Sobrepreço em compras públicas: Metodologia baseada na identificação de valores discrepantes

Diogo Soares[1], João Pedro D. da Silva[1], Andre Wüst Zibetti[1], Simone S. Werner[1]

[1]Departamento de Informática e Estatística (INE),
Univ. Fed. de Santa Catarina (UFSC), Campus Trindade, Florianópolis-SC-Brasil

{diogo.soares, joao}@grad.ufsc.br,
{andre.zibetti, simone.werner}@ufsc.br

Abstract. *One of the aims of public procurement is to avoid overpriced contracts, but identifying them can be difficult. This article presents a methodology to identify overprices, focusing on detecting outliers, considering as a previous step the unit price adjustment by regression models. To illustrate, we used items from Electronic Invoices (NF-e) for gasoline commonly purchased by Public Institutions in Santa Catarina. The results demonstrated that the use of the price modeling process prior to the application of outlier identification techniques is important to avoid the mistaken identification or non-identification of overpriced items.*

Resumo. *Um dos objetivos das licitações públicas é evitar contratações com sobrepreço, no entanto, identificá-las nem sempre é trivial. Este artigo apresenta uma metodologia para identificar sobrepreço, focando na detecção de valores discrepantes, considerando como um passo anterior o ajuste de modelos para descrever o comportamento do preço unitário. Para exemplificar, utilizou-se itens de Notas Fiscais Eletrônicas (NF-e) de gasolina comum, adquirida por Instituições Públicas de Santa Catarina. Os resultados mostraram que a utilização do processo de modelagem do preço anteriormente à aplicação de técnicas de identificação de valores discrepantes é importante para evitar a identificação equivocada ou não identificação notas de itens com sobrepreço.*

1. Introdução

A preocupação em evitar gastos excessivos e garantir a eficiência no uso dos recursos públicos deve ser uma constante na administração pública. A identificação de possíveis casos de sobrepreço em compras públicas é um dos desafios encontrados no processo de aquisição de bens e serviços pela administração pública. O problema central reside na dificuldade de detectar, de maneira eficiente e precisa, os casos de sobrepreço em um grande volume de dados de preços praticados para produtos semelhantes, geralmente com dados de notas fiscais, o que é crucial para garantir a integridade e o uso correto dos recursos públicos.

A análise manual de notas fiscais que são suspeitas de sobrepreço é uma forma ineficiente e suscetível a erros humanos. A aplicação de métodos estatísticos, utilizando uma abordagem sistemática de identificar casos de sobrepreço por meio do mapeamento de valores discrepantes em itens de notas fiscais de licitações é uma forma mais eficaz de

lidar com esta problemática. Dado que, a identificação de sobrepreço pode ser considerado um problema de análise e identificação de valores discrepantes, para o qual diversas metodologias estatísticas já estão estabelecidas [Rousseeuw and Leroy 1987].

A análise de regressão linear associada a detecção de valores discrepantes pode ser uma estratégia eficiente para identificar irregularidades nos preços praticados nas compras públicas e fornecer indícios de possíveis práticas irregulares de sobrepreço, dada a necessidade de ajuste dos valores praticados por outras variáveis que podem influenciar o preço, como quantidade adquirida, índices de preços associados aos produtos, datas de aquisição entre outros.

[Silva et al. 2024] destacam a importância da análise de sobrepreço em licitações. Eles propuseram uma metodologia para padronização de descrições de itens e uma abordagem estatística usando o intervalo interquartílico para detectar sobrepreço.

Este trabalho apresenta uma metodologia para detectar casos de sobrepreço considerando, anteriormente ao processo de identificação de valores discrepantes, a modelagem do preço unitário, por meio de modelos de regressão, e fornece uma base metodológica robusta para futuras análises e aplicações em outras esferas da gestão pública. A utilização da metodologia proposta é exemplificada utilizando dados de preços da gasolina comum.

2. Material e Métodos

Para execução de todas as etapas apresentadas neste trabalho foi utilizada a linguagem de programação *Python* e suas bibliotecas para manipulação e análise estatística de dados. Nesta seção, apresenta-se em sequência, o conjunto de dados utilizado e a forma de tratamento do mesmo, na sequência a descrição da metodologia de detecção de sobrepreço apresentada em [Silva et al. 2024], utilizada neste trabalho como modelo baseline, e a metodologia para detecção de sobrepreço proposta considerando o ajuste do preço unitário.

2.1. Conjunto de dados utilizado para validação

Dados de 63 milhões de notas fiscais emitidas para entes públicos foram fornecidos pelo Ministério Público de Santa Catarina (MPSC), em formato CSV, contendo 13 variáveis, incluindo descrição, data de emissão, valor unitário, situação, código NCM, código GTIN, entre outras.

Utilizou-se um filtro manual para identificação e seleção dos produtos descritos como gasolina comum, que totalizaram um subconjunto de 1 milhão de itens. Anteriormente a análise de preços, os dados foram tratados, incluindo a aplicação de filtros com base em situação, unidades comerciais, período entre 2013 e 2023, tratamento de valores nulos e conversão de tipos de dados. Considerou-se como intervalo válido para o valor unitário, valores entre menos e mais 10 desvios medianos absolutos, os valores fora deste intervalo foram removidos e avaliados separadamente para identificação de possíveis erros de registro.

Dados externos, como o IPCA e o preço diário do barril de petróleo foram integrados ao conjunto de dados original das notas fiscais. Ressalta-se que, dependendo do produto em estudo, outras variáveis externas podem ser utilizadas para assimilação dos dados.

2.2. Metodologias de sobrepreço

A fim de detectar casos suspeitos de sobrepreço com base no valor unitário comercial de notas fiscais vinculadas a licitações, foram aplicadas três metodologias estatísticas para detecção de sobrepreço, M1, M2 e M3.Os dados utilizados neste trabalho não possuem rotulação manual ou anotação prévia, e caberá a equipe técnica do MPSC avaliar os casos detectados.

2.2.1. M1

Esta metodologia foi aplicada por [Silva et al. 2024] e será considerada como modelo de referência. Consiste na determinação de valores discrepantes proposta por [Tukey 1972], e utiliza a distância interquartílica dos valores unitários comerciais para identificar os valores considerados suspeitos (discrepantes). A distância interquartílica (DIQ) é a diferença entre o valor do terceiro quartil ($Q3$) pelo primeiro quartil ($Q1$). Para obter os valores discrepantes com base nessa distância interquartílica foi calculado o limite superior: $Q3 + 1.5 \cdot DIQ$, e o limite inferior $Q1 - 1.5 \cdot DIQ$. Os valores acima do limite superior foram considerados como suspeitos de sobrepreço e os valores abaixo do limite inferior foram considerados apenas como suspeitos. A M1 foi aplicada aos valores agrupados por ano de emissão da nota fiscal. A Figura 1 (a) ilustra o método, conhecido como diagrama de caixas (*box-and-whisker plot*, popularizado por *boxplot*).

2.2.2. Ajuste do modelo

As metodologias M2 e M3 envolvem a criação de um modelo de regressão linear múltiplo usando a técnica de Mínimos Quadrados Ordinários (OLS) [Wooditch et al. 2021], utilizando variáveis das notas fiscais e de fontes externas. A variável dependente é o valor unitário comercial (preço do produto). As variáveis independentes devem ser consideradas de acordo com a natureza da variavél resposta. Para o caso da gasolina comum considerou-se: ano de emissão da nota fiscal, o índice mensal do IPCA, e o preço diário do barril de petróleo.

A equação do modelo de regressão linear múltiplo é dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \tag{1}$$

em que, Y_i é o valor unitário comercial (variável dependente) para a i-ésima observação; β_0 é o intercepto do modelo; $\beta_1, \beta_2, \beta_3$ são os coeficientes das variáveis independentes X_1, X_2, X_3: ano de emissão, índice de inflação IPCA, e preço do barril de petróleo, respectivamente e ϵ_i é o termo de erro ou resíduo da i-ésima observação.

Para seleção e verificação do ajuste utilizou-se análise gráfica. Após ajustar o modelo, os valores preditos ($\hat{Y}_i$) e os resíduos ($\hat{\epsilon}_i = Y_i - \hat{Y}_i$) são calculados. Para avaliação dos valores discrepantes considerou-se o resíduo padrão ($s_{\hat{\epsilon}}$), que é dado por: $s_{\hat{\epsilon}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{\hat{\epsilon}}}$, em que $\hat{\sigma}_{\hat{\epsilon}}$ é o desvio padrão dos resíduos.

Na abordagem M2 e M3, primeiramente são obtidos os valores preditos e resíduos do modelo ajustado. Em seguida, calculam-se os resíduos padronizados para categorizar os itens como suspeitos ou não suspeitos, com base nas metodologias apresentadas.

2.2.3. M2

A segunda metodologia de detecção consiste em utilizar um parâmetro que determina a quantidade de desvios padrão considerada para identificar os valores suspeitos. Assumindo que os resíduos padronizados tenham uma distribuição normal [Wu 2020], considerou-se como suspeitos, os valores que ultrapassarem a quantidade de 3 desvios. O valor de 3 desvios foi considerado pois, sob condições de normalidade, a probabilidade de se obter um valor fora deste intervalo é menor do que 0.27%. A distribuição normal na Figura 1 contém uma área em vermelho que representa os valores considerados discrepantes [Buckland and Buckland 1979], pois ultrapassaram o limite de três desvios-padrão afastados da média. Como neste trabalho utilizou-se o resíduo padronizado, a distribuição esperada é uma distribuição normal padrão, com média zero e desvio padrão um.

Figura 1. Diagrama de caixas (a) e distribuição normal esperada para os resíduos padronizados (b)

2.2.4. M3

A terceira consistiu em calcular a distância interquartílica (DIQ) dos resíduos padronizados e obter valores discrepantes que sejam maiores que o limite superior $Q3 + 1.5 \cdot DIQ$ ou menores que o limite inferior $Q1 - 1.5 \cdot DIQ$, como exposto em M1, no entanto, difere da proposta de [Silva et al. 2024], pois considera a distância interquartílica dos resíduos padronizados resultantes do modelo ajustado, e não diretamente dos valores unitários comerciais dos itens das notas fiscais.

3. Resultados e Discussão

Primeiramente foi aplicado o método M1 para obter os valores suspeitos de sobrepreço no ano de emissão da nota fiscal. Na Figura 2 é apresentada a distribuição dos valores suspeitos em função do ano de emissão da nota fiscal.

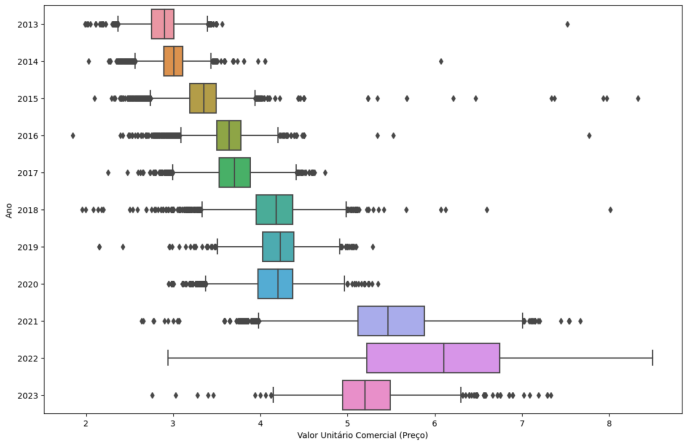

Figura 2. Boxplot dos valores unitários comerciais por ano

A Figura 3 foi construída com o objetivo de investigar o comportamento do valor unitário entre os meses de um mesmo ano, considerando, por exemplo, os anos de 2021 e 2022. Verifica-se, para o ano de 2021, que os pontos identificados como suspeitos no limite superior identificados na Figura 3 em laranja, são valores que foram praticados no final do período anual considerado, com uma clara tendência de aumento do valor unitário. Em outros anos, diferentes tendências são notadas, deixando claro que o procedimento de ajuste do valor do preço unitário por um modelo de regressão pode auxiliar na identificação de valores que realmente se encontram fora do padrão.

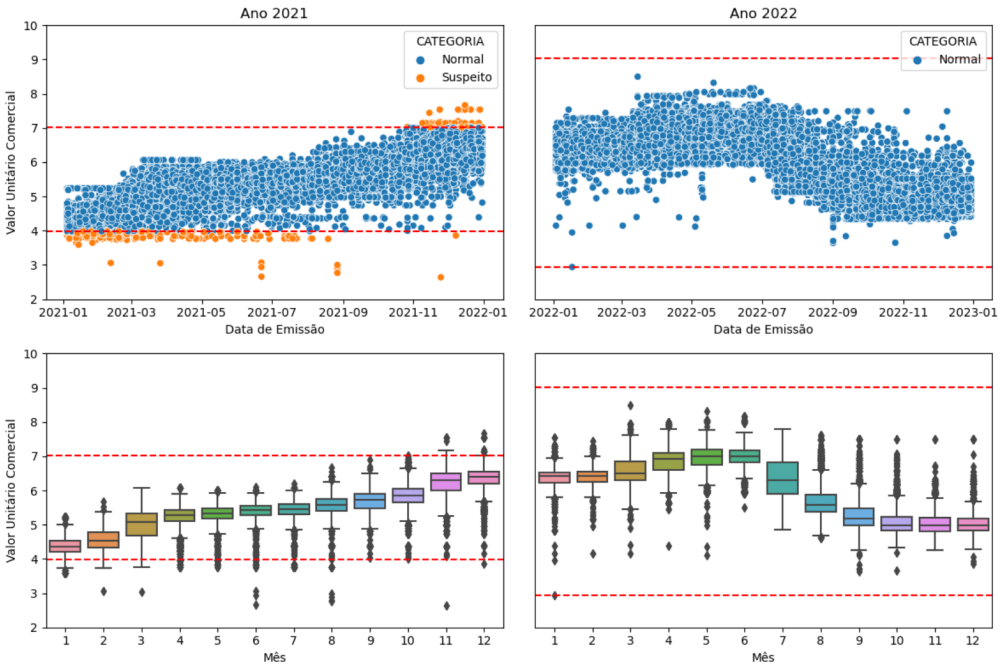

Figura 3. Dispersão dos valores unitários comerciais dos anos 2021 e 2022

Como apresentado na seção 2.2, ajustou-se um modelo de regressão linear múltiplo nos métodos M2 e M3 para obter os resíduos padronizados, incluindo no modelo as variáveis independentes: ano, IPCA e preço do barril de petróleo. A inclusão de

índices e outras covariáveis pode auxiliar no ajuste da tendência do preço unitário, que pode não ser linear, como observado na Figura 3 para 2022. Embora não seja apresentada neste trabalho, a seleção adequada dos modelos para ajuste do preço unitário é de extrema importância. A utilização de modelos inadequados pode levar à indicação incorreta de valores suspeitos.

Para demostrar as diferenças entre os valores suspeitos indicados com as diferentes metodologias elaborou-se a Tabela 1 que compara os preços que serviram como corte para identificação de um item como suspeito ou não e a quantidade de suspeitos separados por ano para cada método aplicados ao conjunto de dados. Ressalta-se que os valores para indicação de suspeitos utilizados nas metodologias M2 e M3 variam ao longo do ano, dado que dependem das covariáveis consideradas no ajuste do modelo, sendo apresentado de forma simplificada o mínimo e máximo do ano.

Ano	**Qtd. Itens**	**Quantidade de suspeitos**			**Preço de corte (R$)**					
					Inferior			**Superior**		
		M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
2013	102711	384	1	2	2.36	1.99	1.99	3.39	3.55	3.55
2014	113261	1629	1	6	2.56	2.03	2.26	3.44	4.05	3.8
2015	112174	2020	28	74	2.74	2.32	2.38	3.93	4.5	4.1
2016	107156	944	4	65	3.07	2.4	2.59	4.2	4.5	4.4
2017	115679	373	8	190	2.99	2.72	2.99	4.41	4.73	4.73
2018	119953	301	47	1255	3.32	3.06	3.31	4.99	5.41	5.13
2019	104771	95	9	134	3.49	3.19	3.38	4.91	5.29	5.06
2020	90119	1164	9	848	3.37	2.98	3.15	4.97	5.35	5.05
2021	98435	1361	8354	22264	3.98	3.49	3.75	7.01	6.56	6.19
2022	97590	0	4859	25736	2.93	4.39	4.46	9.02	7.69	6.82
2023	35653	69	2138	8902	4.12	4.34	4.69	6.3	7.02	6.57

Tabela 1. Limites inferior e superior para identificação dos suspeitos nas 3 metodologias

Verifica-se na Tabela 1 que a quantidade de itens variou significativamente ao longo dos anos. Os preços de corte inferior e superior para cada método apresentaram tendências distintas ao longo dos anos. Destaca-se que o preço de corte superior do método 1 no ano de 2022 foi de 9,02, o que pode indicar algum tipo de anomalia. Em todos os métodos analisados observou-se um aumento nos valores de corte inferior e superior ao longo do tempo, porém nem sempre na mesma proporção.

4. Conclusão

Este trabalho propôs uma metodologia para identificar sobrepreço, focando na detecção de valores discrepantes, considerando como um passo anterior o ajuste de modelos para descrever o comportamento do preço unitário.

Considerando o caso da gasolina comum, ajustou-se um modelo de regressão linear múltiplo utilizando índices de referência como o IPCA e o preço do barril de petróleo que auxiliaram na explicação do comportamento do valor unitário e na identificação de sobrepreço, o que levou a identificação de valores susperitos distintos dos que seriam identificados considerando o método que utiliza os valores unitários não ajustados.

Ressalta-se que o modelo de regressão proposto é apenas um dos possíveis modelos para o ajuste do preço unitário e que modelos inadequados podem levar à identificação equivocada de valores suspeitos. Trabalhos futuros devem explorar técnicas de seleção dos modelos, ajuste de novos modelos para melhorar a qualidade e a flexibilidade do ajuste, tornando o processo de identificação de valores suspeitos mais preciso e contribuindo para a melhor utilização dos recursos públicos.

Agradecimentos

Este trabalho conta com recursos financeiros do projeto *Céos: Inteligência de Dados para a Sociedade*, uma parceria de pesquisa entre a UFSC e o Ministério Público do Estado de Santa Catarina (MPSC) com suporte financeiro do MPSC.

Referências

Buckland, W. R. and Buckland, W. R. (1979). Outliers in statistical data. *Journal of the Operational Research Society*.

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley.

Silva, M. O., Costa, L. G. L., Gomide, L. D., Santos, G. B. D., Oliveira, G. P., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2024). Overpricing analysis in brazilian public bidding items. *Journal of Interactive Systems*.

Tukey, J. (1972). Some graphical and semigraphical displays. *T.A. Bancroft, ed., Statistical Papers in Honor of George W. Snedecor*.

Wooditch, A., Wooditch, A., Johnson, N. J., Johnson, N. J., Solymosi, R., Solymosi, R., Ariza, J. M., Ariza, J. M., Langton, S., and Langton, S. (2021). Ordinary least squares regression. *null*.

Wu, J. (2020). The normal distribution. *Essentials of Pattern Recognition*.