Ciência de Dados e Transparência: Experiências com Dados Públicos do SICOM

Lucas G. L. Costa[1], Marco Túlio Dutra[1,2], Gabriel P. Oliveira[1], Mariana O. Silva[1], Daniela Cruz Soares[3], Luciana de Cassia Silva Faria[3], Wagner Meira Jr.[1], Gisele L. Pappa[1]

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

[2]Universidade Federal de Ouro Preto (UFOP) – Ouro Preto, MG – Brasil

[3]Ministério Público do Estado de Minas Gerais (MPMG) – Belo Horizonte, MG – Brasil

lucas-lage@ufmg.br, marco.dutra@aluno.ufop.edu.br
{gabrielpoliveira,mariana.santos}@dcc.ufmg.br
{daniela,lfaria}@mpmg.mp.br, {meira,glpappa}@dcc.ufmg.br

Abstract. *This paper presents the experience report of the Programa Capacidades Analíticas (PCA) on the use of the public database from the Sistema Informatizado de Contas dos Municípios (SICOM) for applying data science to governmental expenditures. Using the SICOM database, PCA developed advanced techniques in artificial intelligence and data science to analyze public procurement, municipal expenditures, and detect irregularities.*

Resumo. *Este artigo relata a experiência do Programa Capacidades Analíticas (PCA) na utilização da base de dados pública do Sistema Informatizado de Contas dos Municípios (SICOM) para aplicação de ciência de dados no contexto de gastos governamentais. Usando a base de dados do SICOM, o PCA desenvolveu técnicas avançadas de inteligência artificial e ciência de dados para analisar licitações, despesas públicas municipais e identificar irregularidades.*

1. Introdução

Este artigo apresenta o relato de experiência do Programa Capacidades Analíticas (PCA), uma parceria entre o Departamento de Ciência da Computação da Universidade Federal de Minas Gerais (DCC/UFMG) e o Gabinete de Segurança e Inteligência do Ministério Público do Estado de Minas Gerais (GSI/MPMG), na utilização da base de dados do Sistema Informatizado de Contas dos Municípios (SICOM). O PCA está vigente desde janeiro de 2020 e consiste em diversos projetos referentes a pesquisa e desenvolvimento de ferramentas nas áreas de inteligência artificial e ciência de dados para combater a corrupção no setor público.

Cinco dos projetos que compõem o PCA incluem um esforço conjunto para explorar o potencial dos dados do SICOM na identificação de irregularidades em licitações, despesas públicas municipais, análise de sobrepreço, classificação de documentos e curadoria de dados. A iniciativa visa não apenas aplicar técnicas avançadas de análise de dados, mas também desenvolver ferramentas e metodologias que possam ser incorporadas às práticas investigativas e de fiscalização do Ministério Público. Dessa forma, este trabalho contribui para o avanço científico ao enfrentar desafios significativos, como a falta

de padronização e a qualidade variável dos dados disponíveis no SICOM, que frequentemente dificultam análises precisas e eficazes.

2. Sistema Informatizado de Contas dos Municípios (SICOM)

O Sistema Informatizado de Contas dos Municípios (SICOM)[1] é uma plataforma digital desenvolvida pelo Tribunal de Contas do Estado de Minas Gerais (TCE-MG), destinada à centralização e disponibilização de dados financeiros e contábeis dos municípios mineiros. O SICOM desempenha um papel fundamental na fiscalização das contas públicas, facilitando a transparência ao oferecer acesso detalhado a informações sobre receitas, despesas, contratos, licitações e outras operações financeiras. Isso contribui significativamente para a eficiência da gestão pública e fortalece o controle social.

Por ser disponibilizado pelo TCE-MG,[2] o SICOM reúne exclusivamente dados dos municípios de Minas Gerais. Cada município é responsável por alimentar o sistema com seus próprios dados, resultando em diversas fontes com diferentes tipos e características, representando um desafio para validar, assegurar a consistência e garantir a interoperabilidade entre sistemas. Além disso, a extensão desses dados pode afetar a velocidade de processamento e a extração de informações relevantes.

No âmbito do PCA, os dados do SICOM foram obtidos a partir de um armazém de dados estruturados disponível na infraestrutura do MPMG. A base de dados do SICOM engloba informações detalhadas sobre os 853 municípios de Minas Gerais. Até novembro de 2023, o SICOM registra um total de 2.195 órgãos municipais, com um volume considerável de atividades: 508.837 licitações foram realizadas, envolvendo um total de 17.032.217 itens de licitação. Durante esse período, foram contabilizados 35.782.161 lances de fornecedores, resultando em 117.248 contratos formalizados e um total de R$50.712.513 em despesas públicas. Esses dados abrangem o período de 2014 a 2022, oferecendo uma visão abrangente das operações de compras públicas em Minas Gerais.

3. Desafios e Limitações

A partir da experiência adquirida durante os projetos do PCA, foi possível observar que, apesar de ser uma fonte de informações valiosas, os dados disponíveis no SICOM apresentam alguns desafios e limitações que podem impactar suas aplicações. A seguir, são destacados os principais pontos identificados.

Qualidade dos Dados. Os dados referentes a compras públicas apresentam alguns problemas significativos, como anomalias e dados faltantes. Essas inconsistências podem comprometer a precisão das análises realizadas com base no SICOM, dificultando a detecção de padrões de gastos irregulares ou fraudulentos. A presença de licitações com valores extremamente elevados e a falta de licitações registradas na base de dados para determinadas cidades são exemplos de desafios enfrentados na interpretação e na utilização desses dados para promover a transparência na administração pública.

Falta de Padronização. Devido à alimentação descentralizada pelos municípios de Minas Gerais, há uma falta de padronização nos registros, especialmente em campos de

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

[2] Os dados brutos do SICOM estão disponíveis publicamente em https://dadosabertos.tce.mg.gov.br/

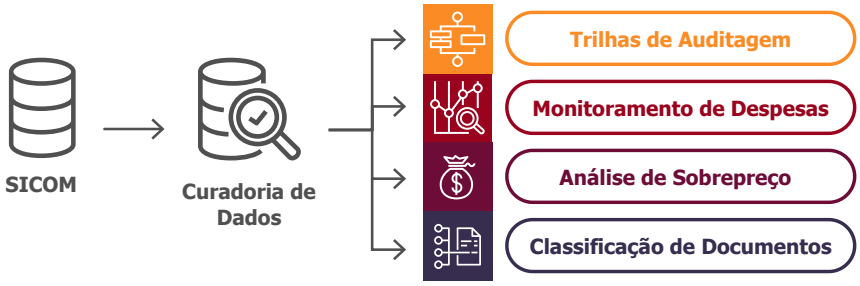

Figura 1. Fluxo de atuações dos cinco projetos nas análises de dados do SICOM.

texto livre como descrições de itens de licitação e suas unidades de medida. Isso dificulta a integração e análise conjunta de dados entre diferentes fontes.

Dados Mascarados. Para proteger a privacidade dos indivíduos, todos os registros de CPF na base do SICOM são mascarados, impedindo a identificação direta de pessoas físicas envolvidas em transações públicas. Essa medida, embora necessária, representa um desafio para análises que dependem da identificação precisa de partes envolvidas.

4. Aplicações

Esta seção apresenta e discute as aplicações desenvolvidas por cinco dos projetos do PCA. Utilizando os dados do SICOM para aplicar ciência de dados no combate à corrupção, foram realizadas explorações de padrões e identificações de anomalias para promover a transparência na gestão pública. A Figura 1 apresenta um fluxo que ilustra as atuações desses cinco projetos, sendo que cada um deles é detalhado a seguir.

4.1. Curadoria de Dados

A partir dos dados brutos do SICOM, um projeto intermediário de curadoria de dados foi necessário devido aos problemas de qualidade descritos na Seção 3. Este projeto visou identificar e corrigir problemas de qualidade nos registros, assegurando que as aplicações subsequentes de combate à corrupção sejam baseadas em dados confiáveis e precisos. Um estudo comparativo realizado por [Oliveira et al. 2022b] avaliou ferramentas *open-source* de qualidade de dados para determinar a melhor opção de implementação para o SICOM. Além do estudo comparativo, também foram implementadas validações de qualidade específicas para dados de licitações do SICOM.

Tais iniciativas foram posteriormente expandidas por [Oliveira et al. 2023a], que estendeu as validações para outras áreas como notas fiscais, contratos e despesas municipais. Além disso, também foram propostas novas métricas de avaliação de qualidade de dados sobre as tabelas do SICOM, considerando o tamanho das tabelas avaliadas e a incidência de problemas referentes a qualidade de dados. A curadoria de dados proposta e implementada em ambos estudos é posteriormente utilizada por [Brandão et al. 2024], que propõe um fluxo semi-automático para a detecção de fraudes em licitações.

Além dos estudos mencionados, o projeto de curadoria de dados também resultou em uma ferramenta denominada Painel de Curadoria de Dados. Essa ferramenta implementa todas as validações de qualidade propostas por [Oliveira et al. 2022b] e [Oliveira et al. 2023a], apresentando visualmente os resultados dessas análises. Isso auxilia os engenheiros de dados na tomada de decisões para mitigar problemas identificados.

O Painel de Curadoria de Dados está disponível exclusivamente para usuários internos do MPMG, com um vídeo institucional disponível para apresentação da ferramenta.[3]

4.2. Trilhas de Auditagem

A partir dos dados referentes a licitações públicas disponibilizados pelo SICOM, diferentes estudos foram realizados para propor e aplicar trilhas de auditagem. Trilhas de auditagem são sequências de passos projetadas para identificar tipos específicos de irregularidades nos dados governamentais [Costa et al. 2022, Oliveira et al. 2023b], visando aumentar a transparência e mitigar práticas fraudulentas.

Inicialmente, [Costa et al. 2022] propuseram 10 trilhas de auditagem, incluindo quatro trilhas sobre empresas que participam de uma mesma licitação com vínculos como sócios, e-mail, telefones ou endereço em comum. As outras seis trilhas focam em características específicas das empresas licitantes, como empresa licitando antes do registro na receita federal, empresa com sanção ativa, empresa com CNPJ inativo, empresa perdedora frequente, empresa vencedora frequente, e empresa participando sozinha da licitação.

Posteriormente, [Oliveira et al. 2022a] introduziram uma abordagem automatizada para detectar incongruências nas licitações presentes na base do SICOM. Essa abordagem usa heurísticas para comparar o segmento de atuação das empresas licitantes com o tipo de item licitado, visando identificar casos em que empresas de um segmento específico estão vencendo licitações de itens não relacionados ao seu ramo de atividade.

O trabalho de [Costa et al. 2022] foi então estendido em [Costa et al. 2023], onde duas novas trilhas foram adicionadas, incluindo uma abordagem semelhante à proposta por [Oliveira et al. 2022a] e uma trilha que investiga licitantes cujos sócios têm vínculos com servidores públicos. Além disso, foi introduzido um sistema de ranqueamento para as licitações enquadradas nas trilhas, priorizando aquelas com maior potencial de irregularidade com base no valor da licitação e na quantidade de trilhas aplicáveis.

De forma similar, [Braz et al. 2023] propuseram duas abordagens para analisar irregularidades em licitações, focando especificamente em empresas de pequeno porte. A primeira identifica licitações onde os licitantes de pequeno porte têm faturamento anual acima do limite estabelecido, enquanto a segunda visa identificar licitantes de pequeno porte vinculados a pessoas jurídicas. Esse estudo foi posteriormente ampliado em [Braz et al. 2024], onde foram adicionadas novas caracterizações dos resultados obtidos.

Os trabalhos de [Costa et al. 2022] e [Costa et al. 2023] também foram estendidos em [Oliveira et al. 2023b], onde sete trilhas de auditagem foram incluídas. Entre elas, destacam-se as duas trilhas focadas em empresas de pequeno porte propostas por [Braz et al. 2023] e três trilhas relacionadas a dados políticos derivadas do estudo de [Mendes et al. 2023], como licitantes que prestaram serviços para campanhas políticas, licitantes cujos sócios são filiados a partidos políticos ou doadores de campanha polıtica. Também foi inclusa uma trilha para detectar licitações onde o menor lance não venceu. A última trilha adicionada visa identificar empresas licitantes em Minas Gerais sem registro de ligação de energia elétrica. Ao todo, 19 trilhas de auditagem foram propostas e detalhadas em [Oliveira et al. 2023b].

[3] Vídeo institucional do Painel de Curadoria de Dados: https://youtu.be/qn4kuvh150Q

Por fim, as trilhas de auditagem são utilizadas como parte do fluxo semi-automático proposto por [Brandão et al. 2024] para detectar fraudes em licitações. Além disso, o desenvolvimento das trilhas resultou na criação de uma ferramenta denominada Painel de Licitações, que implementa todas as 19 trilhas e um sistema de ranqueamento baseado no potencial de fraude de cada licitação. O Painel de Licitações apresenta visualmente os resultados das trilhas e do ranqueamento, facilitando a investigação de fraudes pelos procuradores, promotores e auditores. A ferramenta está disponível apenas para usuários internos do MPMG, com um vídeo institucional disponível para apresentação.[4]

4.3. Monitoramento de Despesas

Um dos projetos do PCA investigou dados relacionados às despesas públicas municipais. Em particular, [Gomide et al. 2023] propuseram uma heurística para a mineração de dados do SICOM, visando identificar possíveis fraudes em municípios cujas despesas estão significativamente acima da média de outros municípios de tamanho semelhante. Esta heurística envolve inicialmente o agrupamento dos municípios com base na faixa populacional do Censo 2010.[5] Uma vez agrupados, é realizado um cálculo estatístico para identificar os municípios com despesas consideravelmente superiores à média para seu respectivo grupo populacional.

A heurística proposta resultou na criação do Painel de Despesas Públicas, uma ferramenta que não apenas incorpora essa metodologia, mas também inclui outras validações e análises temporais das despesas públicas municipais. Desenvolvido exclusivamente para os usuários internos do MPMG, o Painel de Despesas Públicas oferece uma interface visual que facilita a identificação e a análise de potenciais irregularidades financeiras municipais. Um vídeo institucional detalhando o uso da ferramenta está disponível.[6]

4.4. Análise de Sobrepreço

Conforme discutido na Seção 3, um dos principais desafios encontrados nos dados do SICOM é a falta de padronização nas descrições dos itens das licitações. Devido ao campo de texto livre, cada município insere os dados de maneira específica, sem um padrão uniforme, resultando em várias descrições para o mesmo item. Isso dificulta a comparação de itens similares, comprometendo a eficácia da análise de sobrepreço.

Para enfrentar esse desafio, [Silva et al. 2023] propuseram um arcabouço de desambiguação de itens seguido de um cálculo estatístico para identificar produtos com sobrepreço. Esse estudo foi estendido por [Silva et al. 2024b], que validou o arcabouço comparando-o com dados de sobrepreço da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Além disso, [Brum et al. 2024] apresentou um arcabouço mais avançado para a desambiguação de itens de licitações do SICOM, utilizando técnicas de processamento de linguagem natural. Os dados utilizados na desambiguação de itens estão disponíveis no ICPSet, um conjunto de dados estruturados que visa facilitar a análise de compras públicas [Oliveira et al. 2024].

Esses arcabouços de desambiguação e os cálculos de sobrepreço não só foram integrados no fluxo semi-automático para detecção de fraudes em licitações

[4]Vídeo institucional do Painel de Licitações: https://youtu.be/2Kx-f8BiGBI

[5]Censo 2010 - IBGE: https://censo2010.ibge.gov.br/

[6]Vídeo institucional do Painel de Despesas Públicas: https://youtu.be/u1q2o4CusoY

[Brandão et al. 2024], como também resultaram na criação da ferramenta Quanto Custa. Utilizando tais abordagens, a ferramenta permite realizar diversas buscas avançadas sobre os itens das licitações e auxilia na investigação de fraudes e consulta dos preços praticados nas licitações [Costa et al. 2024]. O Quanto Custa está disponível exclusivamente para usuários do MPMG, e um vídeo institucional sobre a ferramenta está disponível.[7]

4.5. Classificação de Documentos

No fluxo semi-automático para detecção de fraudes em licitações proposto por [Brandão et al. 2024], há um módulo dedicado à classificação automatizada de documentos de licitação dos municípios de Minas Gerais. Esses documentos são coletados dos portais da transparência de cada município, e as licitações presentes nesses documentos são as mesmas contidas nos dados de licitações do SICOM. Essa classificação é crucial tanto para o módulo de curadoria de dados desse fluxo quanto para suas aplicações na detecção de fraudes e sobrepreço nas licitações do SICOM.

Além disso, no estudo referente a processamento de linguagem natural conduzido por [Silva et al. 2024a], são avaliados se fatores específicos impactam os resultados de cinco tarefas de classificação textual. Os fatores analisados incluem o tamanho da base de dados, o idioma do modelo de linguagem e o domínio da base de dados usada para o pré-treinamento do modelo. Dessas cinco tarefas, duas utilizam dados de itens de licitações do SICOM. A primeira tarefa classifica se um item de licitação é um produto ou um serviço, enquanto a segunda classifica a natureza da despesa associada a cada item de licitação.

5. Considerações Finais

O presente estudo revelou a complexidade e os desafios inerentes à utilização dos dados do SICOM para a análise e detecção de irregularidades em compras públicas. Apesar das limitações de qualidade dos dados, como inconsistências e falta de padronização, diversas abordagens foram desenvolvidas para mitigar esses problemas. A curadoria de dados mostrou-se essencial como um projeto inicial, identificando registros problemáticos e implementando validações para garantir a confiabilidade dos dados utilizados nas análises. Além disso, os projetos das trilhas de auditagem, dos métodos de análise de sobrepreço e das despesas municipais proporcionaram *insights* valiosos para identificar possíveis fraudes, promovendo maior transparência e eficiência na gestão pública.

O desenvolvimento de ferramentas como o Painel de Curadoria de Dados e o Quanto Custa demonstrou não apenas a capacidade de aplicação prática das metodologias propostas, mas também a necessidade contínua de inovação e aprimoramento tecnológico para enfrentar os desafios em dados governamentais. A integração de técnicas avançadas de processamento de linguagem natural e a colaboração interdisciplinar foram fundamentais para o sucesso desses projetos. A partir dessas experiências, abre-se um caminho promissor para futuras pesquisas e desenvolvimentos na área de análise de dados governamentais, visando sempre a melhoria na gestão e no combate à corrupção.

Agradecimentos. Ao Ministério Público do Estado de Minas Gerais pelo apoio através do Programa Capacidades Analíticas. Ao CNPq, CAPES e FAPEMIG pelo apoio.

[7] Vídeo institucional do Quanto Custa: https://youtu.be/Hm8KHYi-2sc

Referências

Brandão, M. A. et al. (2024). Plus: A semi-automated pipeline for fraud detection in public bids. *Digit. Gov.: Res. Pract.*, 5(1).

Braz, C. S. et al. (2023). Análise de irregularidades em licitações públicas com foco em empresas de pequeno porte. In *WCGE*, pages 94–105. SBC.

Braz, C. S. et al. (2024). Exploring irregularities in brazilian public bids: An in-depth analysis on small companies. *Journal on Interactive Systems*, 15(1):349–361.

Brum, P. P. V. et al. (2024). Unsupervised grouping of public procurement similar items: Which text representation should I use? In *LREC-COLING*, pages 17176–17185. ELRA and ICCL.

Costa, L. G. L. et al. (2024). Quanto Custa: Banco de Preços de Compras Públicas do Estado de Minas Gerais. In *SBBD DS-CoPS*. SBC.

Costa, L. L. et al. (2022). Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In *BraSNAM*, pages 37–48. SBC.

Costa, L. L. et al. (2023). Identification of suspected fraud bids through audit trails. *iSys - Brazilian Journal of Information Systems*, 16(1):13:1–13:23.

Gomide, L. D. et al. (2023). Mineração de dados sobre despesas públicas de municípios mineiros para gerar alertas de fraudes. In *SBBD*, pages 378–383. SBC.

Mendes, B. M. A. et al. (2023). Impacto de Doações Eleitorais no Faturamento de Empresas: Um Estudo nas Eleições Municipais em Minas Gerais. In *SBBD*, pages 420–425. SBC.

Oliveira, G. P. et al. (2022a). Detecting inconsistencies in public bids: An automated and data-based approach. In *WebMedia*, pages 182–190. ACM.

Oliveira, G. P. et al. (2022b). Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa. In *SBBD*, pages 116–127. SBC.

Oliveira, G. P. et al. (2023a). Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management*, 14(1).

Oliveira, G. P. et al. (2023b). Ranqueamento de licitações públicas a partir de alertas de fraude. In *BraSNAM*, pages 1–12, Porto Alegre, RS, Brasil. SBC.

Oliveira, G. P. et al. (2024). ICPSet: Um Conjunto de Dados Estruturados de Itens de Compras Públicas. In *DSW*. SBC.

Silva, M. O. et al. (2023). Análise de sobrepreço em itens de licitações públicas. In *WCGE*, pages 118–129. SBC.

Silva, M. O. et al. (2024a). Evaluating domain-adapted language models for governmental text classification tasks in portuguese. In *SBBD*. SBC.

Silva, M. O. et al. (2024b). Overpricing analysis in brazilian public bidding items. *Journal on Interactive Systems*, 15(1):130–142.