos os conjuntos que não agregam informações às análises. Em geral, tais termos são unidades de medida, como "unidade" e "grande". Além disso, alguns termos aparecem escritos de duas ou

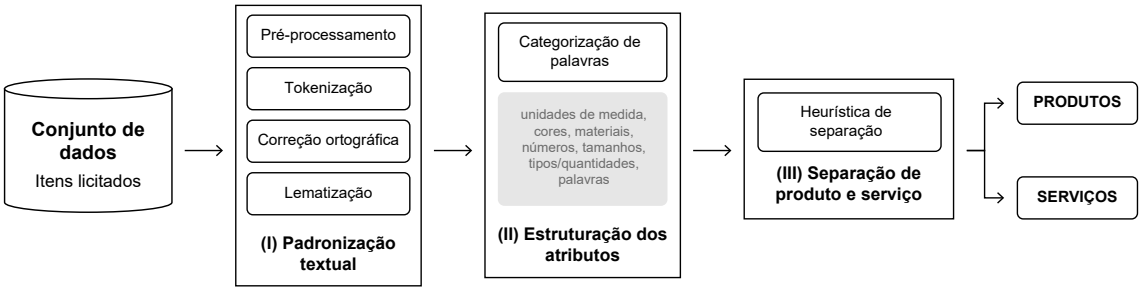

Figura 3. Visão geral da metodologia de tratamento e padronização.

mais formas diferentes, apesar de significarem o mesmo. A partir de tais observações, é necessário aplicar um pré-processamento textual nas descrições dos itens para aumentar a padronização no conjunto de itens.

4. Tratamento e Padronização

Considerando os problemas observados propõe-se, a fim de para padronizar a descrição dos itens licitados, uma metodologia de tratamento e padronização textual que compreende três etapas: (i) padronização textual, (ii) estruturação dos atributos e (iii) separação de produtos e serviços. A Figura 3 ilustra a metodologia aplicada e cada uma das etapas será detalhada e exemplificada a seguir.

4.1. Padronização textual

Na etapa de padronização textual, a descrição dos itens passam por quatro operações: (i) pré-processamento, que envolve a remoção de caracteres especiais, pontuações e *stopwords* (i.e., palavras insignificantes); (ii) tokenização, que divide o texto em *tokens*, ou seja, as palavras que compõem a descrição; (iii) correção ortográfica, que verifica a grafia das palavras e substitui palavras mal escritas ou com erros de digitação por correções; e (iv) lematização, que reduz as palavras à forma básica ou lema, eliminando flexões e variações de gênero, número e tempo. Essas operações são fundamentais para garantir a uniformidade na descrição dos produtos ou serviços licitados, permitindo uma comparação mais precisa dos preços.

Para exemplificar a etapa de padronização textual, considere a descrição de item “OLEO 20W40 MOTOR A GASOLINA 500ML”. Após a aplicação das quatro operações, a saída resultante é uma lista dos *tokens* pré-processados: “oleo”, “20”, “w”, “40”, “motor”, “gasolina”, “500” e “mililitro”. Observe que duas grandes transformações foram realizadas nessa descrição: (i) a palavra “a” presente na descrição original foi removida, por ser considerada *stopword*, e (ii) a palavra “ml” foi substituída por “mililitro” na correção ortográfica.

4.2. Estruturação dos atributos

Após a etapa de padronização textual, a estruturação de atributos é fundamental para a correta identificação das características dos itens descritos nas licitações. Nessa etapa, para cada *token* presente na descrição de um item, são atribuídas categorias de palavras específicas. Aqui, são consideradas sete categorias: *Unidades de medida*, *Cores*, *Materiais*, *Números*, *Tamanhos*, *Tipos/Quantidades* e *Palavras*. A partir dessa estruturação de

Algoritmo 1: Heurística de separação entre produtos e serviços.

Entrada: descrição do item d_i, natureza de despesa nd_i, unidade de medida um_i, *natureza_despesa*, *unidade_medida*, *unigramas* e *bigramas*
Saída: classificação do item como **produto** ou **serviço**

```
início
    se nd_i ∈ natureza_despesa então
        retorna serviço
    se um_i ∈ unidade_medida então
        retorna serviço
    para cada Palavra p em d_i faça
        se p ∈ unigramas então
            retorna serviço
    para cada bigrama b em bigramas faça
        se b ∈ d_i então
            retorna serviço
    retorna produto
```

atributos, é possível obter uma descrição mais precisa dos itens licitados e, consequentemente, realizar uma comparação mais justa dos preços.

A categoria *Unidades de medida* inclui termos como "litro", "grama", "quilograma", entre outros. Já a categoria *Cores* abrange termos relacionados a cores, como "azul", "vermelho", "claro" e "escuro". A categoria *Materiais* engloba termos que descrevem os materiais utilizados na fabricação do item, tais como "aço", "metal", "madeira" e "plástico". A categoria *Números*, por sua vez, inclui todos os termos numéricos presentes na descrição dos itens, enquanto a categoria *Tamanhos* abrange termos que descrevem as variações de tamanho, como "pequeno", "grande", "único", entre outros. A categoria *Tipos/Quantidades* descreve a forma de apresentação do item, bem como sua quantidade, incluindo "pacote", "comprimido" e "unidade". Por fim, a categoria *Palavras* inclui todos os termos que não se enquadram em nenhuma das categorias anteriores.

Como exemplo, considerando a mesma descrição da etapa anterior (i.e., "OLEO 20W40 MOTOR A GASOLINA 500ML"), três categorias foram identificadas: *Unidades de medida*, incluindo os *tokens* "w" e "mililitro"; *Números*, incluindo os *tokens* "20", "40" e "500"; e *Palavras*, incluindo os *tokens* "oleo", "motor" e "gasolina".

4.3. Separação de produto e serviço

Após a padronização textual e a estruturação dos atributos, a próxima etapa da metodologia é a separação dos itens licitados entre produto e serviço. Embora essa informação deveria estar indicada nos dados de entrada, isso nem sempre acontece. Essa separação é fundamental, pois o agrupamento sem distinguir produtos de serviços pode levar a estimativas desbalanceadas de preços. Portanto, foi desenvolvida uma heurística baseada em palavras-chave para realizar essa classificação automaticamente. O Algoritmo 1 descreve a heurística proposta, que recebe como entrada a descrição do item, seus metadados e quatro listas pré-definidas: *natureza_despesa*, *unidade_medida*, *unigramas* e *bigramas*.[6]

As quatro listas foram geradas a partir de descrições e palavras-chave frequentemente utilizadas para descrever itens de serviços. A Tabela 1 apresenta alguns exemplos

[6]O conteúdo completo das quatro listas está disponibilizado em `bit.ly/descricao_listas`.

Tabela 1. Exemplos de descrições e palavras-chave de cada lista pré-definida.

natureza_despesa	locacao de mao, outros servicos de terceiros, servicos de consultoria, obras e instalacoes, servicos, obras
unidade_medida	construcao, diaria, fornecimento, locacao, manutencao, obra, prestacao_servico, procedimento, servico, show, transporte
unigramas	servico, servicos, prestacao, locacao, contratacao, manutencao, construcao, instalacao, consultoria, acessoria, fornecimento
bigramas	transporte escolar, mao obra, show artistico, show musical

presentes em cada um das listas. A lista *natureza_despesa* contém descrições de natureza de despesa associadas a serviços, enquanto as outras listas incluem palavras-chave relacionadas às categorias de *Unidades de medida* e *Palavras*. Assim, para cada item, a heurística verifica se a natureza de despesa do item está contida na lista *natureza_despesa*, ou se sua unidade de medida está contida na lista *unidade_medida*, ou se algum *token* da categoria *Palavras* está contido nas listas *unigramas* e *bigramas*.

Caso alguma dessas condições seja verdadeira, o item é classificado como **serviço**; caso contrário, ele é classificado como **produto**. Para exemplificar a aplicação da heurística, considere o mesmo exemplo apresentado nas etapas anteriores, i.e., "OLEO 20W40 MOTOR A GASOLINA 500ML". A natureza de despesa desse item é desconhecida e a unidade de medida é *unidade*, descartando as condições das linhas 2 e 4. Além disso, nenhum *token* da descrição está contido nas listas *unigramas* e *bigramas*. Portanto, tal item é classificado como **produto**.

Após a aplicação de toda a metodologia de tratamento e padronização, 368.644 (2,88%) itens foram desconsiderados por conter informações ausentes ou descrições sem sentido. Os 12.437.340 itens restantes foram, então, processados e classificados entre produto e serviço. No total, 10.558.948 (84,9%) itens foram classificados como **produto** e 1.878.392 (15,1%) como **serviço**.

5. Identificação de Sobrepreço

A identificação de sobrepreço em compras públicas de bens e serviços pode ser modelada como um problema de detecção de *outliers*, ou seja, identificação de valores que se desviam significativamente do comportamento padrão do conjunto de dados. Nesse contexto, é comum utilizar técnicas estatísticas para detecção de *outliers*, tais como análise de regressão, análise de variância ou métodos baseados em distância. Aqui, propomos uma abordagem estatística baseada na amplitude interquartil, detalhada na Seção 5.1. Além disso, apresentamos os resultados experimentais obtidos na Seção 5.2, bem como exemplos de sobrepreço e anomalia para ilustrar a aplicação prática da metodologia proposta e suas limitações, como descrito na Seção 5.3.

5.1. Abordagem estatística

A amplitude interquartil (IQR) é uma medida robusta que captura a dispersão dos dados em relação à sua mediana. Para isso, é calculada a diferença entre o terceiro e o primeiro quartil do conjunto de dados, resultando em uma faixa que contém a maioria dos valores centrais. Qualquer valor que esteja fora dessa faixa é considerado um possível *outlier*. Para definir um limite superior para o preço de um determinado produto ou serviço, multiplicamos a IQR por um fator de escala igual a 1,5. Além do limite superior, também

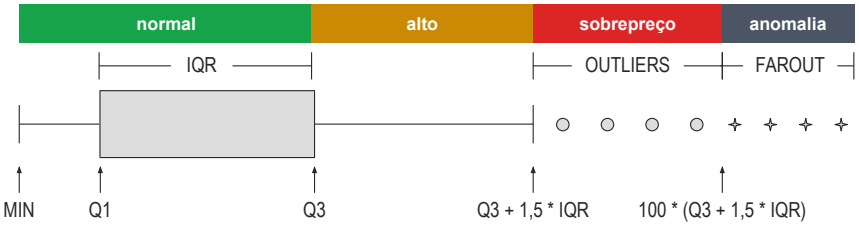

Figura 4. Visão geral da abordagem estatística.

definimos um limite máximo para separar os *outliers* de anomalias, multiplicando o limite superior por um fator de 100.

Conforme mostrado na Figura 4, foram definidos quatro níveis de preço dos itens: *normal*, *alto*, *sobrepreço* e *anomalia*. Esses níveis foram definidos com base em análises empíricas dos dados e na literatura sobre detecção de *outliers*. O nível *normal* se refere aos preços que estão abaixo e dentro da faixa IQR. O nível *alto* se refere aos preços que estão acima da faixa IQR, mas ainda dentro do limite superior. O nível *sobrepreço* se refere aos preços que ultrapassam o limite superior, mas ainda abaixo do limite máximo definido por uma margem de tolerância. Por fim, o nível *anomalia* se refere aos preços que ultrapassam o limite máximo.

É importante ressaltar que essa abordagem não deve ser encarada como uma ferramenta final para detectar sobrepreço, mas sim como um meio para garantir a eficiência e a transparência dos gastos públicos. Ou seja, a partir da classificação de preços suspeitos (*sobrepeço* e *anomalia*), é preciso realizar uma análise mais detalhada para verificar se esses preços são justificáveis ou se indicam a ocorrência de alguma irregularidade.

5.2. Resultados experimentais

Para aplicar a abordagem estatística proposta a um determinado produto ou serviço, primeiramente, é necessário agrupar o conjunto de dados pela descrição do item. Nesse caso, diferentes formas de agrupamento podem ser realizadas, incluindo agrupar pela descrição original, pela descrição tratada e estruturada ou agrupar por uma ou mais categorias resultantes da etapa de estruturação de atributos. Vale ressaltar que a escolha da forma de agrupamento tem um impacto direto na qualidade e confiabilidade dos resultados obtidos. Um agrupamento inadequado pode gerar estimativas de preço menos confiáveis e, consequentemente, comprometer a detecção de sobrepreço ou indícios de fraude.

Portanto, três experimentos foram realizados para analisar a abordagem estatística proposta e a metodologia de tratamento e padronização, alternando a forma de agrupar os produtos e serviços. Para se ter uma referência de comparação, a primeira forma de agrupamento considera a descrição original dos itens, sem qualquer tipo de tratamento ou padronização de texto. Já a segunda forma de agrupamento utiliza todos os termos presentes na categoria *Palavras*, enquanto a terceira utiliza apenas o primeiro termo dessa categoria. Em todas as três formas, foram considerados também a comarca,[7] o ano e o mês da licitação, visto que a localidade e o período podem influenciar significativamente os preços dos produtos e serviços.

[7]Comarca é um termo que caracteriza a divisão de uma região onde existem fronteiras, ou seja, onde as divisões territoriais são de responsabilidade de um ou mais juízes de direito.

Tabela 2. Comparação dos resultados de cada agrupamento.

	# grupos		# grupos suspeitos		% sobrepreço		% anomalia	
	produto	**serviço**	**produto**	**serviço**	**produto**	**serviço**	**produto**	**serviço**
Descrição original	4.244.671	1.179.081	2.600	548	0,072%	0,036%	0,000%	0,001%
Palavras	2.069.335	796.475	20.003	4.857	0,579%	0,719%	0,001%	0,004%
Primeiro termo	3.365	16.710	2.274	2.011	4,171%	3,544%	0,019%	0,189%

Com base nos resultados apresentados na Tabela 2, pode-se observar que a estratégia de agrupamento baseada na descrição original dos itens apresenta uma maior quantidade de grupos, tanto para produtos quanto para serviços. Esse resultado era esperado visto que essa abordagem representa uma descrição mais detalhada e com maior ruído, onde diferentes descrições podem definir o mesmo item. Portanto, em comparação com as demais formas de agrupamento, tal abordagem resulta em grupos mais específicos e menores, e, consequentemente, apresenta uma menor taxa de sobrepreço e anomalia. Em relação à abordagem de agrupamento por *Palavras*, os resultados indicam uma quantidade intermediária de grupos. Já a forma de agrupamento por apenas o primeiro termo dessa categoria apresentou a menor quantidade de grupos, sugerindo uma perda significativa de informações e uma tendência a agrupar itens distintos em um mesmo grupo.

Apesar da estratégia *Palavras* ter gerado um número intermediário de grupos, em relação aos casos suspeitos, o número de grupos foi consideravelmente maior do que nas outras duas abordagens. Isso indica que a estratégia de agrupamento por *Palavras* pode gerar grupos mais representativos e, como resultado, é capaz de identificar mais casos suspeitos de sobrepreço e anomalias. Mesmo com uma redução na quantidade de grupos, em comparação com a abordagem baseada na descrição original dos itens, a estratégia *Palavras* se mostrou mais eficiente na detecção de casos suspeitos, apresentando uma taxa de sobrepreço e anomalias consideravelmente maior.

De fato, os resultados indicam que a abordagem de agrupamento por *Palavras* apresenta uma taxa de anomalia e sobrepreço maior do que a estratégia baseada na descrição original, mas menor do que a abordagem que considera apenas o primeiro termo dessa categoria. Isso sugere que a utilização de todas as palavras na categoria *Palavras* pode ajudar a refinar a descrição dos itens e a reduzir a presença de anomalias e sobrepreços em comparação com a abordagem que utiliza apenas o primeiro termo. No entanto, ainda assim, a perda de informações é menor do que a abordagem que utiliza apenas o primeiro termo, o que a torna uma opção intermediária viável em situações onde a descrição original é muito ruidosa ou pouco detalhada.

Após a análise dos diferentes agrupamentos, também foi avaliada a distribuição de preços em cada um deles para identificar possíveis diferenças significativas. A Figura 5 apresenta a distribuição de preços (A-B) produtos e (C-D) serviços nos níveis de sobrepreço e anomalia, respectivamente. Os resultados indicam que, em relação ao nível de sobrepreço, há diferenças significativas nas distribuições de preços dos diferentes agrupamentos, independentemente da natureza do item. Notavelmente, a abordagem *Palavras* apresenta preços menores, em média, em comparação com as outras formas de agrupamento. Tal resultado pode ser justificado pelo fato de que a abordagem pode levar a perda de informação e, consequentemente, à redução da variância dos dados.

Em relação ao nível de anomalia, a distribuição de preços é semelhante para a mai-

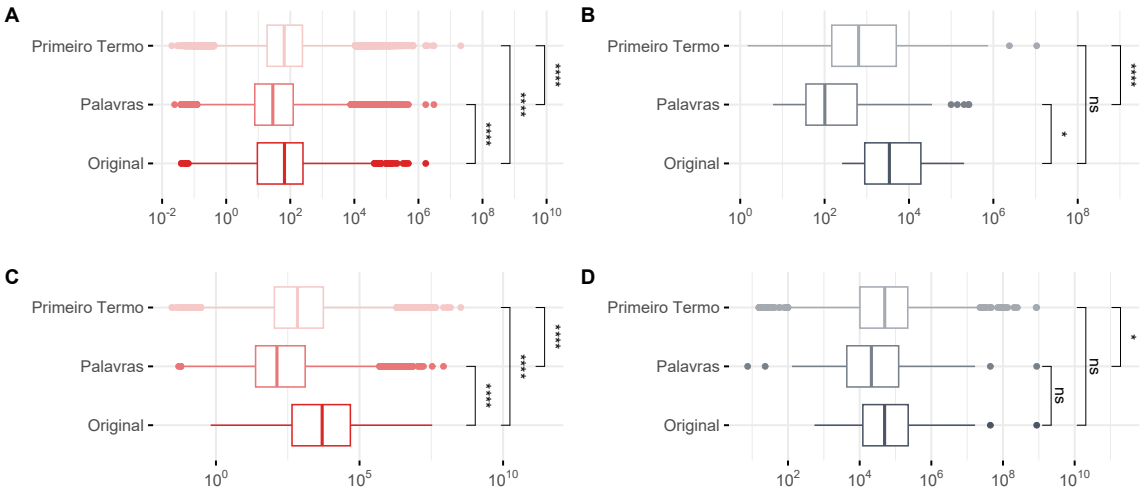

Figura 5. Distribuição de preços para (A-B) produtos e (C-D) serviços nos níveis de sobrepreço e anomalia, respectivamente. Os níveis de *p-value* são simbolizados como (1) ns e *: $p > 0,01$, (2) ***: $p < 0,001$, (3) ****: $p < 0,0001$.

oria dos casos, indicando que a presença de anomalias não é influenciada pela forma de agrupamento utilizada. No entanto, é importante destacar que, mesmo na presença de anomalias, a abordagem estatística proposta e a metodologia de tratamento e padronização são capazes de identificar e tratar esses valores discrepantes de maneira adequada, garantindo a confiabilidade dos resultados obtidos. Portanto, no geral, a análise dos resultados indica que a abordagem estatística proposta e a metodologia de tratamento e padronização são eficazes em identificar possíveis sobrepreços e anomalias em itens de produtos e serviços. Além disso, os resultados sugerem que a abordagem de agrupamento por *Palavras* pode ser uma opção viável para a detecção de possíveis irregularidades em compras públicas, pois apresenta uma boa relação entre quantidade de grupos e capacidade de identificar casos suspeitos.

5.3. Exemplos de sobrepreço e anomalia

Para cada estratégia de agrupamento, foram exemplificados os top três produtos e serviços classificados como sobrepreço e anomalia, ou seja, os três itens com maior discrepância entre o preço registrado e o limiar de sobrepreço/anomalia (conforme explicado na Seção 5.1). As Tabelas 3 e 4 apresentam esses resultados detalhadamente. Observa-se que os itens classificados como sobrepreço apresentam valores muito superiores ao limiar, identificando possíveis irregularidades nos preços praticados. No entanto, é importante ressaltar que os resultados de agrupamento podem variar entre as abordagens, o que ressalta a necessidade de uma avaliação cuidados desses resultados.

Já para os itens classificados como anomalias, é possível notar uma diferença de magnitude significativa entre o preço registrado e o limiar de anomalia em todos os casos. Isso sugera a presença de possíveis erros de digitação na descrição ou na precificação do produto/serviço. Além disso, é comum encontrar descrições pouco claras, como em "EVENTOS" ou "CONTRATACAO DE EMPRESA", ou escritas de maneira não usual, como o caso de "LIVRO OBRA LITERARIA PARA COMPOSICAO DE KIT ESCOLAR 3...", o que pode estar relacionado a problemas no momento do registro do item. Esses resultados ressaltam a importância do tratamento e padronização da descrição de

Tabela 3. Top três produtos e serviços classificados como sobrepreço. Os termos utilizados nas estratégias *P* e *PT* foram sublinhados.

	Produto			Serviço		
	Descrição original	**Preço (R$)**	**Limiar (R$)**	**Descrição original**	**Preço (R$)**	**Limiar (R$)**
DO	TECIDO TIPO TNT VERDE	485.000,0	59.780,4	PRESTACAO DE SERVICOS DE RECAPEAMENTO ASFALTICO...	871.685.340,0	8.010.707,4
	EQUIPAMENTOS HOSPITALARES	392.000,0	41.600,0	EVENTOS	44.348.000,0	217.000,0
	OLEO DIESEL S10 4892018 447337	1.688.830,8	1.378.620,0	33903923 OUTROS SERVICOS DE TERCEIROS PESSOA JURIDICA...	34.000.000,0	10.500.000,0
P	PECAS	3.036.463,9	199.000,0	PRESTACAO DE SERVICOS DE RECAPEAMENTO ASFALTICO...	871.685.340,0	8.010.707,4
	TECIDO TIPO TNT VERDE	485.000,0	59.780,4	EVENTOS	44.348.000,0	217.000,0
	EQUIPAMENTOS HOSPITALARES	392.000,0	41.600,0	33903923 OUTROS SERVICOS DE TERCEIROS PESSOA JURIDICA...	34.000.000,0	10.500.000,0
PT	LIVRO OBRA LITERARIA PARA COMPOSICAO DE KIT ESCOLAR 3...	202.696,3	49,6	PRESTACAO DE SERVICOS DE RECAPEAMENTO ASFALTICO...	871.685.340,0	1.514.915,3
	CONCESSAO DE DIREITO REAL DE USO REFERENTE 01 (UM) TERRENO...	21.000.000,0	2.250.000,0	CONTRATACAO DE EMPRESA PARA CONTINUACAO DA OBRA...	841.365.170,0	32.224,0
	EXAMES LABORATORIAIS ITEM GENERICO-378218	10.643.942,0	152,5	CONTRATACAO DE EMPRESA	336.330.975,6	4.591.747,5

DO = Descrição Original *P* = Palavras *PT* = Primeiro termo

Tabela 4. Top três produtos e serviços classificados como anomalia. Os termos utilizados nas estratégias *P* e *PT* foram sublinhados.

	Produto			Serviço		
	Descrição original	**Preço (R$)**	**Limiar (R$)**	**Descrição original**	**Preço (R$)**	**Limiar (R$)**
DO	LIVRO OBRA LITERARIA PARA COMPOSICAO DE KIT ESCOLAR 3...	202.696,3	4.957,5	PRESTACAO DE SERVICOS DE RECAPEAMENTO ASFALTICO...	871.685.340,0	801.070.738,5
	LIVRO OBRA LITERARIA PARA COMPOSICAO DE KIT ESCOLAR 3 A 5...	8.722,0	2.667,0	EVENTOS	44.348.000,0	21.700.001,5
	BRONZE DO PARAFUSO DA COROA	1.349,8	944,5	CONTRATACAO DE EMPRESA	16.626.939,6	100,0
P	LIVRO OBRA LITERARIA PARA COMPOSICAO DE KIT ESCOLAR 2...	263.712,1	3.577,5	PRESTACAO DE SERVICOS DE RECAPEAMENTO ASFALTICO...	871.685.340,0	801.070.738,5
	LIVRO OBRA LITERARIA PARA COMPOSICAO DE KIT ESCOLAR 3...	256.230,4	3.577,5	EVENTOS	44.348.000,0	21.700.001,5
	GASOLINA COMUM	99.500,0	396,0	CONTRATACAO DE EMPRESA	16.626.939,6	100,0
PT	EXAMES LABORATORIAIS ITEM GENERICO-378218	10.643.942,0	15.250,0	CONTRATACAO DE EMPRESA PARA CONTINUACAO DA OBRA...	841.365.170,0	3.222.400,0
	EIXO VIRABREQUIM 366	2.398.025,0	94.868,5	PRESTACAO DE SERVICOS DE RECAPEAMENTO ASFALTICO...	87.1685.340,0	151.491.532,5
	TUBO ACELERADOR DE PARTICULAS, REF. 11303910, PARA ACELERADOR...	747.957,6	8.302,5	PRESTACAO DE SERVICO - IMPLANTACAO DA PLATAFORMA...	214.692.000,0	20.050,0

DO = Descrição Original *P* = Palavras *PT* = Primeiro termo

itens licitados e a necessidade de avaliar criteriosamente as anomalias identificadas.

Ameaças à validade. Abordagens que lidam com processamento de texto enfrentam muitos desafios devido à falta de padronização, o que pode afetar a qualidade dos resultados. Uma ameaça identificada na heurística proposta é que, em alguns casos, a discrepância pode ser derivada de problemas na heurística de separação. Por exemplo, na análise do item "EXAMES LABORATORIAIS ITEM GENERICO", a descrição indica claramente que se trata de um serviço laboratorial, mas a heurística de separação o classificou de forma incorreta. Isso pode levar a uma estimativa imprecisa do limiar de anomalia e, consequentemente, gerar um falso positivo. No entanto, é importante lembrar que erros de classificação são esperados em uma heurística e que é necessário considerar esses erros ao interpretar os resultados.

Outra ameaça identificada é a presença de diferenças na magnitude entre o limiar e o preço do item, bem como padrões confusos e não usuais na descrição dos itens, o que pode afetar a precisão da detecção de sobrepreço. Esses ruídos podem estar relacionados a problemas de registro e à abordagem de agrupamento. Portanto, para uma aplicação real da análise de sobrepreços, é importante tratar esses casos de forma especial, removendo

itens errôneos ou ajustando a abordagem de agrupamento.

6. Conclusão

Este trabalho propôs duas abordagens para detectar sobrepreço em compras públicas: uma metodologia de tratamento e padronização da descrição de itens licitados e uma abordagem estatística baseada na amplitude interquartil para identificar itens com preços acima do esperado. Foram realizados três experimentos com dados reais de compras públicas, comparando diferentes formas de agrupamento dos itens: considerando a descrição original dos itens; utilizando todos os termos presentes na categoria *Palavras*; e com apenas o primeiro termo dessa categoria. Os resultados mostraram que a abordagem que utiliza a categoria *Palavras* apresentou melhores resultados na identificação de sobrepreço, reduzindo a variância dos dados.

No geral, este trabalho apresentou estratégias promissoras para a detecção de sobrepreço e anomalias em compras públicas de produtos e serviços. Embora existam desafios a serem enfrentados, como a falta de padronização nas descrições dos itens licitados, os resultados obtidos sugerem que essas abordagens podem auxiliar na identificação de possíveis irregularidades. No entanto, é importante destacar que as abordagens não são uma solução definitiva, sendo necessário o acompanhamento contínuo e aprimoramento das técnicas utilizadas. Portanto, como trabalhos futuros, planeja-se submeter as licitações identificadas como suspeitas para a avaliação de especialistas, visando melhor validar a abordagem estatística proposta.

Agradecimentos. Ao Ministério Público de Minas Gerais (MPMG) pelo apoio através do Projeto Capacidades Analíticas. Ao CNPq, CAPES e Fapemig pelo apoio aos pesquisadores envolvidos.

Referências

Constantino, K. et al. (2022). Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 304–316, Porto Alegre, RS, Brasil. SBC.

Gabardo, A. C. and Lopes, H. S. (2014). Using social network analysis to unveil cartels in public bids. In *2014 European Network Intelligence Conference*, pages 17–21.

Luna, R. and Figueiredo, D. (2022). Caracterização das licitações públicas no estado do rio de janeiro: Diversidade, licitantes Únicos e redes. In *WCGE*, pages 145–156, Porto Alegre, RS, Brasil. SBC.

Oliveira, G. P. et al. (2022). Detecting inconsistencies in public bids: An automated and data-based approach. In *WebMedia*, pages 193–201, Porto Alegre, RS, Brasil. SBC.

Pereira, A. et al. (2022). Usando redes complexas na identificação de empresas fraudulentas em licitações públicas. In *WCGE*, pages 13–24, Porto Alegre, RS, Brasil. SBC.

Pereira, G. et al. (2021). Classificação taxonômica de categorias de serviços públicos para aplicações digitais. In *WCGE*, pages 119–130, Porto Alegre, RS, Brasil. SBC.

Silva, M. O. et al. (2022). Lipset: Um conjunto de dados com documentos rotulados de licitações públicas. In *SBBD DSW*, pages 13–24, Porto Alegre, RS, Brasil. SBC.