Detecção de anomalias em sistemas de administração de frotas públicas municipais

Cesar R. V. Trindade[1], **Pablo S. Chaves**[2], **Marcelo Teixeira**[1]

[1] Departamento de Informática – Universidade Tecnológica Federal do Paraná (UTFPR)
Pato Branco – PR – Brasil

[2]Eficax Consultancy Ltda.

crvtrindade@gmail.com, eficaxconsultancy@gmail.com, mtex@utfpr.edu.br

Abstract. *Systemic corruption is difficult to detect and treat, making a solid link to impunity. In Information Systems, it is possible that legitimate data are manipulated with non-legitimate intent, shadowing reality with the purpose of circumventing the law. Detecting such patterns of illegitimacy implies tracking and crossing a wide and complex chain of information, of multiple natures, which is manually unfeasible. This article applies local density concepts to extract patterns and trends of irregularities in databases of supply transactions in municipal fleets. In the evaluated period, there were 27 abnormal detections for gasoline refueling and 90 for diesel oil.*

Resumo. *A corrupção sistêmica é de difícil detecção e elo sólido de impunidade. Em Sistemas de Informação, é possível que dados legítimos sejam manipulados com intenção não legítima, sombreando a realidade com o propósito de burlar a lei. Detectar tais padrões implica em rastrear e cruzar uma complexa cadeia de informações, de múltiplas naturezas, o que é manualmente inviável. Este artigo aplica conceitos de densidade local para extrair padrões e tendências de irregularidades em bases de dados com transações de abastecimento em frotas municipais. No período avaliado, foram 27 detecções anormais para abastecimentos com gasolina e 90 para óleo diesel.*

1. Introdução

Práticas de corrupção sistêmica, diferente do que em geral se toma como corrupção, não incidem apenas sobre a esfera pública-política. O senso comum passa a ideia de que o cidadão não se nega a obter vantagem quando lhe é conveniente, indicando que o *jeitinho* e a *malandragem* são heranças culturais (corrupção endêmica). Isso, em conjunção com a ineficácia no cumprimento das leis que disciplinam e penalizam atos de corrupção, estruturam a impunidade no país.

Junto à dimensão alarmante da corrupção no país [Bochenek and Pereira 2018], a quantidade de valores ilegalmente tomados dos cofres públicos é reflexo da ousadia praticamente irrefreável de corruptos e corruptores, afetando principalmente aos menos favorecidos, que mais dependem de recursos e serviços públicos. Dados do IPC 2020 [IPC 2020], mostram que grande parte dos países progrediu pouco ou nada na última década, em relação ao combate à corrupção. Mais de dois terços dos países obtiveram uma pontuação abaixo de 50, em uma escala que vai de 0 a 100, em que 0 significa

altamente corrupto e 100 significa muito íntegro. O Brasil figura, nessa pesquisa, na 94ª posição num ranking de 180 países, com 38 pontos, frente a 35 pontos em 2019, o que sugere uma estagnação no combate à corrupção no país.

Na prática, a corrupção produz cinco efeitos devastadores [Warde 2018]: transforma o Estado e as suas funções em coisas no mercado; desnatura as demais instituições para as submeter aos fins próprios da corrupção; usurpa, ao se apropriar do Estado, a energia vital dos trabalhadores; falseia a concorrência entre os agentes econômicos, para incrementar o poder de mercado de uns em detrimento de outros, até o seu expurgo dos mercados e, por fim, é um obstáculo ao desenvolvimento das nações, promovendo a pobreza e afrontando a dignidade das pessoas.

O combate à corrupção deve ser, portanto, visto como um ato de intolerante reação à impunidade e aos desvios de conduta com a coisa pública, passível de ampla e severa punição, forçando assim aqueles que ocupam posições oficiais de trabalho ao exercício digno e incorruptível de suas atribuições. Entretanto, a tarefa de reagir aos desvios de conduta e punir a corrupção esbarra, sobretudo, em detectar atos ilícitos. Quando praticados de modo explícito, a detecção costuma ser simples e acompanhada de provas cabais e, em geral, materiais. Porém, para uma vasta classe de atos contemporâneos de ilicitude, eventuais indícios são de ordem não material, muitas vezes mascarados em fragmentos de dados cuja derivação do valor semântico depende de complexo processamento computacional. Dessa forma, os avanços nas áreas de hardware e software, muito embora agreguem vantagens sem precedentes à sociedade [da Silva Eleutério and Machado 2011], também subsidiam novos mecanismos que podem ser usados para práticas ilegais.

De fato, avanços na área da computação e telecomunicações culminaram na geração de uma superabundância de dados, de modo que a capacidade de registrar dados supera a habilidade de análise e extração do conhecimento destes dados [de Castro and Ferrari 2016]. Nesse sentido, a possibilidade de aplicar técnicas e ferramentas que convertam, de forma automática e prática, dados em informações úteis, tem se tornado estratégica. A literatura expõe um movimento significativo na ampliação das ferramentas voltadas ao combate a fraudes [Lopes 2019, Santos et al. 2017]. Contudo, aplicações em dados municipais são ainda incipientes, principalmente pelo fato de que esse domínio não possui o mesmo grau de transparência em comparação a dados federais ou mesmo estaduais. Dados municipais são, muitas vezes, armazenados localmente ou, quando não locais, em ambientes de difícil acesso ao cidadão, e apenas amostragens genéricas (em geral obrigatórias) e de difícil compreensão comum, são expostas.

É dessa lacuna que emerge o objetivo deste trabalho, que aplica o algoritmo Local Outlier Factor (LOF) [Breunig et al. 2000], baseado no conceito de densidade local através dos *k*-vizinhos mais próximos (k-Nearest Neighbors – kNN), em que a distância é usada para avaliar a densidade, sobre bases de dados públicos relativos à manutenção de frotas municipais. É mostrado como certos padrões de anomalia podem ser detectados e associados a padrões de ilegalidade. Para isso, serão aplicadas técnicas de engenharia de dados para preparar morfologicamente os dados a serem explorados, relativos à manutenção de frotas de uma prefeitura municipal. Sobre os dados pre-processados, aplica-se então o método de Descoberta de Conhecimentos em Bases de Dados (Knowledge Discovery in Databases – KDD), que inclui os algoritmos kNN, por meio da ferramenta *Rapidminer*. A pesquisa é, portanto, de natureza quantitativa e estatística, de

posição epistemológica principal positivista-crítica, que emprega métodos experimentais sobre um estudo de caso real.

O restante deste documento está estruturado da seguinte forma: a Seção 2 apresenta uma breve tomada do estado da arte; a Seção 3 apresenta os conceitos centrais que sustentam os resultados principais apresentados na Seção 4. Por fim, a Seção 5 discute as conclusões e aponta para algumas direções em aberto.

2. TRABALHOS RELACIONADOS

Na Política de Dados Abertos, instituída pelo Poder Executivo Federal [Planalto 2016], dados abertos são aqueles acessíveis ao público, em meio digital, processáveis por máquina, referenciados na internet e disponibilizados sob licença que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte [Pansani and Ferneda , e Doraliza Monteiro e Anderson Reis 2020]. Análises sobre dados abertos não costumam ser prioritárias pela gestão pública, mas se tornam exequíveis a qualquer indivíduo, viabilizando, por exemplo, a construção de ferramentas de reconhecimento de fraudes e investigação criminal [Santos et al. 2017].

Iniciativas nessa direção incluem a "Operação Serenata de Amor" [Brasil 2021], um projeto de inteligência artificial que usa a ciência de dados para fiscalizar gastos públicos e compartilhar estas informações com os cidadãos; e "Rosie", seu desmembramento, uma máquina inteligente que analisa gastos reembolsados por deputados federais e senadores durante o exercício de sua função, pela Cota para Exercício da Atividade Parlamentar, identificando transações suspeitas e estimulando a população a contestá-las. Complementarmente, outras abordagens focam na identificação e combate às irregularidades já na origem da ação. Por exemplo, os Robôs "Mônica" (Monitoramento Integrado para o Controle de Aquisições), "Sofia" (Sistema de Orientação sobre Fatos e Indícios para o Auditor) e "Alice" (Análise de Licitações e Editais), aplicam inteligência artificial em dados do Tribunal de Contas da União (TCU) para identificar e combater irregularidades em licitações. Embora essas soluções sejam adotadas como forma de evitar o desvio de recursos e dano ao erário, a operação Serenata de Amor tem um caráter mais voltado para a moralidade e honestidade, enquanto que as tecnologias empregadas pelo TCU são mais direcionadas ao controle de operações.

Outras abordagens exploram mineração de dados para identificar anomalias semânticas, o que se alinha aos propósitos deste artigo, embora não sejam específicos para o domínio de aplicação aqui considerado. Por exemplo, [Lubambo 2008] explora aprendizado supervisionado ao propor um modelo capaz de classificar empresas suspeitas de operarem exportações fictícias, fundamentando-se na metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM). O estudo explora os algoritmos de indução de regras APRIORI e TERTIUS e é ilustrado por meio de experimentos sobre dados reais da Secretaria da Fazenda de Pernambuco.

Já [Kintopp 2017] identifica possíveis anomalias nos dados disponíveis no portal da transparência. Para isso, este trabalho aplica a metodologia de KDD por meio do algoritmo LOF , similar ao modelo aplicado neste artigo, porém com o objetivo de atribuir uma pontuação de anormalidade para cada instância da base, utilizando técnicas baseadas nos k-vizinhos mais próximos. Em direção similar, mas se utilizando de outras técnicas e outras bases de dados públicos, [Lopes 2019] visa classificar gastos públicos para a

detecção de possíveis fraudes. Para tal, o trabalho compara a performance das técnicas como *Regressão Logística*, *Árvore de Decisão* e *Gradient boosting*.

Já o trabalho de [Assunção et al. 2016] descreve o INFOSAS, um sistema interativo de detecção de anomalias no sistema de pagamento aos prestadores de serviços ao Sistema Único de Saúde (SUS) para posterior auditoria. O sistema procura detectar dois tipos de discrepâncias: um valor médio excessivo cobrado por procedimentos dentro de um alvo e um número excessivo na produção de um procedimento por parte de um estabelecimento. Um dos resultados mais importantes do INFOSAS é a sua capacidade de análise do volume total de produção de serviços de saúde em busca de anomalias, considerando cada prestador de serviço.

O trabalho de [Silva 2020] estudou o comportamento de usuários de empresas de telefonia ao longo de um mês fazendo uso de métodos de detecção de anomalias de modo a encontrar padrões de eventos que divergem do comportamento normal dos usuários. Através deste estudo, com o emprego dos métodos *Boxplot*, *Isolation Forest*, *DBSCAN* e *KMeans*, desenvolveu um classificador para auxiliar na verificação da confiabilidade do usuário a fim de viabilizar o redirecionamento de fluxos alternativos.

Constata-se, na literatura, um movimento significativo na ampliação das ferramentas voltadas ao combate a fraudes; contudo, aplicações em dados municipais são ainda incipientes, principalmente pelo fato de que esse domínio não possui o mesmo grau de transparência em comparação a dados federais ou mesmo estaduais. Fato é que dados municipais são, muitas vezes, armazenados localmente ou, quando não locais, em ambientes de difícil acesso ao cidadão, e apenas amostragens genéricas (em geral obrigatórias) e de difícil compreensão comum, são expostas. É dessa lacuna que emerge o objetivo deste trabalho, o qual envolve os conceitos descritos na seção seguinte.

3. REFERENCIAL TEÓRICO

Um aspecto importante da extração de conhecimento está atrelado à habilidade para detectar instâncias ou certos eventos de dados cujo teor destoe de um determinado padrão esperado, ou seja, se diferenciam de outros elementos de uma mesma amostra. Quando detectado, esse fenômeno é associado a uma *anomalia* [Hodge and Austin 2004], que pode ou não possuir significado prático, mas sobre o qual se deseja investigar. Existem diversas abordagens para detecção de anomalias e a escolha depende das características do domínio de aplicação sob análise e dos aspectos a serem ponderados [de Castro and Ferrari 2016]. Porém, o fator que mais influencia uma abordagem de detecção de anomalias está relacionado à disponibilidade e ao uso dos rótulos de dados. Desta forma, sendo a detecção de anomalias parte integrante do aprendizado de máquina, existem duas grandes divisões de formas pela qual o aprendizado pode ocorrer: supervisionado e não supervisionado.

O aprendizado supervisionado consiste em reconhecer objetos anômalos e normais por meio de um modelo aplicado sobre um conjunto de treinamento com dados rotulados. Diferentemente, no aprendizado não supervisionado, usado nesse trabalho, não existe nenhuma proposição sobre o rótulo de dados e as anomalias são identificadas sem o conhecimento prévio das classes ditas normais e anômalas. Desta forma, os dados mais distantes de certo padrão informado serão reconhecidos como possíveis anomalias.

3.1. Algoritmo Local Outlier Factor - LOF

Algoritmos baseados em distância local consistem, em geral, em explorar as distâncias entre objetos do conjunto de dados sob análise, para estimar desvios entre determinado objeto e seus vizinhos. Tais desvios podem ser associados, por exemplo, a outliers ou certas anomalias que se queira detectar. Dentre os algoritmos desta natureza, cita-se o LOF [Breunig et al. 2000]. Esse algoritmo consiste de um conjunto de dados de entrada, de uma etapa de pré-processamento, visando o ajuste dos dados para o propósito de detecção de distância, e do processamento do resultado final, cuja precisão está em geral associada à qualidade dos dados de entrada e do ajuste de pre-processamento.

Dado um conjunto de dados de entrada, a etapa de pré-processamento é ainda subdividida em subetapas. A primeira visa eliminar dados ditos sujos ou desnecessários; a segunda implementa a transformação dos dados, ou seja, modificações que permitam processar cálculos numéricos; por fim, a terceira envolve a padronização dos dados para escalas normalizadas a fim de evitar que valores de atributos que destoem de outros valores possam deturpar o processamento de resultados.

Nesse contexto, a técnica de *Max-Min* [Han and Pei 2012] é uma das formas mais comuns de normalização e é utilizada neste trabalho. Sua aplicação consiste em transformar atributos em um mesmo intervalo de valores, por exemplo, [0, 1]. A equação do *Max-Min* é mostrada na Eq. (1).

$$a' = \frac{a - min_a}{max_a - min_a}(new\_max_a - new\_min_a) + new\_min_a \tag{1}$$

A partir de uma entrada pré-processada, o algoritmo LOF explora a distância entre a vizinhança de amostras. O método realiza uma marcação em cada ponto de dados e efetua um cálculo da proporção das densidades médias dos vizinhos do ponto, com a densidade do próprio ponto. A densidade estimada de um ponto p é o número de vizinhos de p dividido pela soma das distâncias até os seus vizinhos, como mostra a Eq. (2):

$$f(p) = \frac{k}{\sum_{x \in N(p)^{d(p,x)}}} \tag{2}$$

em que:

- $N(p)$ o conjunto de vizinhos do ponto p;
- k é o número de pontos desse conjunto;
- $d(p, x)$ é a distância entre os pontos p e x.

O LOF de um objeto é calculado da seguinte forma: havendo uma quantidade k de vizinhos mais próximos, utilizar a Eq. (3) para encontrar o LOF de cada objeto da base. Serão consideradas anomalias todos os objetos $x_i = 1, \ldots N$, cujo $\mathrm{LOF}_k(x_i) >> 1$.

$$\mathrm{LOF}_k(x_i) = \frac{1}{lrd(x_i)} \cdot \frac{\sum_{x_j \in N_{k(x_i)}{}^{lrd(x_j)}}}{k} \tag{3}$$

O valor de LOF é aproximadamente 1 para objetos que se encontram dentro de um grupamento (cluster) e variam entre limites superior e inferior para os outros objetos,

que então são interpretados através de uma heurística de classificação de objetos por seu maior valor de LOF no intervalo. É possível reconhecer anomalias em um conjunto de dados que não se enquadrariam como anomalias em outra região do mesmo conjunto, devido à abordagem local. Um objeto a uma curta distância de um cluster com grande densidade é considerado anomalia, ao passo que outro objeto no interior de um cluster disperso pode apresentar distâncias parecidas com seus vizinhos e por consequência não ser caracterizado como anomalia.

4. Metodologia e resultados

Este trabalho analisa uma parcela do domínio em que a detecção de anomalias se enquadra. O conjunto de dados extraídos contém 22447 lançamentos de registros reais relacionados à frota de uma Prefeitura Municipal, cuja identificação é preservada por questões de exposição e eventual conflito com os propósitos deste trabalho, que é o de revelar possíveis anomalias nas transações sem, no entanto, associar tais anomalias com atos de ilicitude. A metodologia proposta para tal, é sumarizada na Figura 1.

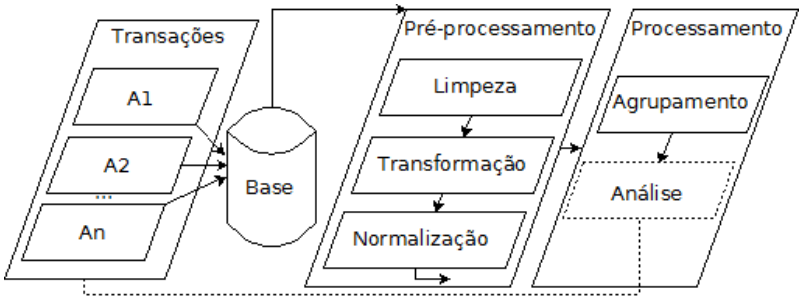

Figura 1. Sumário da metodologia proposta.

No banco de dados em questão, a entidade MovimentoVeiculo armazena registros de consumo de combustíveis e lubrificantes, serviços a executar e troca de pneus. Sobre essa tabela foi executada uma filtragem a fim de retornar um subconjunto T de atributos sobre os quais se tem interesse em determinar possíveis anomalias. Dessa forma, definiu-se

$$\mathrm{T} = \langle \mathsf{cdVei}, \mathsf{cdTipoMov}, \mathsf{dtEv}, \mathsf{hrEv}, \mathsf{vlVal}, \mathsf{qtMov}, \mathsf{dsHis} \rangle.$$

Nesse esquema resultante, cdVei é o código do veículo, cdTipoMov se refere ao tipo de combustível usado, dtEv representa a data em que ocorreu o evento de abastecimento, hrEv representa o horário em que ocorreu o abastecimento, vlVal é o valor pago pelo combustível abastecido, qtMov se refere à quantidade abastecida de combustível e dsHis representa a descrição da quantidade e do tipo de combustível.

Na etapa de preparação da base de dados criou-se uma rotina computacional para reduzir a dimensão e selecionar apenas os dados que contém instâncias de abastecimentos de gasolina (2640 instâncias) e óleo diesel (8307 instâncias). Objetos com dados ausentes se mostraram presentes no atributo vlVal (referente ao valor pago pelo abastecimento) em 53 instâncias de abastecimento com gasolina e em 94 de abastecimento com óleo diesel. Observou-se que os dados ausentes se concentravam no período de 18 de novembro de 2010 a 1º de dezembro de 2010. Utilizando-se da imputação de valores do tipo *hot-deck*, os valores ausentes foram complementados com o valor do mesmo atributo de um

objeto similar selecionado, neste caso, o valor do preço do combustível à época. A frota cadastrada na base de dados, 56 veículos com o combustível gasolina e 35 com óleo diesel teve seu domínio discretizado em 18 grupos conforme a capacidade do tanque a fim de permitir uma melhor análise dos dados. Esse procedimento foi realizado durante a seleção dos atributos para a análise.

A limpeza dos dados, com o intuito de imputar valores ausentes, reduzir a dimensão da base, bem como adequar a morfologia dos dados para o processo de análise, foi empregada em especial nos atributos dtEv que se refere a data em que ocorreu o abastecimento e dsHis referente ao volume de combustível abastecido. Aplicou-se a técnica de normalização *Max-Min* referenciada na Eq. (1) com um intervalo de valores [0,1] no atributo qtMov. A distância selecionada no algoritmo foi a euclidiana.

Para a etapa da mineração de dados foi empregado o software *Rapidminer*. No experimento, calculamos os outliers locais com LOF > 1.5. A entidade Item representa as características dos veículos e possui quinze atributos. Tendo em vista que a capacidade do tanque de combustível não se faz presente na entidade, tal informação, de valor fundamental para análise dos dados após a aplicação do algoritmo LOF , foi incluída na base de dados tendo como parâmetro os valores existentes nas fichas técnicas de cada veículo. Essas informações foram colhidas nos sites do fabricante.

A Tabela 1 apresenta a quantidade de combustível consumida pelos veículos e o valor pago pela Prefeitura no período de 22 de outubro de 2010 a 30 de julho de 2016.

Tipo	**Qtd veículos**	**Litros**	**Valor**
Óleo diesel	35	855.370,02	R$ 1.824.943,61
Gasolina	56	73.039,64	R$ 198.947,76

Tabela 1. Quantidade de combustível por tipo e valor pago.

Os resultados mostram que a aplicação do algoritmo LOF para identificar anomalias como possíveis indícios de fraudes apresentou objetos Falsos Positivos (FPs), em virtude da variação entre o menor e o maior número de volume de combustível. Abastecimentos com óleo diesel ficaram dispostos entre 5 e 800 litros (Figura 2), ao passo que aqueles com gasolina ficaram entre 5 e 65 litros (Figura 3). Para uma melhor visualização dos gráficos, o volume de combustível foi dividido em múltiplos de 5.

Entretanto, o algoritmo também reconheceu instâncias Verdadeiros Positivos (VPs), sugerindo a existência de registros de abastecimentos acima do limite do tanque de combustível. Ao separarmos as instâncias por veículos, de maneira mais específica, isto é, que possuem um certo padrão na distribuição dos valores no abastecimento, o algoritmo LOF distinguiu os valores anormais.

Para o veículo identificado como Cod74, abastecido com óleo diesel, do total de 21 abastecimentos o algoritmo detectou 13 instâncias caracterizadas como outliers, sendo 10 VPs e 3 FPs, com uma acurácia de 85% e precisão de 77%. Se verificarmos que, para este veículo, a capacidade máxima de combustível no tanque é 275 litros, os outliers identificados como VPs caracterizam, em tese, uma provável fraude já que é impossível inserir um volume de combustível além da capacidade de um reservatório. A Tabela 2 mostra o total de outliers identificados, bem como as instâncias consideradas FPs.

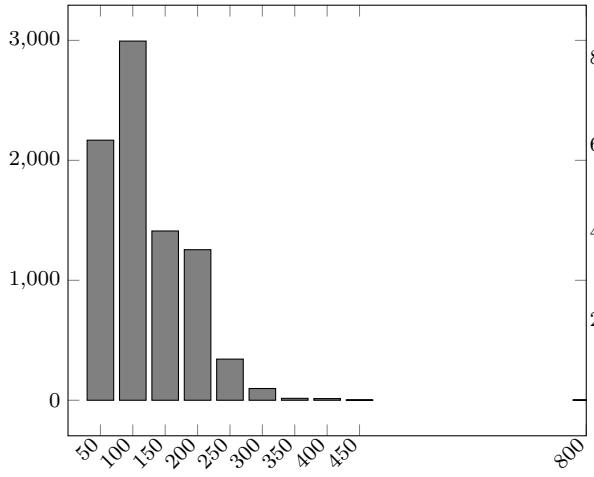

Figura 2. Óleo diesel.

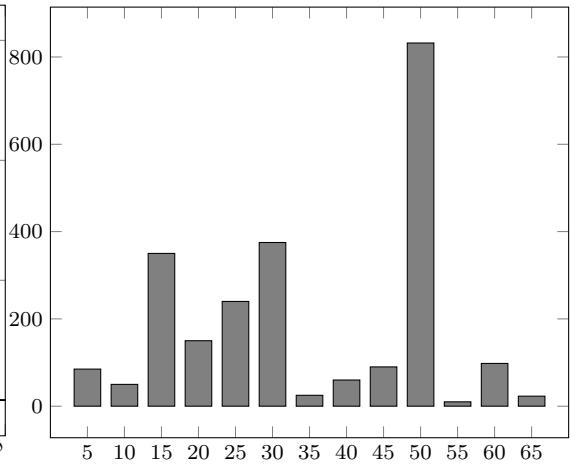

Figura 3. Gasolina.

Veículo Cod74 - diesel				Veículo Cod04 - Gasolina			
Abast.	lts	Outlier	Detecção	Abast.	lts	Outlier	Detecção
2	80	2.559	normal	1	60	10.174	outlier
1	250	2.002	normal	1	58	8.570	outlier
6	400	6.831	outlier	2	3	6.563	outlier
1	600	2.344	outlier	1	7	4.732	outlier
3	800	3.516	outlier	1	6	3.979	outlier

Tabela 2. Anomalias detectadas para o abastecimento de dois veículos.

Para o veículo identificado como Cod04, abastecido com gasolina, em que o tanque permite no máximo 50 litros, os maiores outliers representaram 6 abastecimentos. Da mesma maneira, foram detectadas anomalias para pequenos abastecimentos, principalmente em veículos pesados, os quais possuem um reservatório de combustível de grande capacidade, tendo em vista as horas de trabalho do equipamento e o grande consumo de combustível. De alguma forma, foge aos padrões estes abastecimentos porque equipamentos deste tipo habitualmente necessitam de uma grande quantidade de combustível para operação. A Tabela 3 apresenta estes dados.

Cód.	**Abast.**	**Combustível (lt)**	**Outlier**	**Tanque (lt)**
8	1	10	15.703	160
10	1	10	6.980	160
11	1	10	4.919	340
17	2	10	3.648	150
39	1	10	4.585	105
46	1	10	3.604	103
47	1	10	4.424	103
63	1	5	4.593	380
73	1	10	4.909	160

Tabela 3. Anomalias detectadas para o caso de pequenos abastecimentos.

Por fim as Figuras 4 e 5 ilustram as anomalias detectadas pelo algoritmo LOF na base de dados para os veículos abastecidos com óleo diesel e gasolina. O eixo x caracteriza a quantidade de combustível e o eixo y as anomalias. Observamos que as

maiores anomalias estão localizadas precisamente na área do gráfico que marca as maiores quantidades de combustível, algumas delas acima da capacidade do tanque.

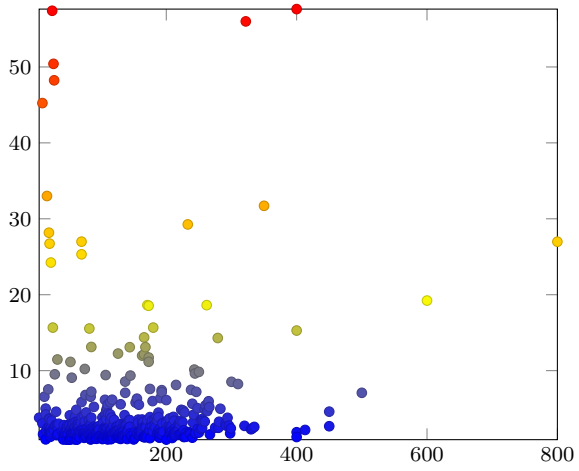

Figura 4. Anomalias - diesel.

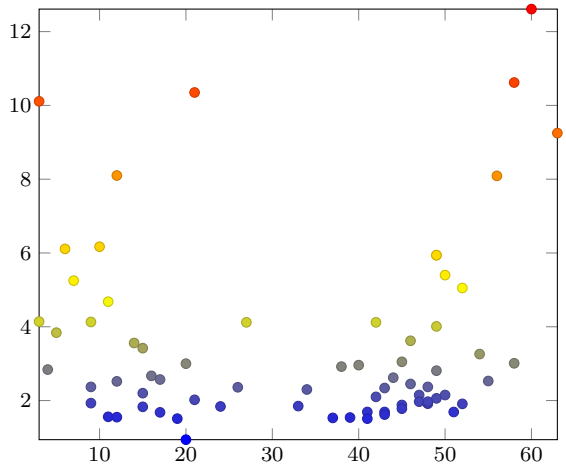

Figura 5. Anomalias - gasolina.

Concluímos que, de 10947 registros de abastecimentos, 117 foram identificados como VPs, sendo 1,07% do total, com uma média de 19 abastecimentos acima da capacidade do veículo a cada ano, no período de 22 de outubro de 2010 a 30 de julho de 2016. Não podemos afirmar que houve fraude nem qualquer tipo de irregularidade nos abastecimentos. Tendo em vista a atual legislação Brasileira, qualquer software de análise de dados será tão somente formulador de indícios de um possível desvio de conduta perante a um padrão pré concebido legalmente, sempre sujeito a análise da autoridade competente para determinar se há (ou não) desvio de caráter e ilegalidades. Assim, não é intenção deste trabalho inferir fraude. A intenção é tão somente detectar e alertar para o desvio de certos comportamentos que possam despertar a curiosidade do gestor em auditar.

É plausível concluir que o algoritmo empregado indica que alguns veículos apresentaram quantidade de combustível abastecida além da capacidade do reservatório e consequentemente estão em desarmonia em relação aos demais. Igualmente, deve ser observado que, em se tratando de máquinas e equipamentos, é usual conduzir recipientes com combustível para situações de trabalho em áreas distantes de recursos de reabastecimento.

5. Conclusões

Este trabalho apresentou uma abordagem de ciência de dados para a detecção de padrões de irregularidades em bases públicas municipais. A abordagem permite que uma classe particular de possíveis atos ilícitos, geralmente oculta a olho humano em meio à complexidade e quantidade de dados públicos, seja explorada de modo automático, revelando-se uma ferramenta online de importância fundamental para a rotina operacional de organizações públicas, em particular as prefeituras.

Um experimento envolvendo dados reais de controle de frotas municipais foi conduzido para ilustrar a abordagem. Resultados mostram que 27 detecções anormais para abastecimentos com gasolina e 90 para óleo diesel, no período avaliado. Isso, porém, não implica necessariamente em fraudes, mas sugere que a transação não seguiu um padrão esperado e é plausível de averiguação. Trabalhos futuros visam integrar a abordagem

com *web crawling*, a fim de que os parâmetros usados para detectar ou não uma anomalia sejam dinâmicos e flutuem conforme dados da Internet.

Referências

Assunção, R. M., Carvalho, O. S., Prates, M. O., and Campos, M. A. (2016). Detecção de anomalias nos pagamentos do SUS. *J. health inform*, pages 459–468.

Bochenek, A. C. and Pereira, J. L. (2018). Revista Jurídica do Ministério Público do Paraná - Corrupção sistêmica no Brasil.

Brasil, O. K. (2021). Operação serenata de amor. https://serenata.ai/.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104.

da Silva Eleutério, P. M. and Machado, M. P. (2011). *Desvendando a computação forense*. Novatec, São Paulo, 6 edition.

de Castro, L. N. and Ferrari, D. G. (2016). *Introcução à mineração de dados*. Editora Saraiva, São Paulo, 1 edition.

e Doraliza Monteiro e Anderson Reis, A. S. (2020). Qualidade da informação dos dados governamentais abertos: análise do portal de dados abertos brasileiro. *Revista Gestão em Análise*, 9(1).

Han, J. and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier, 3 edition.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, pages 85–126.

IPC (2020). Índice de percepção da corrupção. https://cutt.ly/WQUSLbW.

Kintopp, P. M. (2017). Aplicação de técnicas de aprendizado de máquina em dados públicos para detecção de anomalias. B.S. thesis, Universidade Tecnológica Federal do Paraná.

Lopes, M. A. (2019). Aplicação de aprendizado de máquina na detecção de fraudes públicas. B.S. thesis, Universidade de São Paulo.

Lubambo, S. W. (2008). Processo de mineração de dados como apoio à decisão no controle de gastos públicos. B.S. thesis, Universidade Federal de Pernambuco.

Pansani, E. A. and Ferneda, E. Dados governamentais abertos: uma análise da qualidade dos dados em portais de transparência brasileiros. In *Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação*, pages 5023–5046, Londrina, PR, Brasil.

Planalto (2016). Decreto nº 8777 de 11 de maio de 2016. https://cutt.ly/MvG9rOD.

Santos, R., Nunes, F., Oliveira, M., and Júnior, M. (2017). Um survey sobre a utilização de técnicas de data mining e data analytics por agências de investigação criminal do brasil. In *Simpósio Brasileiro de Sistemas de Informação*, pages 593–600. SBC.

Silva, A. G. G. d. (2020). Detecção de anomalias em um serviço de operadora de telefonia móvel. B.S. thesis, Universidade Federal Fluminense.

Warde, W. (2018). *O espetáculo da corrupção: como um sistema corrupto e o modo de combatê-lo estão destruindo o país*. Leya, Rio de Janeiro, 1 edition.