Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Robson Duarte Xavier

Análise de informações das despesas com combustível das entidades públicas municipais do Paraná

Curitiba
2019

Robson Duarte Xavier

Análise de informações das despesas com combustível das entidades públicas municipais do Paraná

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. José Luiz Padilha da Silva

Curitiba
2019

Análise de informações das despesas com combustível das entidades públicas municipais do Paraná

Information analysis of fuel expenses by municipal public entities of Paraná State

Robson Duarte Xavier[1]

[1]Especialização em Data Science & Big Data, Universidade Federal do Paraná, Curitiba, PR, Brasil *

Resumo

Dentre os muitos dados abertos disponíveis, este trabalho tem como tema o uso e a respectiva análise de informações das despesas com combustível declaradas pelas entidades públicas do estado do Paraná (e.g. Prefeituras e Câmaras). Estima-se despesas de R$ 380 milhões anuais e um milhão de registros de dados ao ano. São aplicadas técnicas estatísticas e de aprendizagem de máquina para identificar transações incomuns. Assim, são apresentadas e utilizadas Lei de Benford, decomposição de série temporal via STL e o algoritmo de detecção de valores discrepantes *Robust Random Cut Forest* (RRCF), combinando os *scores* com *Borda count*. Os resultados obtidos foram estatísticas descritivas e exploratórias de interesse, uso dos resíduos dos métodos aplicados para classificação e encontrar o limiar – *threshold* – e lista(s) de anomalias para investigação aprofundada (*ranking*).

Palavras-chave: dados públicos, detecção de anomalias, lei de Benford, decomposição de séries temporais, aprendizado de máquina, *robust-random-cut-forest*

Abstract

Among many public data available, this work uses and analysis expenses declared by public institutions of Paraná State (for example, City Halls and Chambers). Expenses are estimated around R$ 380 million per year and one million data records per year. Statistics and machine learning technics are applied to identify unusual transactions. Thus, Benford Law, STL time series decomposition and Robust Random Cut Forest (RRCF) outliers detection algorithm are introduced and applied, combining each scores using Borda Count. As results, were obtained descriptive analysis, data exploration of interest, use of each method's residuals (for ranking and threshold) and finally list(s) of anomalies for more in-depth investigation.

Keywords: public data, detect outliers, Benford's law, temporal series decomposition, machine learning, robust-random-cut-forest

1. Introdução

1.1. Contexto

Com o advento da transparência pública brasileira, consagrada pelas leis de Acesso à Informação e Transparência, muitos dados públicos encontram-se divulgados – os também chamados dados abertos – que preconizam a disponibilização em tempo real, em meios eletrônicos de acesso público, de informações pormenorizadas sobre a execução orçamentária e financeira da União, dos Estados, do Distrito Federal e dos Municípios, para liberação ao pleno conhecimento e acompanhamento da sociedade [1].

Neste contexto, este trabalho tem como temática a análise de informações de despesas com combustíveis, de registros de consumo, de hodômetro e horímetro (respectivamente quilômetros e horas de funcionamento) dos veículos e equipamentos das entidades públicas municipais do Paraná como Prefeituras, Câmaras e Autarquias.

Estas informações são declaradas pelas entidades e de sua responsabilidade, e recebidas mensalmente e republicadas pelo Tribunal de Contas do Paraná como dados abertos para o Controle Social [2][3]. A obtenção desses dados tem por base a despesa realizada e o controle do maquinário e da frota de veículos em funcionamento, objetivando aferir aquisições, o consumo e os estoques de combustíveis [4].

*robsondxavier@gmail.com

Neste cenário, a expressividade dos valores dessas despesas com combustíveis é enfatizada, valores estes estimados em aproximadamente R$ 380 milhões anuais em média (considerando os anos entre 2011 e 2018), contabilizando os dados agregados de 827 entidades públicas municipais que disponibilizaram as informações.

1.2. Problema

Analisar, descrever, expor visualmente e procurar anomalias é uma tarefa árdua e complexa, principalmente considerando a miríade de dados, recursos limitados e a dificuldade nos procedimentos de exame de informações públicas, nas atividades de controle externo e do próprio controle social das entidades públicas. Menciona-se que a detecção de anomalias é o problema de encontrar padrões nos dados que apresentam desconformidade com um comportamento ou modelo esperado.

O quadro a seguir resume a quantidade de registros de liquidações (despesas), consumos, hodômetros, declarados ano a ano, com notável taxa de crescimento.

Ano	Liquidações	Consumos	Hodômetros
2011	212.665	274.446	320.837
2012	257.242	353.944	486.156
2013	357.812	338.145	540.914
2014	375.083	369.211	597.843
2015	385.686	380.147	663.245
2016	439.019	390.111	721.705
2017	467.657	387.645	773.324
2018	487.178	412.005	834.558

É necessário explicar as informações disponibilizadas pelas entidades. Por exemplo, demonstrando as suas propriedades de tendência, sazonalidade e anomalias no tempo. Essas características são atualmente desconhecidas, já que todos trabalhos anteriormente desenvolvidos foram executados em situações esporádicas ou específicas, de forma manual, relatórios pontuais ou em poucas amostras *adhoc*.

Ademais, inexiste um exame aplicado nesses dados para identificar indícios para posterior inspeção minuciosa com indicação de grau de risco, dificultando o planejamento e a alocação de esforços de trabalho.

Logo, este presente trabalho apresenta a seguinte pergunta de pesquisa aplicada:

> É possível sistematizar a identificação de anomalias nos dados disponibilizados das despesas com combustível das entidades públicas municipais do Paraná?

1.3. Motivação

Desta forma, se motiva a conceber um conjunto de análises, demonstrações, estatísticas descritivas, mineração de dados e visualizações acerca dessas informações do tema, para subsidiar os procedimentos de análise de dados, que respondam a pergunta deste trabalho indicando quais os casos atípicos, propiciando posteriormente uma investigação aprofundada a partir desses indícios.

1.4. Objetivos

1.4.1. Objetivo Geral

Aplicar técnicas da estatística e aprendizado de máquina para identificar transações não usuais nos registros de informação de despesa, quantidades adquiridas, consumo, quilometragem e horas de uso dos veículos das entidades públicas municipais do Paraná.

1.4.2. Objetivos Específicos

- Avaliar e investigar a composição dos dados para consolidá-los (agregação dos dados);
- Definir um ou mais métodos que identifiquem dados anômalos para o contexto;
- Aplicar algoritmo(s) de identificação de dados anômalos que identifique grau de risco ou um *score*;
- Expor graficamente os dados;
- Identificar o grau de risco de cada caso conforme técnicas utilizadas, sendo um caso a estratificação por entidade declarante do dado, ano, mês e categoria de combustível adquirido.

1.5. Hipótese

Assim sendo, este presente trabalho apresenta a seguinte hipótese:

> A aplicação dos métodos de aprendizado de máquina e de estatística auxiliam na identificação de anomalias nos dados disponíveis de despesas com combustível.

1.6. Contribuições e Delimitações

Durante a execução deste trabalho, foram investigadas e eleitas três técnicas para aplicação ao problema, a saber: Lei de Benford [5], *Seasonal-Trend Decomposition Procedure Based on Loess* (STL) [6] e *Robust Random Cut Forest* (RRCF) [7]. O propósito foi incluir técnicas, conjuntamente, que provessem estatísticas descritivas e perfil dos dados, que propiciassem a análise levando em consideração os dados conforme gerados no tempo e também ao menos um método multivariado e de aprendizado de máquina. Com a aplicação dessas técnicas eleitas é possível produzir uma lista de casos para um posterior exame aprofundado. Essas técnicas serão detalhadas na seção Estado da Arte.

Isso posto, como delimitações deste trabalho, se registra que a pesquisa e aplicação de técnicas não foram exaustivas. Ainda, acerca das variáveis captadas nos dados abertos disponíveis, é importante registrar que outras variáveis que poderiam explicar os fenômenos não foram angariadas, não estão disponíveis, são de difícil captação ou extrapolam a execução e prazo deste trabalho. Exemplificando: geografia e malha viária dos municípios; mudanças como da legislação (em saúde, educação e no transporte escolar), de políticas de trabalho, de gestor, de tipo de contratação das entidades (e.g. contrato global para controle da frota); negociações de compra (e.g. descontos, abatimentos); controle e uso de estoque dos combustíveis; contratação e treinamento de motoristas; fatores sazonais como feriados, férias; ou atípicos como greves e mutirões. Esses tipos de dados são custosos para detectar e incluir nas análises, entretanto podem, em certa medida, serem identificados na análise exploratória ou como anomalias.

1.7. Organização do documento

Este trabalho está dividido da seguinte forma: na Seção 2 são expostas as técnicas utilizadas. Logo após, é explicado o processo de construção deste trabalho, no qual se aplicam as técnicas eleitas, na seção 3. A partir disto, são expostos os resultados obtidos na seção 4. Por fim, a seção 5 encerra o trabalho e sugere temas de continuidade do mesmo.

2. Estado da técnica e da arte

A literatura e o ferramental que está a disposição sobre o problema de detecção de anomalias em dados é vasta. Pesquisas de visão geral e aprofundadas, *surveys*, classificação de *outliers*, descrições e taxinomias das técnicas empregadas sobre o tema podem ser encontradas nos trabalhos [8], [9], [10] e [11] e suas correlatas referências.

A detecção de anomalia é a averiguação de padrões nos dados procurando divergências em relação ao padrão [10]. De forma complementar, o dado anormal resulta no indicativo em que o dado foi gerado por um mecanismo diferente, sendo o desvio entre o resultado do mecanismo padrão (da expectativa ou do dado esperado) e o dado real como uma das principais formas de identificar uma anomalia [8][9].

2.1. Decomposição de séries temporais com STL

A ideia central em uma série temporal é que é possível isolar padrões de regularidade nos dados, sejam eles tendência e sazonalidade, considerando dados que representam fenômenos sequenciais e correlacionados no tempo. Tudo aquilo que o modelo temporal não explica via padrões encontrados (e.g. decomposição) é resíduo, por consequência possibilitando a identificação de irregularidades quando os resíduos são suficientemente grandes [6], [12] e [13].

O método para decomposição de séries temporais utilizado neste trabalho é o STL (*Seasonal-Trend decomposition using Loess*), que em tradução livre significa decomposição de sazonalidade e tendência utilizando *Loess* (este sendo um método de suavização estatístico). STL é um método estatístico não-paramétrico, e propicia configurações de período, tendência e sazonalidade, podendo ser robusto a *outliers*, característica primordial a este trabalho (robustez proporciona um modelo menos sensível a dados anômalos, incrementando os resíduos da decomposição). É um método para séries temporais univariadas [6].

A Figura 1 expõe a decomposição utilizando STL para todos os dados captados das despesas de 2011 a 2018 das entidades públicas municipais do Paraná. De cima para baixo, os retângulos dessa figura mostram os valores informados mensalmente, a sazonalidade, a tendência e por fim os resíduos da decomposição.

2.2. Lei de Benford

A distribuição de Benford é uma distribuição que trata dos primeiros dígitos significativos dos números. De forma simples, o primeiro dígito significativo de um número é o dígito mais à esquerda em sua representação decimal, seguido do segundo dígito significativo e assim sucessivamente [5].

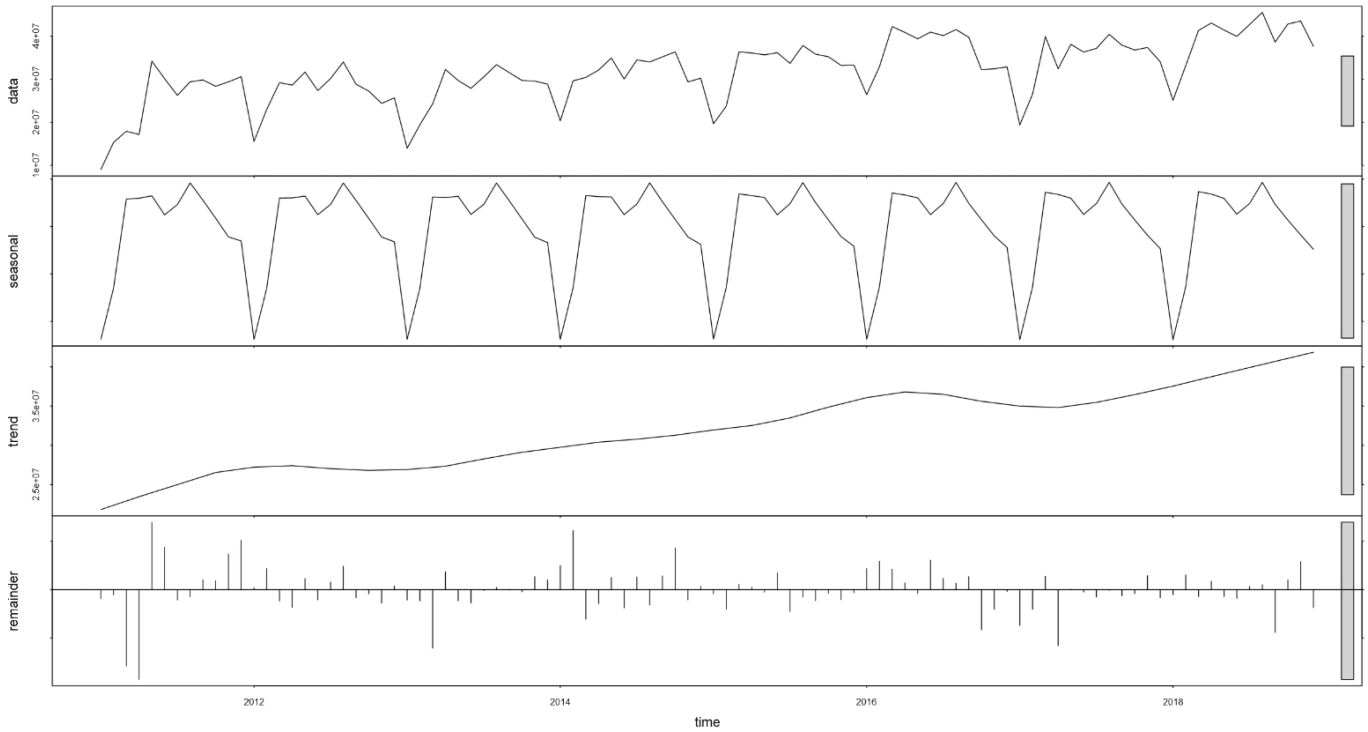

Figura 1: STL - Decomposição dos valores de liquidação de 2011 a 2018. São encontrados características sazonais típicas e tendência crescente durante os anos.

A Lei de Benford define a distribuição, sendo d os n primeiros dígitos significativos, como:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{1}$$

Assim, quanto maior o numeral presente nos dígitos iniciais, menor a probabilidade na distribuição, identificando um modelo univariado baseado na frequência dos números.

Tem presença ubíqua na literatura e no uso prático, sendo utilizada em testes de conformidade com dados amostrais, igualmente na suspeita e detecção de fraudes e anomalias, erros de digitação, duplicação de lançamentos, valores monetários, sendo relativamente comum em auditorias contábeis [14][15].

Sua principal proposta é detectar quando um conjunto de dados desvia da distribuição de Benford indicando os dados suspeitos que necessitam verificação [14][15].

A Figura 2 mostra a distribuição de todos dígitos significativos para todas as as despesas dos dados captados das entidades públicas municipais do Paraná, entre 2011 e 2018. No eixo y está a frequência acumulada números nos dados, no eixo x os numerais ou dígitos propriamente ditos (análogos a categorias). A linha representa a distribuição teórica de Benford.

2.3. Random Robust Cut Forest

O método *Random Robust Cut Forest* [7] – RRCF – é um algoritmo de *machine learning* não supervisionado, multivariado e não paramétrico do tipo *ensemble* (classe de algoritmos que se utilizam de um conjunto de vários modelos básicos combinados buscando resultados ótimos).

É oriundo do desenvolvimento das técnicas de árvores, *Random Forest* e *Isolation Forest* [16] (este último também um algoritmo cujo objetivo é detecção de *outliers*), e foi projetado tanto para processamento para dados em *batch* (lote) quanto em *streaming* (fluxo).

A principal diferença deste tipo de algoritmo é que ele busca explicitamente isolar anomalias, ao invés de construir regiões de conjuntos de dados próximos entre si (como na clusterização). O algoritmo é adequado para alta dimensionalidade de dados, robusto a dimensões de dados irrelevantes e os autores demonstraram melhores comparativos em relação a outros algoritmos e bons resultados empiricamente.

Em primeiro lugar, o RRCF produz uma floresta composta de árvores de busca binária, sendo a largura de um ponto particular na árvore seu equivalente em largura em *bits* (número de *bits* necessários para armazenar o ponto). A complexidade do modelo pode ser

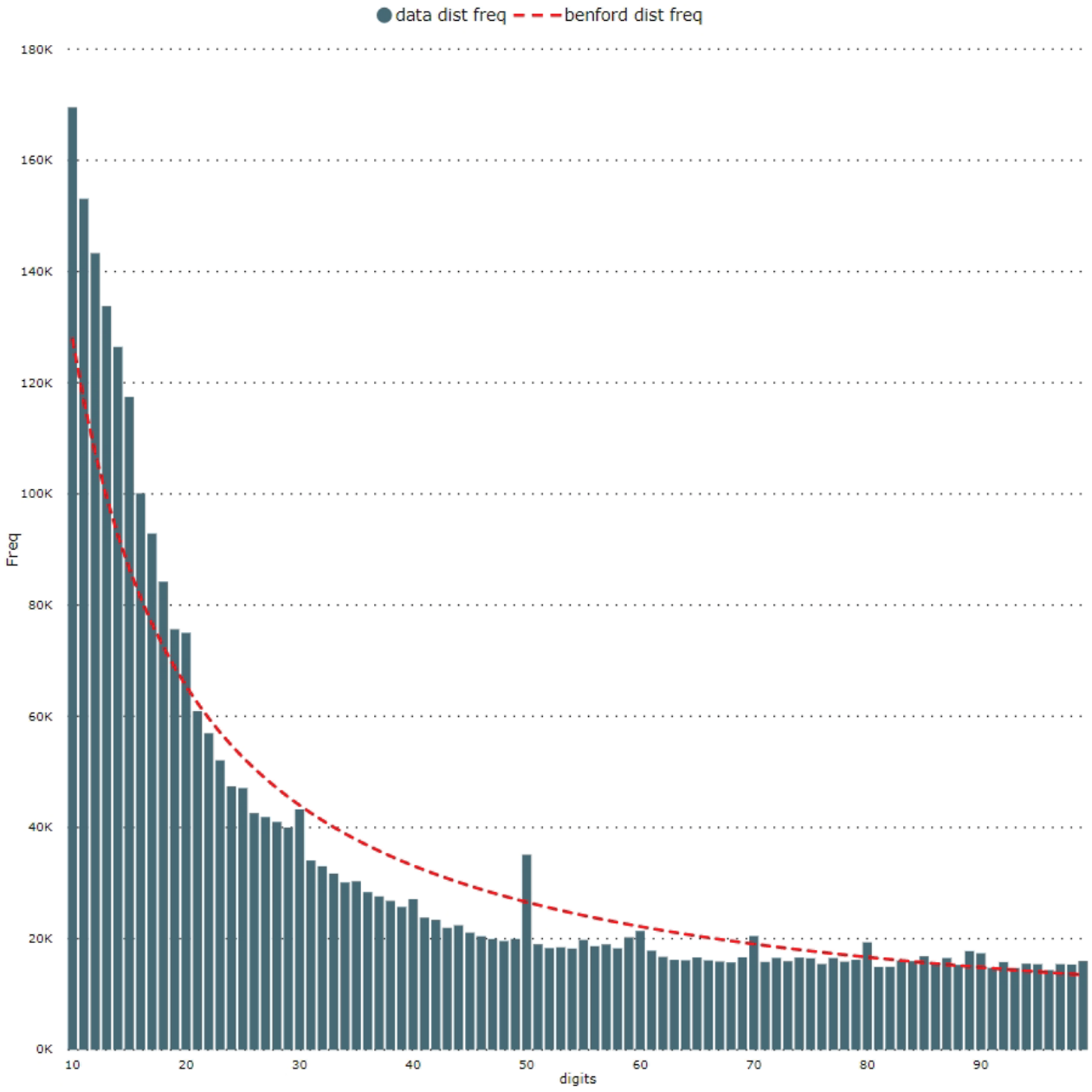

Figura 2: Valores de liquidação de 2011 a 2018 (frequência dos dois primeiros dígitos) versus Benford. Quanto maior o numeral presente nos dígitos iniciais, menor a probabilidade na distribuição.

representada pela soma dos *bits* de largura de todos os pontos da árvore. Dentro deste contexto, um *outlier* é definido como um ponto que significantemente aumenta a complexidade da árvore (i.e. largura em *bits*), caso esse ponto seja adicionado ou removido.

Desse modo, a métrica resultante do algoritmo é o deslocamento colusivo – *collusive displacement* – (CoDisp), que os autores relatam embasamento estatístico, calculando-a para cada amostra em relação a complexidade da mesma no modelo. Para tanto e de modo sumário, calcula-se a média do máximo deslocamento ao excluir determinado ponto e seus vizinhos dividindo pelo número de pontos excluídos (sendo o deslocamento esperado causado por um ponto é o número de pontos no nó irmão da folha contendo o próprio ponto). Essa métrica estende a noção de deslocamento por contabilizar e tratar os dados duplicados e os dados próximos dos duplicados que podem mascarar a presença de anomalias (*masking effect* [8][10]). *Outliers* equivalem a um grande CoDisp [7].

2.4. Técnicas preteridas

Durante a investigação das técnicas, cabe aqui documentar as que foram preteridas.

A primeira delas foi a clusterização por aglomeração [17] [18] (e.g. utilizando *single linkage*). Os ensaios iniciais demonstraram que seria necessário investir muito tempo e esforço técnico-computacional para calcular a matriz de distâncias e na composição do método aglomerativo, particularmente pela impossibilidade de diminuir a dimensionalidade de dados monetários declarados pelas entidades (e.g. 3.089.188 despesas registradas de 2011 a 2018, ou 214.580 despesas agregadas em temporalidade mensal no mesmo período).

Em seguida, o método de série temporal SARIMA [12] implicou que que cada estratificação tivesse o seu próprio modelo auto-regressivo e de médias móveis, neste sentido as suposições ou seriam muito genéricas para todas as entidades ou que cada combinação de 827 entidades versus uma categoria de combustível teriam o seu próprio e diferente modelo, o que dificultaria a comparação entre os estratos calculando um grau de risco. Por outro lado, o método de decomposição STL providencia um modo relativamente simples para realizar a generalização das suposições de frequência e período da série temporal dos dados identificando *outliers* [8][12].

E a terceira e última técnica preterida foi a aplicação de Rede Neural Auto-Associativa e Algoritmos Genéticos [13]. A técnica não foi eleita pois os dados não possuem rótulo e pela dificuldade de gerar dados fictícios que sejam *outliers* sem um histórico de anomalias já identificadas.

3. Metodologia e Materiais

Nesta seção descreveremos os procedimentos realizados para a construção da sistematização da análise e busca de anomalias nos dados públicos de despesas municipais com combustível.

A princípio, compete descrever os passos inerentes e transversais (i.e. em todas fases ou etapas), executados sempre que necessários, ao processo de desenvolvimento deste trabalho.

Foram utilizadas diferentes composições e combinações, agregações e desagregações dos dados; análises, estatísticas exploratórias e diversos gráficos; pesquisa e revisão da literatura em busca de alternativas; aprofundamento nos conhecimentos necessários do contexto e do tema para entendimento do problema. Esses passos podem ser usuais em todo trabalho analítico ou aplicado de ciência de dados.

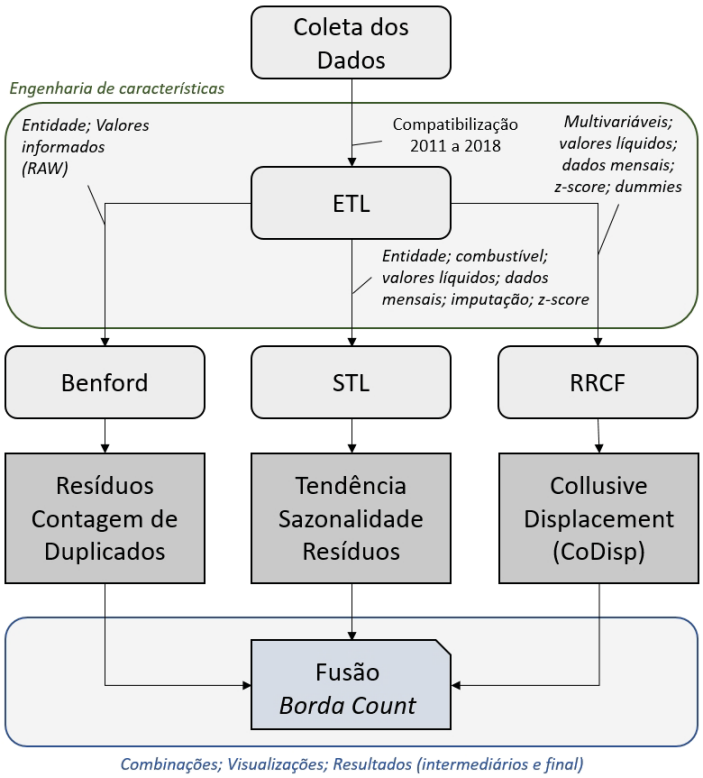

Figura 3: Fluxograma da metodologia empregada.

Em seguida, são elencados sucintamente todos os passos da metodologia do trabalho: coleta de dados; pré-processamento dos dados (e.g. ETL) e engenharia de características; aplicação de STL, Benford e RRCF; resultados de cada modelo (intermediários na metodologia); combinação, visualização e análise desses resultados intermediários; fusão dos resultados via *Borda Count*; visualização e análise dos resultados finais.

O diagrama exposto na Figura 3 demonstra o fluxo da metodologia aplicada neste trabalho. Nele, se enuncia a separação em três partes para cada técnica e seus respectivos resultados no encadeamento (caixas cinza escuro), culminando na fusão de Borda. Também se mostram a primeira área realçada (borda verde) contendo as divisões de informações e alguns detalhes anteriores para cada técnica, e a última área (borda azul) de ênfase nas visualizações e análise de cada técnica (intermediário), bem como do resultado combinado (final).

Para a coleta dos dados foram extraídas as bases de dados de Controle Social, Gastos Municipais e de *Download* de Dados, incluindo as bases dos dados do período de 2011 até 2012 [2] e as bases de dados de 2013 até os dias presentes [3]. Em seguida, foi pro-

duzida a união de ambas as bases dos dois períodos distintos, a limpeza dos dados (tipagem, retirada de caracteres especiais, transformação em modelo tabular), compatibilização dos valores numéricos, padronização das categorias de combustível.

Como delineamento, foram angariados todos os dados que ocorreram até o final de 2018, com dados informado até o período de junho de 2019. Esta linha de corte foi necessária para obter uma base de trabalho estática e sem atualizações, já que as entidades podem continuar informando dados a posteriori (há entidades que demonstraram atraso no envio das informações).

As principais variáveis angariadas são listadas na tabelas 1 para as despesas registradas, na tabela 2 para os registros de consumo dos veículos por mês e na tabela 3 para os registros de hodômetro e horímetro. A variável que dá nome para as Entidades foi anonimizada. Registra-se as categorias genéricas e não usuais encontradas: Outros combustíveis e Veículos Flex e Assemelhados (não delimitam uma categoria de fato).

Estão elencadas aqui somente as variáveis principais, sendo suprimidas variáveis acessórias como as categorias das entidades, categorias de unidades de medida, estornos dos registros (acréscimos, alterações ou exclusões dos valores), placas dos veículos, identificação dos equipamentos, códigos sequenciais de transação, outras variáveis de contagem, textuais ou descritivas.

Em seguida, as variáveis foram trabalhadas para compor os conjuntos de dados e/ou variáveis de trabalho. O primeiro conjunto é o conjunto de dados tal como informados (brutos) pelas entidades (no entanto limpos, tabulados, adequados e compatibilizados), primordialmente para uso no modelo de Benford (e.g. variável bruta de despesa na tabela 1).

O segundo conjunto de dados de trabalho foi composto do cálculo dos valores líquidos (contabilmente) das despesas, considerando os eventos de lançamento (liquidação) e os respectivos estornos (e.g. variáveis líquidas na tabela 1). A seguir, foram agregados os dados para compor uma série de dados no tempo (em termos de mês e ano), ou seja, dados em temporalidade mensal. Dessa forma, reduzindo a dimensionalidade dos dados, para cada entidade, mês, ano, categoria de combustível e respectivos valores numéricos (e.g. despesas monetárias e de quantidade). Do mesmo modo, foram criadas duas variáveis de contagem para cada evento contábil que ocorrem em cada mês (despesa e estorno de despesa). Esse segundo conjunto de dados foi empregado nas técnicas de STL e RRCF.

Ademais, dados que não são informados para uma entidade, mês, ano e categoria de combustível foram completados com zeros (*imputação*). Isto é necessário pois uma entidade que não comprou um categoria de combustível em um dado mês ou ano não precisa informar dados neste caso, e para realizar uma análise ao longo do tempo é necessário que a série esteja completa (STL). A tabela 4 lista estas variáveis.

Nas próximas subseções serão explicados os três caminhos da metodologia referentes a cada técnica empregada e a fusão por consenso dos resultados.

3.1. Aplicação de decomposição temporal STL

Para a aplicação do método estatístico de decomposição temporal STL, o segundo conjunto de trabalho foi dividido, para cada iteração, por entidade e categoria de combustível. Se assume que cada categoria de combustível tem um comportamento próprio e individual, pois cada entidade ou possui em seu patrimônio ou contrata veículos de classes de combustíveis diferentes, com respectivo comportamento inerente a cada grupo.

A variável de interesse foi o valor monetário despendido mensalmente pela entidade (univariado), devidamente padronizado (*z-score*), bem como configurado como parâmetros do método o período anual (12), o componente sazonal (janela) igual a treze (13) para cada iteração de cada estratificação e aplicada a robustez. O parâmetro de janela *periódico* não se mostrou adequado, pois quando aplicado resulta nas médias simples para cada ponto no tempo da série, dessa forma deixando de aplicar a suavização Loess.

Como resultado intermediário, este método forneceu para cada caso (entidade, mês, ano, categoria de combustível) os dados de tendência, sazonalidade, e primordialmente os resíduos.

3.2. Aplicação da Lei de Benford

Já para a aplicação da Lei de Benford, o primeiro conjunto de trabalho foi dividido por entidade, para cada iteração. O foco do uso do método foi nos dados monetários tal como informados pelas entidades (univariado), para cada transação (para cada registro único de despesa de cada entidade). Para este método, isto é primordial, pois o seu uso preconiza que os valores comparados com a distribuição de Benford serão os dados amostrados originais.

Para isto, foram utilizados os dois primeiros dígitos significativos. Com um único número significa-

Nome da Variável	Descrição	Tipo/Formato	Exemplo
Entidade	Nome da Entidade anonimizado	Texto	Entidade 15789487
Data da Liquidação	Dia em que ocorreu a despesa	Data formato DD/MM/AAAA	03/01/2018
Valor liquidado (bruto)	Valor informado da despesa em reais	Número com duas casas decimais (R$)	100,00
Valor liquidado (líquido)	Valor calculado contabilmente da despesa considerando estornos em reais	Número com duas casas decimais (R$)	94,00
Quantidade (bruto)	Quantidade adquirida bruta em litros ou m³ adquiridos	Número com três casas decimais	25,500
Quantidade (líquido)	Quantidade calculada líquida em litros ou m³ adquiridos considerando estornos	Número com três casas decimais	23,500
Combustível	Categorias de combustível adquirida	Biodiesel Diesel Etanol Gasolina GNV Outros combustíveis Querosene Veículos Flex e Assemelhados	Gasolina
Preço unitário	Preço calculado (valor liquidado líquido / quantidade líquido)	Número com duas casas decimais em reais (R$)	4,00

Tabela 1: Variáveis de despesa com combustível. Ocorrem *n* registros em uma data (dia/mês/ano).

Nome da Variável	Descrição	Tipo/Formato	Exemplo(s)
Veículo	Um automóvel ou equipamento	Texto	Ônibus Mercedez-Bens Roçadeira
Quantidade	Consumo em litros ou m³.	Número com três casas decimais.	30,500
Quantidade calculada	Consumo em litros ou m³ considerando estornos	Número com três casas decimais.	35,500
Combustível	Categorias de combustível consumida	Biodiesel Diesel Etanol Gasolina GNV Outros combustíveis Querosene Veículos Flex e Assemelhados	Gasolina
Mês e ano	Registro de consumo em determinado mês e ano	Mês específico formato MM/AAAA	05/2018

Tabela 2: Variáveis dos registro de consumo. Ocorrem *n* registros em um mês (mês/ano).

tivo (somente de zero a nove), a análise poderia ficar muito genérica. Com três dígitos significativos, os valores abaixo de R$ 100,00 seriam desprezados, sendo esses valores muito comuns nos dados. Foi utilizada somente a parte inteira do valor monetário.

Como resultado intermediário, o método gerou os dados de resíduos (entre a distribuição esperada e cada amostra dos dados) baseada na frequência dos dois primeiros dígitos e também a contagem de duplicados (a contagem dos dígitos frequentes pressupõe a contagem dos duplicados).

Este método não considera diretamente as amostras distribuídas no tempo, contudo são identificados corretamente com as amostras referentes ao período em que ocorreram.

Nome da Variável	Descrição	Tipo/Formato	Exemplo
Veículo	Um automóvel ou equipamento	Texto	Ônibus Mercedez-Bens Roçadeira
Hodômetro ou horímetro inicial (1)	Início do contador de quilometragem ou horas de trabalho	Número com três casas decimais (km ou h)	10.000,000
Hodômetro ou horímetro final (2)	Final do contador de quilometragem ou horas de trabalho	Número com três casas decimais (km ou h)	11.000,000
Quilometragem ou horas calculadas	Hodômetro ou horímetro final menos inicial (2-1)	Número com três casas decimais (km ou h)	1.000,000
Tipo de medidor	Hodômetro ou horímetro instalado no veículo/equipamento	Hodômetro Horímetro	Hodômetro
Mês e ano	Registro de hodômetro ou horímetro em determinado mês e ano	Mês específico formato MM/AAAA	05/2018

Tabela 3: Variáveis dos registros hodômetro e horímetro. Ocorre um único registro em um mês específico (mês/ano) por veículo/equipamento.

Nome da Variável	Descrição	Tipo/Formato	Exemplo
Entidade	Nome da Entidade anonimizado	Texto	Entidade 15789487
Mês e ano	Registro de despesas em determinado mês e ano	Mês específico formato MM/AAAA	05/2018
Valor liquidado (líquido) mensal	Valor acumulado da despesa mensal em reais	Número com duas casas decimais (R$)	1.000.000,00
Quantidade (líquido) mensal	Quantidade acumulada líquida em litros ou m³ mensal	Número com três casas decimais	255.000,000
Combustível	Categorias de combustível adquirida	Biodiesel Diesel Etanol Gasolina GNV Outros combustíveis Querosene Veículos Flex e Assemelhados	Gasolina
Preço unitário (médio)	Preço mensal médio (valor liquidado líquido / quantidade líquido) em reais	Número com duas casas decimais (R$)	3,92

Tabela 4: Variáveis de despesa com combustível em temporalidade mensal (agregação dos dados em mês e ano).

3.3. Aplicação do algoritmo RRCF

Dando sequência, ao aplicar o método RRCF o segundo conjunto de trabalho foi processado integralmente pelo algoritmo (multivariado). A preparação específica acrescentou a padronização *z-score* recomendada pela literatura e a transformação da variável categórica combustível (variáveis *dummy*).

Após alguns ensaios e testes, o algoritmo foi configurado com os parâmetros de 1.000 árvores para a floresta, 1.000 nós de tamanho das árvores e de 173 a 1.000 o número de amostras sem repetição para montar cada árvore. Ensaios com menor número de árvores e tamanho tenderam não explicitar adequadamente dados diferentes. De mesmo modo, conforme recomenda a literatura, um número maior de árvores e tamanho propicia melhor identificação de dados atípicos, em contrapartida a um desempenho menor em termos de tempo de execução.

Como resultado intermediário do processo, este método forneceu a métrica *CoDisp* para cada caso (entidade, mês, ano, categoria de combustível). Este método também não considera diretamente as amostras distribuídas no tempo, contudo seus resultados são identificados ao período em que ocorreram.

3.4. Fusão dos *scores* via *Borda Count*

Muito embora os resultados intermediários de cada um dos métodos sejam muito úteis para a análise e resposta desse trabalho, os seus respectivos resultados também foram fundidos utilizando o método de Borda [19].

Borda *count* é um método bastante usado para fusão de resultados de algoritmos quando os mesmos fornecem um lista ordenada. É um método simples, tendo como procedimento somar a pontuação em *ranking* de cada algoritmo, resultando assim em um *ranking* unificado (também chamado de votação por consenso).

Ao resultado de cada método ou algoritmo foi atribuído uma pontuação ordenada de forma crescente conforme cada caso (entidade, ano, mês e categoria de combustível) e seu respectivo *score* (i.e. *ranking*). Depois essa pontuação foi somada, decorrendo que quanto menor a contagem de Borda, mais alto no *ranking* gerado pela fusão. Ou seja, quanto menor o número resultante, mais alto o risco e maior o consenso entre as três técnicas.

3.5. Finalização: visualização, análises e recursos computacionais

Tanto a partir dos resultados intermediários na metodologia, bem como dos resultados da fusão por Borda *count*, foram combinados e produzidos visualizações em gráficos e tabelas. Após a aplicação de cada método foi necessário compatibilizar corretamente os resultados (e.g. resíduos ou *score*) com o período em que se referem.

Os principais recursos computacionais utilizados foram:

- R [20]
  - Versão 3.5.1;
  - Pacotes `benford.analysis` e `stl`;
  - IDE RStudio Desktop 1.2.1335.
- Python [21]
  - Versão 3.6.5;
  - Biblioteca `rrcf`;
  - IDE Jupyter Notebook 5.5.0.

Na próxima seção são analisados, demonstrados e argumentados os resultados obtidos.

4. Análise e Resultados

Preliminarmente, compete expor que as técnicas aplicadas resultaram em mais de uma maneira de sistematizar e identificar anomalias na análise dos dados das despesas, cada qual com suas características. Resultando em listas e *rankings* de uma série de informações que podem ser inspecionadas (i.e. possíveis anomalias ou indícios), verificando assim a empregabilidade da metodologia adotada neste trabalho.

Em seguida, serão descritas a análise e resultados obtidos, de forma pormenorizada para cada técnica dentro da metodologia empregada, sobre a descrição dos dados na primeira subseção 4, logo depois os resultados intermediários para STL no tópico 4.2.1, Benford em 4.2.2 e após RRCF em 4.2.3, findando com a fusão em Borda na subseção 4.3. As análises e resultados mostram uma visão global de todo o período de 2011 a 2018, e são explicitadas pontualmente para o ano de 2018.

4.1. Análise exploratória e estatística descritiva

A análise exploratória e estatística descritiva foram de suma importância para realização deste trabalho. Somente com sua aplicação que foi possível esquadrinhar os dados para tomada decisão nas abordagens aplicadas neste projeto.

Para começar, a seguir se apresenta a volumetria final dos dados de despesas, considerando a linha de corte de dezembro/2018 e a redução de dimensionalidade.

- 3.089.188 dados de despesas (registro a registro);
- 214.580 registros agregados mês a mês das despesas (entidade, ano, mês e categoria de combustível);
- 28.409 meses completados com zero (imputação dos meses não informados pelas entidades);
- 827 entidades;
- 96 meses, de janeiro de 2011 a dezembro de 2018.

Na tabela 5 são listadas as estatísticas descritivas considerando todos os registros de despesas. Neste momento, se frisa o valor de moda R$100,00 e a quantidade máxima de mais de 68 milhões de litros em um único lançamento (provável erro na informação).

Já na tabela 6 são apresentadas as estatísticas descritivas considerando os valores agregados mensalmente, para a categoria de combustível Diesel no ano de 2018, totalizando gastos na ordem de R$ 332.119.745,92. Esta

Descrição	Valor Liquidado (R$)	Quantidade (l)
Mínimo	-14,60	-47.830,000
Primeiro Quartil	120,20	38,000
Mediana	221,40	75,000
Média	1.036,60	622,000
Moda	100,00	40,000
Terceiro Quartil	635,80	215,000
Máximo	413.851,80	68.198.310,000

Tabela 5: Sumário dos dados das despesas considerando as 3.089.188 transações de 2011 a 2018.

categoria representa o maior gasto de combustível nos dados, totalizando R$ 2.168.602.006,27 em todo período (2011 a 2018). Pode-se perceber valores destoantes para o mínimo e máximo do preço unitário (R$ 0,05 e R$ 15,69 respectivamente) e também para o número máximo de estornos (59) em relação as respectivas métricas relacionadas.

Nesta circunstância, se analisa resumidamente o perfil do dados indicando as medidas de centralidade (média, mediana, moda) para referencial com uma análise dos valores atípicos encontrados nas técnicas utilizadas (STL, Benford e RRCF). A mediana, sendo mais robusta aos *outliers* e a moda para contagem dos eventos mais recorrentes são especialmente úteis. Enfatiza-se as diferenças entre os valores de mediana, média e moda (tanto conceitualmente quanto nas diferenças numéricas entre si), a qual propiciam cada qual análises diferentes.

Ademais, são elencados também alguns resultados de interesse, como os quantitativos de valores negativos (possível ajuste na despesa agregada mensal), valores zerados (liquidações sem valor ou que foram totalmente estornadas), valores irrisórios (possível litragem com valor incomum, ou também ajuste na despesa mensal) e meses com gastos irrisórios, ou sem despesas ou sem lançamentos.

- 332 liquidações de valor negativo;
- 19.343 liquidações de valor zerado;
- 10.635 liquidações de valor menores que R$ 10,00 dez reais;
- 38.580 liquidações de quantidade menor que 1 litro ou m³;
- 3.365 meses com lançamentos de valor agregado mensal menor ou igual a R$ 100,00 para uma categoria de combustível;
- 108 meses com lançamentos de valor agregado mensal zerado para uma categoria de combustível (efetivamente informados pelas entidades, excluindo os meses imputados).

4.2. Resultados Intermediários

4.2.1. Resultados: STL

O principal resultado na busca de valores atípicos, ao aplicar a decomposição STL, foram os resíduos. A síntese dos resíduos consta na tabela 7, considerando o limiar (*threshold*) de referência. O *threshold* é o valor de limiar para os resíduos acima do 99,5 percentil a partir do número das amostras. É relevante sublinhar que na aplicação da decomposição foi aplicado *z-score* nos valores de despesa, para evitar influência dos altos valores monetários.

A partir dos resíduos é possível se construir *rankings* para cada caso (entidade, mês, ano e categoria de combustível) de interesse na busca por valores atípicos, para todo período ou para períodos determinados (inclusive recalculando o *threshold* caso desejado). A Figura 4 mostra os valores para os maiores casos de 2018, empilhando os valores dos resíduos de todas as categorias do caso/mês.

Além dos resíduos, outros produtos importantes para a análise da STL foram os números e gráficos para os valores de tendência e sazonalidade, para análises do comportamento ao longo do tempo das despesas de uma entidade versus uma determinada categoria de combustível. Um caso de indício atípico, angariado a partir do *ranking* de 2018, é exposto na Figura 5, no qual é possível identificar uma mudança atípica no comportamento de despesas de etanol no período.

Descrição	Valor Liquidado (R$)	Quantidade (l ou m³	Preço unitário médio (R$)	Liquidações	Estornos
Mínimo	31,40	10,020	0,05	1	0
Primeiro Quartil	27.535,80	8.270,100	3,17	8	0
Mediana	50.530,90	15.157,260	3,30	18	0
Média	60.363,50	18.376,300	3,32	40,79	0,3873
Moda	170,00	10.000,000	3,29	1	0
Terceiro Quartil	80.297,60	24.345,690	3,46	42	0
Máximo	654.287,20	268.273,930	15,69	590	59

Tabela 6: Sumário dos dados das despesas, considerando a temporalidade mensal (segundo conjunto de dados) para a categoria de combustível Diesel, para o ano de 2018.

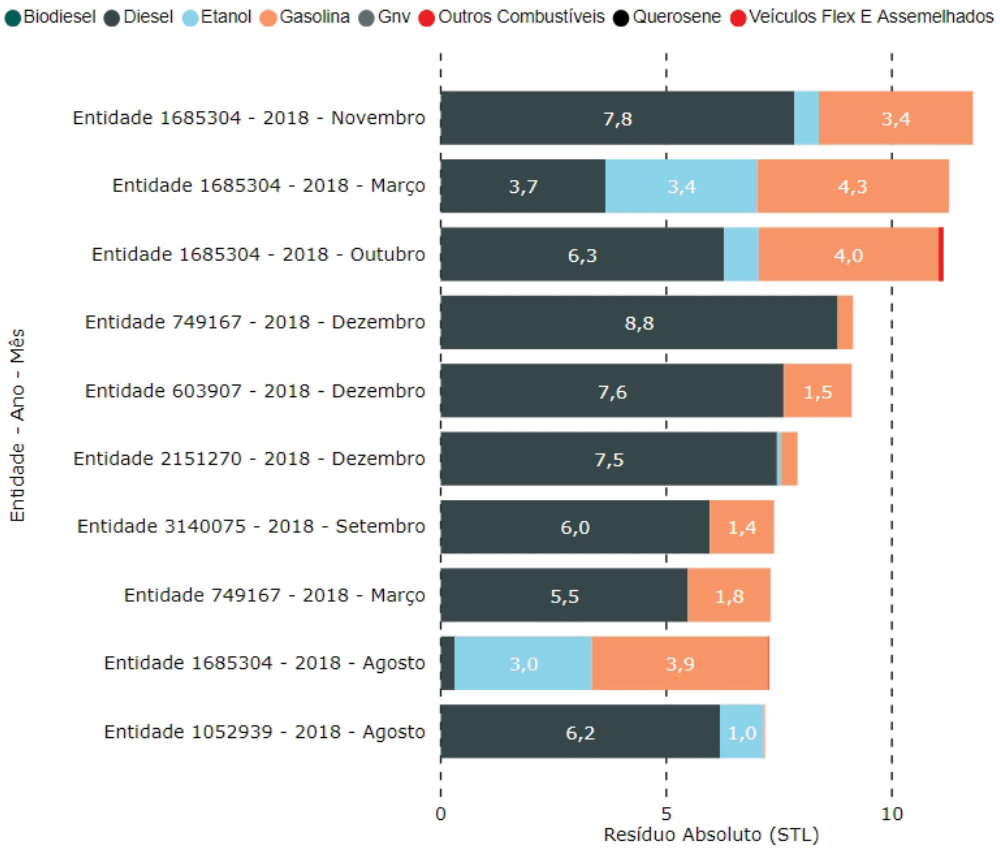

Figura 4: Top 10 casos em 2018 com maiores resíduos absolutos empilhados em STL.

Descrição	Valor
Mínimo	-10,23853
Primeiro Quartil	-0,01278
Mediana	0,0000
Média	0,01534
Terceiro Quartil	0,01892
Limiar (*Threshold*)	**± 0,5863**
Máximo	18,54713

Tabela 7: Sumário dos resíduos para a aplicação de STL em todo período 2011-2018, destaque para o *threshold*.

4.2.2. Resultados: Benford

Dois produtos são os principais resultados ao aplicar o método de comparação da distribuição da amostra com a distribuição de Benford, na busca de valores atípicos. A síntese dos resíduos, o primeiro desses produtos, consta na tabela 8, considerando também um limiar (*threshold*) de referência (linha de corte para resíduos acima do 99,5 percentil a partir do número das amostras).

Da mesma forma, a partir dos resíduos em Benford é possível se construir *rankings* na busca por valores atípicos. A Figura 6 mostra os valores para os maiores

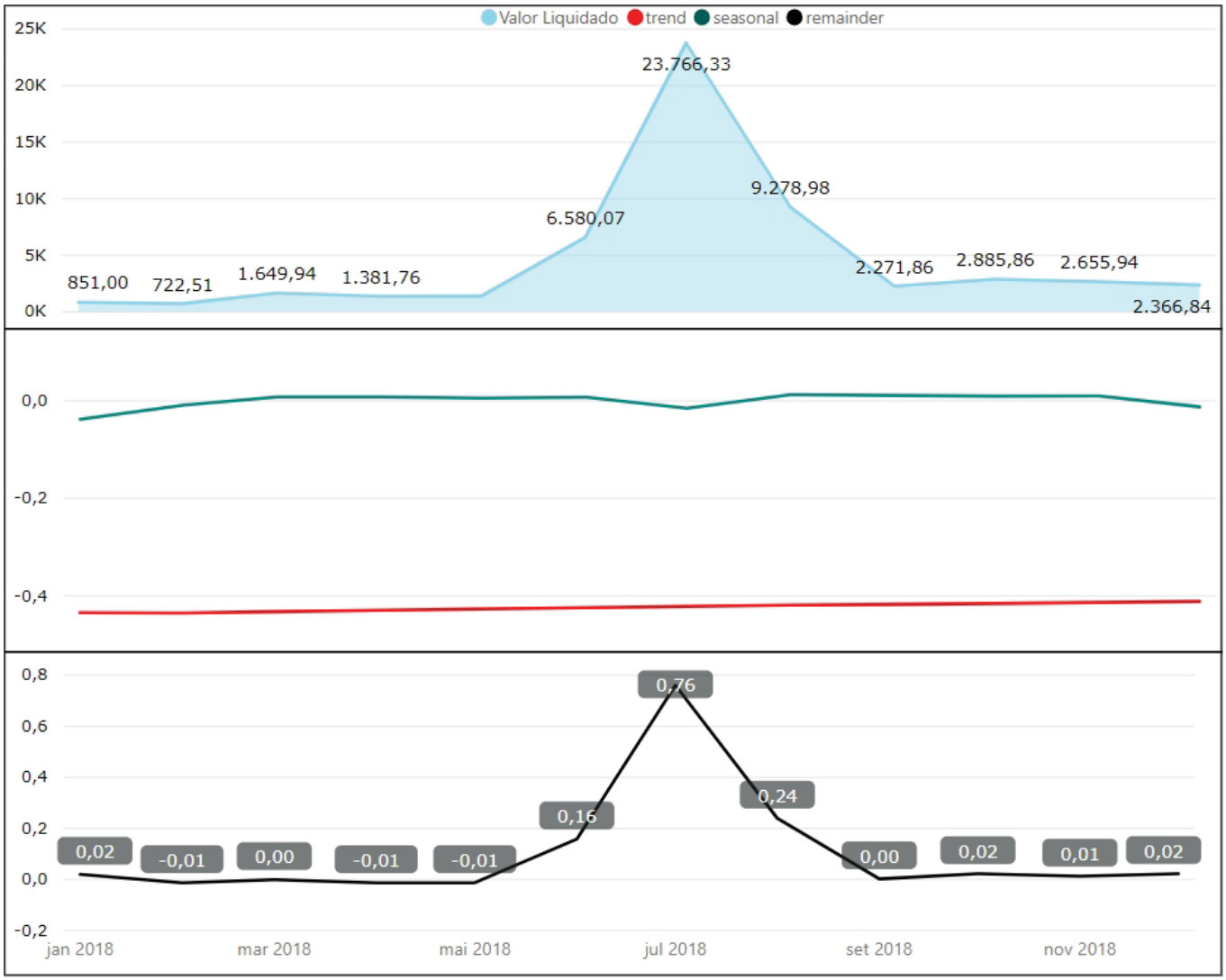

Figura 5: Um indício para avaliação usando método STL. Valor liquidado atípico de R$ 23.766,33 em julho de 2018 com resíduo da decomposição STL de 0,76.

Descrição	Valor
Mínimo	0,00
Primeiro Quartil	22,10
Mediana	117,39
Terceiro Quartil	537,55
Média	2.465,88
Limiar absoluto (*Threshold*)	**11.504,68**
Máximo	479.515,28

Tabela 8: Sumário dos resíduos para a aplicação de Benford (agregado para um mês/ano) em todo período 2011-2018, destaque para o *threshold*.

casos de 2018, empilhando os valores dos resíduos de todas as categorias do caso/mês.

O segundo produto é a listagem dos números duplicados, pois as contagem de frequência das ocorrências dos valores naturalmente resultam nessa contagem. É um resultado muito útil e complementar ao perfil e descrição dos dados. A tabela 9 elenca um *ranking* para valores duplicados. Muitas transações idênticas (duplicadas) podem demonstrar despesas sobrepostas e não efetuadas de fato, indícios que podem ser verificados.

De forma complementar, transações raras e de alto valor podem também indicar despesas que também podem ser averiguadas. Por exemplo, a transação de valor R$ 298.874,99 ocorreu para uma mesma entidade duas vezes. A tabela 10 lista um *ranking* para valores informados e raros, e nenhuma das maiores transações de 2018 figuram como as maiores transações de todo o conjunto de dados.

Um caso para melhor avaliação em Benford, classificado em quarta posição no Top 10 do ranking em

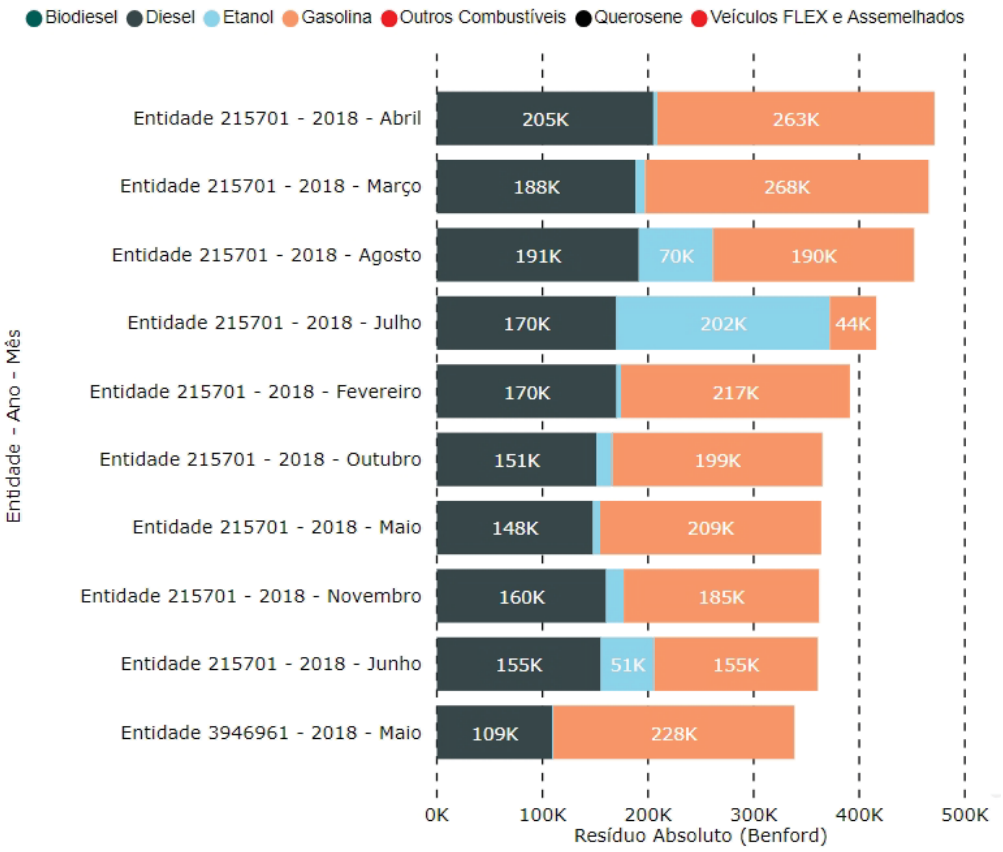

Figura 6: Top 10 casos em 2018 com os maiores resíduos empilhados em Benford.

Valor (R$)	Duplicados	Valor (R$) (2018)	Duplicados (2018)
100,00	14.613	100,00	2.481
50,00	11.482	50,00	1.754
150,00	5.505	150,00	1.269
70,00	4.754	200,00	656
120,00	3.896	70,00	554
80,00	3.438	80,00	475
200,00	3.304	120,00	439
60,00	3.067	140,00	408
30,00	2.984	97,00	377
90,00	2.563	30,00	368

Tabela 9: Tabela com os Top 10 casos de maiores repetições dos valores. Todo o período 2011-2018 e somente o ano de 2018.

Valor (R$))	Ocorrências	Valor (R$) (2018)	Ocorrências (2018)
413.851,83	1	114.240,00	1
359.950,00	1	113.977,26	1
298.874,99	2	103.005,00	1
200.000,00	1	102.970,00	1
199.000,00	1	96.600,00	1
195.000,00	1	96.470,54	1
165.135,38	1	93.951,42	1
162.210,00	1	91.476,00	1
151.350,00	1	89.760,00	1
151.240,00	1	89.448,00	1

Tabela 10: Tabela com os Top 10 casos raros e de alto valor. Todo o período 2011-2018 e somente o ano de 2018.

2018 em Beford, é Entidade 215701 - 2018 - Julho, apresentando alta diferença em relação à distribuição de Benford. Detalhadamente se apresentam aqui as transações duplicadas para o mês do caso em questão na tabela 11.

Uma das possibilidades de investigação é procurar as datas correspondentes de todas transações de mesmo valor, relacionando-as conforme a datas iguais ou datas muito próximas, como exposto na tabela 12.

4.2.3. Resultados: RRCF

Para a pesquisa de valores atípicos, o principal produto resultante ao aplicar o algoritmo RRCF é o *CoDisp*. O resumo do *score* consta na tabela 13, estimando também o limiar (*threshold*) proposto, conforme literatura de referência, considerando o *CoDisp* acima do 99,5 percentil a partir do número das amostras).

Assim como em STL e Benford, a partir do *CoDisp* é possível construir *rankings* na busca por valores anô-

Valor (R$)	Quantidade	Preço unitário (R$)	Ocorrências
128,45	48,01	2,68	2
122,16	46,01	2,66	2
114,23	42,00	2,72	2
109,74	41,01	2,68	2
100,65	37,01	2,72	2
85,65	31,49	2,72	2
79,66	30,00	2,66	2
77,95	30,00	2,60	2
77,94	30,00	2,60	2

Tabela 11: Tabela com as Top 10 transações duplicadas no mês julho/2018 para a Entidade 215701 com despesas de etanol.

Transação/ Data	Valor (R$)	Quantidade	Preço unitário (R$)
1 19/07/2018	122,16	46,01	2,66
2 19/07/2018	122,16	46,01	2,66

Tabela 12: Transações duplicadas na mesma data.

Descrição	*CoDisp*
Mínimo	1,495578
Primeiro Quartil	3,804129
Mediana	5,482057
Terceiro Quartil	10,378071
Média	12,981544
Limiar *CoDisp* (*Threshold*)	**32,00**
Máximo	996,625000

Tabela 13: Sumário dos *scores CoDisp* para a aplicação de RRCF em todo período 2011-2018, destaque para o *threshold*.

malos. A Figura 7 mostra os valores para os maiores casos de 2018, empilhando os valores dos resíduos de todas as categorias do caso/mês.

Como o algoritmo RRCF é multivariado, possibilita a análise de outras variáveis em conjunto com o *CoDisp* e dos valores das despesas. É significativo assinalar que as categorias genéricas de combustível (Outros Combustíveis e Veículos FLEX e Assemelhados identificados pela cor vermelha na Figura 7) se manifestaram no *ranking*.

Dois casos selecionados de indícios atípicos, posicionados em primeiro e sexto no *ranking* RRCF de 2018, são tabulados na tabela 14. É peculiar a quantidade de estornos atípica (*CoDisp* = 714) e preço médio unitário e quantidade atípicas (*CoDisp* = 694). Conjectura-se as possibilidades de erros ou atualizações de valores exageradas no primeiro caso, e de correção no estoque de combustível, ou correção de valores a pagar no mês no segundo.

	CoDisp = 714	*CoDisp* = 694
Valor (R$)	11.455,82	30.283,30
Quantidade	2.662,20	1,94
Preço unitário médio (R$)	4,30	15.609,94
Liquidações	126	3
Estornos	35	0

Tabela 14: Dois casos na aplicação de RRCF para combustível Gasolina. Primeiro e sexto lugares no Top 10 RRCF em 2018.

4.3. Fusão dos resultados

A partir de cada *ranking* de STL, Benford e RRCF, e a fusão utilizando a contagem de Borda, é possível também mesclar explicitamente as análises dos casos na busca por dados anômalos. A tabela 15 demonstra os dez primeiros lugares via *Borda Count* para 2018.

Em seguida, se explora o primeiro lugar no *ranking* de Borda. O mesmo possui *Rank STL = 300* e *Rank RRCF = 904*. Utilizando a análise de duplicados em Benford, a tabela 16 demonstra cinco transações idênticas para o caso.

Na tabela 17 listamos todas as datas da transações, cujos valores são iguais a R$ 307,80, que mais ocorreram no mês (oito vezes). Importante salientar que a análise e investigação pode se estender também para as transações em valores únicos (não duplicadas), atentando além das ocorrências idênticas.

O caso selecionado de anomalia em questão, apresentado na tabela 18, tem *score CoDisp* = 199,11, com valores expressivos em valor monetário despendido e número de liquidações (acima do terceiro quartil para 2018 conforme apresentado na tabela 4). Assim como individualmente as análises podem se complementar (de forma eventual), como nos caso para Entidade 215701 - Julho - 2018 (e.g. Figura 5 em anomalia identificada via STL; e tabela 11 de transações duplicadas via Benford; e demais considerações), a fusão por Borda Count funde efetivamente as técnicas para se tornarem complementares. Ambas as utilizações são úteis para análises dos valores atípicos.

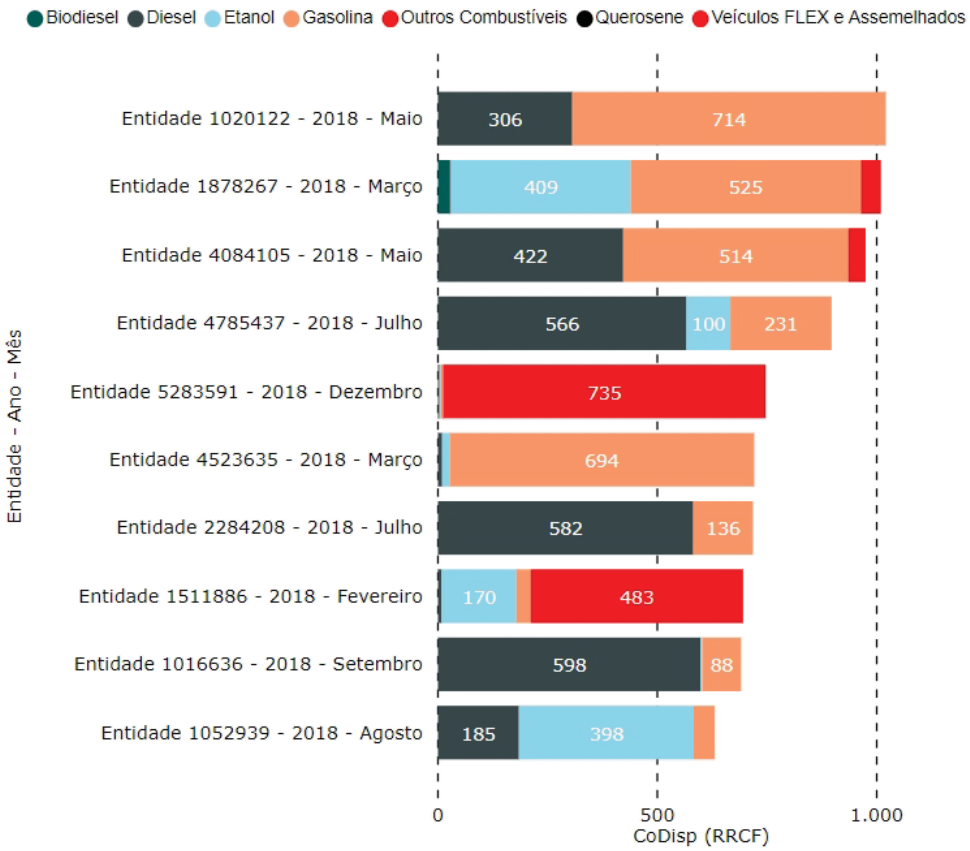

Figura 7: Top 10 casos em 2018 com os maiores *scores* empilhados em RRCF.

Borda Count	Entidade	Mês	Combustível	Valor (R$)	Quantidade	Preço unitário médio (R$)	Rank STL	Rank Benford	Rank RRCF
1	4545654	Novembro	Diesel	140.157,84	41.120,14	3,41	3.573	300	904
3	994189	Outubro	Diesel	203.179,89	57.411,70	3,54	977	864	4.124
2	215701	Fevereiro	Diesel	154.807,50	52.151,97	2,97	4.310	162	1.304
4	4545654	Abril	Diesel	149.793,99	47.838,62	3,13	5.928	310	669
5	5247562	Maio	Diesel	138.951,23	30.140,39	4,61	1.552	1.646	4.226
6	994189	Dezembro	Diesel	214.575,71	63.808,92	3,36	269	456	6.792
7	3218786	Agosto	Diesel	169.515,54	50.451,80	3,36	5.159	1.854	706
8	1052939	Agosto	Etanol	71.335,57	25.106,86	2,84	4.826	2.863	245
9	3946961	Agosto	Diesel	248.438,17	76.826,28	3,23	7.237	307	704
10	1052939	Maio	Etanol	86.421,53	28.634,84	3,02	1.866	4.267	2.353

Tabela 15: Top 10 casos em 2018 com os maiores *scores* em Borda count.

5. Conclusões e Trabalhos Futuros

Neste fechamento do trabalho, sinteticamente, são reiterados as possibilidades de resoluções ao problema exposto. As técnicas STL, Benford e RRCF isoladamente, ou em conjunto no consenso de Borda, viabilizam levantamentos (único ou múltiplos, em lista ou *ranking*) para perscrutar as despesas informadas com combustível.

Com sua aplicação, foram identificados dados atípicos como indícios que podem ser erros nos dados, dados fora de parâmetros sazonais ou de tendência, dados duplicados que chamam atenção, e valores discrepantes. Dessa forma, as técnicas estatísticas e de aprendizagem de máquina cumpriram os objetivos.

Vale assinalar que os resultados fornecem indícios que ensejam análises e captação de outras informações, dentre outros procedimentos, e não forçosamente

Valor (R$)	Quantidade	Preço unitário (R$)	Ocorrências
307,80	90,00	3,42	8
233,10	70,00	3,33	5
174,42	51,00	3,42	3
198,36	58,00	3,42	3
313,20	90,00	3,48	3

Tabela 16: Primeiro lugar no rank *Borda Count*. Top 5 transações duplicadas no mês novembro/2018 para a entidade 4545654.

Data	Transação (*)	Valor (R$)	Quantidade	Preço unitário (R$)
19/11/18	10	307,80	90,00	3,42
19/11/18	11	307,80	90,00	3,42
20/11/18	12	307,80	90,00	3,42
20/11/18	13	307,80	90,00	3,42
25/11/18	14	307,80	90,00	3,42
27/11/18	15	307,80	90,00	3,42
29/11/18	16	307,80	90,00	3,42
30/11/18	17	307,80	90,00	3,42

Tabela 17: Transações duplicadas na mesma data ou datas muito próximas. (*) Número de transação anonimizado.

	CoDisp = 199,11
Valor (R$)	140.157,84
Quantidade	41.120,14
Preço unitário médio (R$)	3,41
Liquidações	468
Estornos	0

Tabela 18: RRCF: primeiro lugar no ranking Borda Count em novembro/2018 da entidade 4545654.

configuram achados efetivos. Assim, ilustrando, se faria jus a inspeções dessas informações pelas instituições públicas competentes ou pedidos de acesso à informação se valendo da Lei de Acesso à Informação e Transparência.

As recomendações e sugestões que abrem perspectivas para novas abordagens no contexto deste trabalho são inúmeras. Em termos de continuidade nas mesmas concepções deste trabalho, podem ser acrescentadas nos estudos o restante dos dados (registros de consumos, hodômetro, horímetro), as variáveis exógenas como as já citadas nas delimitações (seção 1.6), e angariar e analisar as despesas relacionadas neste conjunto com outros insumos como pneus e lubrificantes.

Sobre a aplicação das técnicas, podem ser evoluídos o uso e análise dos resíduos para a aplicação de MAD (*Mean Absolute Deviation* - desvio médio absoluto) ou chi-quadrado na análise de Benford; a aplicação de MAD (*Median Absolute Deviation* - desvio absoluto pela mediana) em STL e a aplicação do RRCF em fluxo dos dados (*streaming*) assim considerando a correlação das variáveis no tempo. Ademais, a detecção pode ser amplificada para diferenciar tipos de *outlier* como condicional (ou contextual) e global, pois neste trabalho os tipos não foram diferenciados.

Além disso, com enfoque em mudar ou estender a forma de solução aplicada neste trabalho, as possibilidades seriam empregar técnicas de econometria, atualizações dos valores utilizando índices econômicos, comparações com preços de bases externas (e.g. ANP), aplicar técnicas estatísticas de séries temporais multivariadas ou modelos geo-temporais, e examinar os dados para análise de eficiência e desempenho e comparação entre as entidades.

6. Agradecimentos

Agradeço imensamente a todos os professores dos Departamentos de Estatística e Computação da UFPR quanto a criação, coordenação e manutenção da Especialização em Data Science & Big Data, e final e especialmente ao orientador Prof. Dr. José Padilha.

Referências

[1] BRASIL. *Lei Complementar n. 131, de 27 de maio de 2009*. Determina a disponibilização, em tempo real, de informações ao acesso público. Brasília, DF, maio 2009. http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp131.htm

[2] TCE-PR. *Gastos municipais com combustível 2011 a 2012*. Acessado em: 31/05/2019. Disponível em: `http://www1.tce.pr.gov.br/conteudo/combustiveis/235523/area/250`

[3] TCE-PR. *Portal Informação para Todos. Download de Dados. Consolidado 2013 a 2018*. Acessado em: 31/05/2019 Disponível em: `http://servicos.tce.pr.gov.br/TCEPR/Tribunal/Relacon/Dados/DadosConsulta/Consolidado`

[4] Santos, Lincoln José dos. (2015). *A Importância Do Principio Da Eficácia Para Os Tribunais De Contas - Caso Dos Combustíveis Nos Municípios Do Estado Paraná*. Trabalho de Conclusão de Curso. (2015) (Bacharelado em Direito) Faculdade De Educação Superior Do Paraná – Fesp.

[5] Benford, Frank. (1938) *The Law of Anomalous Numbers*. Proceedings of the American Philosophical Society, vol. 78, no. 4, 1938, pp. 551–572. JSTOR, www.jstor.org/stable/984802.

[6] B Cleveland, Robert & S Cleveland, William & E McRae, Jean & Terpenning, Irma. (1990). *STL: A Seasonal-Trend*

Decomposition Procedure Based on Loess. Journal of Official Statistics. 6. 3-33.

[7] Guha, Sudipto, N. Mishra, G. Roy, & O. Schrijvers. (2016). *Robust Random Cut Forest Based Anomaly Detection on Streams*. Proceedings of the 33rd International conference on machine learning, New York, NY, 2016 (pp. 2712-2721).

[8] Aggarwal, Charu C. *Outlier Analysis - Second Edition*. IBM T. J. Watson Research Center, Yorktown Heights, New York, November 25, 2016

[9] Ben-Gal, I. *Outlier detection*. Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

[10] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. *Anomaly detection: A survey*. ACM Comput. Surv., 41:15:1–15:58, 2009.

[11] Pimentel, Marco A.F., David A. Clifton, Lei Clifton, Lionel Tarassenko *A review of novelty detection*. Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK, http://dx.doi.org/10.1016/j.sigpro.2013.12.026

[12] Hochenbaum, J., Vallis, O.S., & Kejariwal, A. (2017). *Automatic Anomaly Detection in the Cloud Via Statistical Learning*. CoRR, abs/1704.07706.

[13] Melo, Diemisom Carlos Romano de. (2015). *Inteligência computacional aplicada à detecção e correção de outliers em séries temporais: estudo de caso em consumo de energia elétrica*. Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Belém, 2015. Programa de Pós-Graduação em Engenharia Elétrica.

[14] Lu, Fletcher, J. Efrim Boritz, and Dominic Covvey. (2006.) *Adaptive Fraud Detection using Benford's Law*. Conference Paper in Lecture Notes in Computer Science · June 2006 DOI: 10.1007/1176624730

[15] Nigrini, M. J. (2012). *Benford's Law: Application for Forensic Accounting, Auditing and Fraud Detection*. Wiley and Sons: New Jersey.

[16] Liu, F.T., Ting, K.M., & Zhou, Z. (2012). *Isolation-Based Anomaly Detection*. TKDD, 6, 3:1-3:39.

[17] Barai, Anwesha (Deb), Lopamudra Dey. (2017). *Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering*. World Journal of Computer Application and Technology 5(2): 24-29, 2017 DOI: 10.13189/wjcat.2017.050202

[18] Souza, Helds de Medeiros. (2017) *Detecção de transações não usuais em lançamentos contábeis através de técnicas estatísticas multivariadas*. Trabalho de Conclusão de Curso. (Graduação em Estatística) - Universidade Federal do Paraná. Orientador: Elias Teixeira Krainski.

[19] M. van Erp, L. Vuurpijl and L. Schomaker. (2002) *An overview and comparison of voting methods for pattern recognition* Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition, Niagara on the Lake, Ontario, Canada, 2002, pp. 195-200.

[20] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[21] Python Software Foundation. Python Language Reference, version 3.6.5. Available at http://www.python.org