UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Raissa Costa Brizeno

Identificação de *Outliers* para Detectar Riscos de Gestão

Recife

2018

Raissa Costa Brizeno

Identificação de *Outliers* para Detectar Riscos de Gestão

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Orientador: Cleviton Monteiro

Coorientador: Rinaldo Lima

Recife

2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

B862i Brizeno, Raissa Costa
Identificação de outliers para detectar riscos de gestão / Raissa Costa Brizeno. – 2018.
52 f. : il.

Orientador: Cleviton Monteiro.
Coorientador: Rinaldo José de Lima.
Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) – Universidade Federal Rural de Pernambuco, Departamento de Informática, Recife, BR-PE, 2018.
Inclui referências e apêndice(s).

1. Outliers (Estatística) 2. Estatística – Métodos gráficos 3. Estatística – Cartas, diagramas, etc. 4. Decisão estatística 5. Processo decisório I. Monteiro, Cleviton, orient. II. Lima, Rinaldo José de, coorient. III. Título

CDD 004

RAISSA COSTA BRIZENO

IDENTIFICAÇÃO DE OUTLIERS PARA DETECTAR RISCOS DE GESTÃO

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 02 de Setembro de 2018.

BANCA EXAMINADORA

Cleviton Vinicius Fônseca Monteiro (Orientador)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Rinaldo José de Lima
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Gabriel Alves de Albuquerque Junior
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

À Deus e minha família, essenciais para o meu sucesso.

Agradecimentos

Mais um ciclo se encerra e gostaria de deixar meus sinceros agradecimentos à todas as pessoas incriveis que me acompanharam por toda esta jornada.

Agradeço primeiramente à Deus e aos meus pais, Irani, Rosa, Inês e Ricardo por todo o amor, incentivo e apoio em todos os momentos que impulsionaram a minha educação.

Agradeço ao meu irmão Daniel por me ensinar diariamente o sentido de paciência, pureza e amor e à minha irmã Mayza que sempre esteve comigo, me aconselhando, amando e abraçando em todos os momentos de tristezas e alegrias, e também por dividir todos os períodos difíceis da minha vida. Amo muito vocês!

Ao meu orientador Cleviton Monteiro e coorientador Rinaldo Lima pelo suporte, paciência, dedicação, correções e incentivo.

Aos meus colegas de classe e amigos que fizeram parte da minha formação, tornando tudo mais divertido e colaborativo. Gostaria de agradecer de forma especial à Panda, Matheus, Igor, Daivid, Blenda, Delando, Daniel, François, Guilherme Matheus, Rafa Montenegro, Netinho, Demis, Jorge, Edvan, Rodolfo e João.

De forma especial gostaria de agradecer à Lili e Gabi, pela amizade sincera, por todo apoio, noites sem dormir estudando e por dividir todos os momentos fossem eles de aperto, tristezas ou alegrias. Vocês foram demais meninas!

Agradeço à UFRPE pelo ambiente criativo, e na maioria das vezes, amigável que nos proporciona.

Agradeço também à todos os professores por me proporcionarem conhecimento, em especial Taciana Pontual, Ceça Moraes e Gabriel Alves que foram inspirações pra mim.

Também gostaria de prestar minha gratidão à Carol, rainha de BSI, pelo acolhimento, principalmente nos momentos de sufoco.

Agradeço aos meus companheiros de trabalho em especial ao Raphael Brito pela paciência e todo o suporte que me prestou nos meus momentos de descabelamentos com os codiguinhos.

E por fim, agradeço à todos aqueles que, mesmo indiretamente, contribuíram não só para a realização deste trabalho como também por toda esta longa caminhada.

”In a dark place we find ourselves, and a little more knowledge lights our way”
(Yoda)

Resumo

Os *outliers* são valores que não convergem com o restante dos dados de uma série. Estes valores quando surgem no contexto financeiro podem representar problemas que influenciam diretamente na saúde de um empreendimento e na tomada de decisão pelos gestores. Diante disto pretendeu-se com este trabalho identificar anomalias em lançamentos financeiros advindos contas contábeis de empresas reais. Para isto, realizou-se análises estatísticas dos lançamentos para que técnicas de detecção de *outliers* pudessem ser escolhidas e, posteriormente, comparadas com a detecção de *outliers* de avaliadores. Dentre a grande variedade de técnicas foram escolhidos os métodos de Boxplot, Boxplot ajustado, MAD e desvio padrão. Os resultados obtidos mostram que a maioria das séries não seguiam uma distribuição normal, e os resultados experimentais das comparações entre os métodos automáticos e os avaliadores demonstraram diferenças substanciais.

Palavras-chave: Outliers, métodos de identificação de outliers, avaliação por especialistas.

Abstract

Outliers are values that doesn’t converge with the rest of the data series. These values when they arise in financial context can represent problems that have a direct influence on the health of an enterprise and the decision-making by the managers. In view of this, it was intended with this work identify anomalies in financial launches arising from the accounts of real companies. For this, statistical analyzes of the launches were fulfilled in order that outliers detection techniques could be chosen and then compared with the outliers detection of evaluators . Among the great variety of techniques were chosen the methods of Boxplot, Boxplot adjusted, MAD and standard deviation. The results show that most of the series didn’t follow a normal distribution, and the experimental results of the comparisons between the automatic methods and the evaluators showed substantial differences.

Keywords: Outliers, outliers identification methods, expert evaluation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DP	Desvio Padrão
IQR	Intervalo Interquartil
Q1	Quartil Inferior
Q3	Quartil Superior
MC	Medcouple
MAD	Median Absolute Deviation

Sumário

1 Introdução

1.1 Contexto e Motivação do Trabalho

O sucesso ou fracasso de uma empresa tem relação direta com a sua gestão financeira esta é responsável pelo controle sobre as questões que envolvem os recursos financeiros do empreendimento, desde o valor de entrada e saída no caixa aos lucros e investimentos. Para este controle é necessário que se gere relatórios financeiros, que poderão auxiliar na tomada de decisões estratégicas para maximizar os resultados econômicos e garantir vantagem competitiva sobre a concorrência, além de reduzir o risco de que o empreendimento venha a falência. Uma forma de tentar minimizar os riscos e falhas na tomada de decisão é buscar identificar se está havendo algum gastos fora do normal através do controle e acompanhamento financeiro. Este acompanhamento pode ser atraves de sistemas, processos ou aplicação de algum método de análise escolhido pela gestão de acordo com as necessidades da empresa.

Quando estes processos não são automatizados, identificar lançamentos financeiros que sejam discrepantes em séries advindas de contas contábeis (categorias) pode ser bastante demorado e no contexto empresarial a agilidade e precisão na obtenção de resultados para que se possa tomar uma decisão estratégica é fundamental.

Sendo assim, observou-se a oportunidade de investigar meios automatizados de identificação de valores anormais em lançamentos financeiros. Ou seja, identificar lançamentos que fujam do comum em uma dada conta contábil. Esses valores anormais, também chamados de *outliers*, podem representar uma mudança nos gastos da empresa, um gasto irregular ou erros de pagamento ou lançamento. Todos esses eventos representam riscos de gestão que merecem ser identificado e avaliado pelos gestores.

Outliers são pontos incomuns em uma série de dados, sua detecção é uma questão de extrema importância, pois podem conter observações relevantes em relação aos dados, por exemplo mudanças no cenário, fraude, erro no cadastro de lançamentos financeiros, entre outros. Pode-se aplicar estas informações em várias vertentes, como no auxilio na tomada de decisões estratégicas em uma empresa. A partir da identificação de um *outlier* se pode indicar possíveis anomalias ou falhas que ocorreram durante algum processo e que podem ser prejudicial para o empreendimento. A identificação desses pontos é um grande desafio, pois existem vários métodos que se adequam a diversos tipos de séries de dados e com isto há possibilidade de estes métodos serem aplicados de forma equivocada, influenciando erroneamente na tomada

de decisão pelos gestores ou apontando episódios que não correspondem com a situação real do empreendimento.

Diante disto, este trabalho tem por finalidade analisar quais técnicas de detecção de *outliers* melhor se adequam à identificação de lançamentos financeiros fora da normalidade em empresas reais, através de análise das distribuição dos dados a fim de avaliar a sensibilidade das técnicas na capacidade de diagnosticar corretamente os pontos que são discrepantes.

1.2 Objetivos: Geral e Específicos

Este trabalho tem como objetivo geral revisar e comparar métodos de identificação de *outliers* para analisar o que melhor se adequa ao problema em questão.

Para atingir o intuito citado acima os seguintes objetivos específicos foram delimitados:

- Analisar a distribuição de dados dos lançamentos financeiros das empresas;
- Selecionar métodos de detecção de *outliers* adequados à distribuição dos dados das empresas;
- Analisar qual método de detecção de *outliers* melhor se adequa à identificação de pontos discrepantes em lançamentos financeiros de contas contábeis, através de análises teóricas e quantitativas;

1.3 Organização do Trabalho

O restante do trabalho encontra-se organizado em mais 5 Capítulos.

O Capítulo 2 apresenta todo o referencial teórico utilizado, abordando conceitos de trabalhos relacionados e as principais técnicas em que este trabalho se apoia.

No Capítulo 3, são descritos os procedimentos, entre eles as técnicas utilizadas para detecção de *outliers*.

No Capítulo 4, apresenta-se os resultados obtidos em cada etapa do procedimento e realiza-se a análise e interpretação desses.

Por fim, o Capítulo 5 apresenta as conclusões deste trabalho.

2 Referencial Teórico

Essa seção apresentará o arcabouço teórico utilizado na pesquisa, compreendendo conceitos ligados à gestão financeira e métodos de identificação de *Outliers*.

2.1 Gestão Financeira

A gestão financeira envolve várias práticas, ações e processos administrativos. Também planeja e faz análises das atividades financeiras de um empreendimento visando maximizar os seus resultados econômicos (SEBRAE, 201-?). A gestão financeira em um empreendimento abrange todo o processo de planejamento financeiro, desde a capitação e investimentos de recursos até o faturamento. Para isso, uma das ferramenta para planejamento e controle financeiro é chamada de fluxo de caixa. O objetivo dela é apurar e fazer projeções do saldo que tem disponível para que a empresa sempre possua capital de giro, para caso necessite de fazer alguma aplicação ou caso surja algum gasto eventual. Além disso, quando se elabora um fluxo de caixa, pode-se ter a visão geral da situação atual do negócio e também dando a possibilidade da realização de projeções futuras para que possíveis dificuldades financeiras possam ser evitadas ou minimizadas (SEBRAE, 2018).

Em conjunto com a necessidade acima, também é imprescindível que se possua um sistema para o gerenciamento de informações que suporte os dados financeiros e o provisionamento de relatórios dos custos, para que este evidencie as despesas de forma a estar sempre atualizada, garantindo também uma maior compreensibilidade e rapidez na geração e acesso das informações e com isto fortalecer o plano de atuação da empresa, garantindo uma estrutura de gestão diferenciada, que pode resultar em vantagem competitiva sobre a concorrência (BAZZOTTI; GARCIA, 2006).

> *"A análise do fluxo de caixa permite traçar estratégias para o crescimento da empresa ou reverter as situações negativas"(SEBRAE, 2018).*

Nas tomada de decisão empresarial é primordial o controle financeiro e o acompanhamento diário das informações, e quando estas são geradas representam a primeira etapa da gestão do capital de giro (SEBRAE, 2016), sendo assim, pode-se inferir que a gestão de finanças significa se preocupar com a entrada e saída (lançamentos) de recursos (CHENG; MENDES, 1989).

A FIGURA 1, é um exemplo retirado do sistema Flowup que trás algumas categorias que foram utilizadas neste trabalho, onde estas representam a variação e

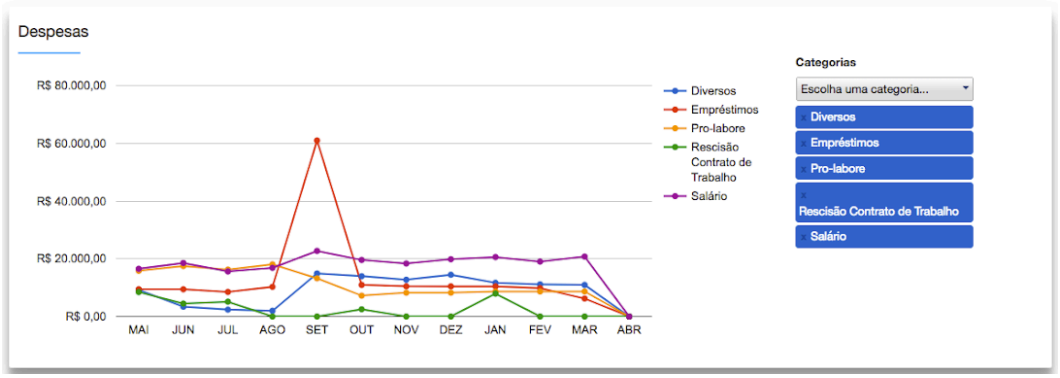

Figura 1 – Exemplo retirado do Flowup.

evolução das despesas de uma empresa ao longo do tempo, nesta imagem pode-se observar que no mês de setembro houve um aumento considerável com empréstimos e isto pode ter uma justificativa plausível, como por exemplo foi causado por alguma necessidade real da empresa ou em contrapartida pode representar alguma anomalia prejudicial para a organização ou apenas pode ter sido um erro de categorização.

Segundo Anastácio (2004), ao analisar os dados de uma empresa, é importante que não sejam desconsiderados os indicadores econômico-financeiros pois podem ser usados para projeção do futuro da empresa. Os indicadores econômicos tem por propósito mensurar o desenvolvimento econômico da empresa, sendo expressos em índices que demonstram parâmetros e permitem comparativos de desempenho periódicos (SEBRAE, 2017). Em qualquer análise de dados a detecção de valores atípicos é uma parcela bastante significativa (ZHANG et al., 2007). Analisar os lançamentos para detectar valores que não estejam de acordo com o restante da série que podem representar anomalias é importante em muitas aplicações além da detecção de fraudes, como também assistência médica, segurança e proteção pública, detecção de danos na indústria e de intrusões (HAN et al., 2012).

No contexto desse trabalho, os valores atípicos podem representar erro operacional de categorização, prejuízos, aumento do gasto com um determinado recursos, entre outros.

2.2 Identificação de *Outliers*

Outliers são pontos discrepantes em uma série de dados que podem possuir vários sentidos, o qual dependem das hipóteses relativas à estrutura dos dados e às técnicas aplicadas para detecção. Contudo, existem definições generalizadas como as de Hawkins, Barnet e Lewis e a de Johnson (BEN-GAL, 2005).

> *"Outliers são observações ou medidas suspeitas porque são muito menores ou muito maiores do que a grande maioria das observações"* (COUSINEAU; CHARTIER, 2010, tradução nossa) [1].

Outliers podem ocorrer por diferentes razões, seja ele a indicação de um erro ou um evento. O erro, também conhecido como anomalia, indica observações discordantes, exceções, falhas, defeitos, aberrações, ruído, danos ou contaminantes (ZHANG et al., 2007 apud HAN et al., 2012). É caracterizado evento quando gerado através de um "mecanismo diferente", indicando que esse tipo de *outlier* pertence a padrões inesperados que não estão em conformidade com o comportamento normal e podem incluir informações úteis sobre eventos que ocorrem raramente em inúmeras aplicações. Identificar esses *outliers* para que seja possível uma verificação adicional é de extrema importância, além disso, no dia-a-dia, com tantas despesas, é difícil perceber que o custo de alguma conta contábil está subindo demais e a organização pode ter problema de lucratividade e ter prejuízos (ZHANG et al., 2007 apud HAWKINS, 1980).

> *"Estas observações são problemáticas porque podem não ser causadas por o processo mental sob pesquisas minuciosas ou pode não refletir a habilidade sob análise. O problema é que alguns* outliers *são por vezes suficientes para distorcer os resultados do grupo (alterando o desempenho médio, aumentando a variabilidade etc)"(COUSINEAU; CHARTIER, 2010, tradução nossa).* [2]

A detecção de *outliers* é muito útil em diversas aplicações, mas enfrenta algumas dificuldades. Como refere-se (HAN et al., 2012) nos itens abaixo:

- Modelar efetivamente valores normais e *outliers*: É desafiador construir um modelo para dados normalizados, pois é difícil enumerar todos os possíveis comportamentos. A fronteira entre dados normais e anormais não é tão clara na maioria das vezes. Consequentemente, enquanto alguns métodos de detecção atribuem a cada valor no conjunto de dados um rótulo de "normal" ou "*outlier*", outros métodos atribuem uma pontuação medindo o grau de ser um *outlier*.
- Detecção de *outliers* específicos em uma aplicação: escolher a medida de similaridade e o modelo de relacionamento para descrever o conjunto de dados é critico na detecção de *outliers*, pois aplicações diferentes podem ter requisitos diferentes. Por exemplo:

[1] *"Outliers are observations or measures that are suspicious because they are much smaller or much larger than the vast majority of the observations"(COUSINEAU; CHARTIER, 2010)*

[2] *"These observations are problematic because they may not be caused by the mental process under scrutiny or may not reflect the ability under examination. The problem is that a few outliers is sometimes enough to distort the group results (by altering the mean performance, by increasing variability, etc)"(COUSINEAU; CHARTIER, 2010).*

> *"Na análise de dados clínicos, um pequeno desvio pode ser importante o suficiente para justificar um outlier. Em contraste, na análise de marketing, os objetos geralmente estão sujeitos a flutuações maiores e, consequentemente, um desvio substancialmente maior é necessário para justificar um outlier"* (HAN et al., 2012, tradução nossa) [3]

A alta dependência da detecção de valores discrepantes no tipo dos dados impossibilita o desenvolvimento de um método de detecção de exceções universalmente aplicável. Em vez disso, métodos de detecção de *outliers* individuais dedicados a casos mais específicos devem ser desenvolvidos.

- Manipulação de ruído na detecção de *outliers*: É inevitável que existam ruídos em dados coletados, estes são pontos que precisam de uma atenção pois podem representar algo danoso. Ele pode estar presente como desvios nos valores ou até mesmo como valores ausentes. Um desafio na detecção de *outliers* é a distorção dos dados que desfoca a distinção entre valores normais e *outlier*. Além disso, ruídos e dados perdidos podem "ocultar" *outliers*, quando um método de detecção pode erroneamente identificar um ruído como um *outlier*.

- Compreensibilidade: Em alguns cenários, um usuário não apenas pode detectar *outliers*, mas também entender por que os pontos detectados são *outliers* para isso algum método de detecção atípica precisa fornecer alguma justificativa para tal detecção.

2.2.1 Tipos de Análises

Análises de dados podem ser denominadas como univariadas ou multivariadas. Na análise univariada estuda-se apenas uma variável por vez, isoladamente. Se um ponto é determinado *outlier* pelos valores de seus atributos, um dado univariado é aquele que possui um único atributo (ZHANG et al., 2007), por exemplo, numa análise para identificar *outliers* em uma série de dados da conta contábil Salário, apenas os lançamentos dos custos bastariam para se analisar e identificar *outliers*.

Em contrapartida, na análise multivariada é feita uma análise em um grupo de variáveis em diferentes contextos, por exemplo análise de estruturas de covariancia, tecnicas de classificação e agrupamentos, entre outras. Essas variáveis quando associadas podem influenciar na variável resposta (PAES, 2018), por exemplo para se realizar uma projeção de lucros de uma empresa, é necessário análisar os lançamentos dos custos ao longo do tempo, então seria necessário ao menos estas duas variáveis para esta análise. Dados multivariados são os que possuem vários atributos que

[3] *"In clinic data anal- ysis, a small deviation may be important enough to justify an outlier. In contrast, in marketing analysis, objects are often subject to larger fluctuations, and consequently a substantially larger deviation is needed to justify an outlier"(HAN et al., 2012).*

podem ser estudados simultaneamente, eles dependem entre si para um valor ser identificados como atípico, pois apenas quando mais de um atributo é posto junto revelam valores distoantes, mesmo que nenhum desses seja um valor anormal individualmente (ZHANG et al., 2007).

2.3 Métodos de Detecção de *Outliers*

Existem várias abordagens para detecção de outliers que funcionarão de formas diferentes para cada conjunto de dados específico em termos de precisão e tempo de execução (ZHANG et al., 2007). Segundo (SEO, 2006) existem dois tipos de métodos para detecção de *outliers*: testes formais, também chamados de testes de discordância ou paramétricos e os testes informais, comumente chamados de métodos de rotulagem ou não paramétricos.

Para a maioria dos testes formais é necessário testes estatísticos de hipótese, que permitem tomar uma decisão entre duas ou mais hipóteses, utilizando os dados de um experimento, esses testes geralmente são baseados em alguma distribuição assumida normal e testam se o valor é um *outlier* na distribuição, ou seja, se este valor desvia-se da distribuição presumida (SEO, 2006). Segundo (ZHANG et al., 2007) abordagens paramétricas assumem que os dados podem ser modelados por uma distribuição estatística padrão, como numa distribuição normal, por exemplo para em seguida poder calcular-se os parâmetros dessa série baseando-se nas médias e covariância dos dados originais e se um ponto se desvia desse parâmetro ele é considerado um *outlier*. Esta abordagem é pertinente para distribuições de dados já conhecidas e quando as configurações de parâmetro foram previamente determinadas, contudo em situações reais, nem sempre se conhece a distribuição de dados e calcular os parâmetros não é uma tarefa simples.

A distribuição normal se apresenta uma curva em formato de sino, também conhecida como a Curva de Gauss, pode-se observar um exemplo na FIGURA 2.

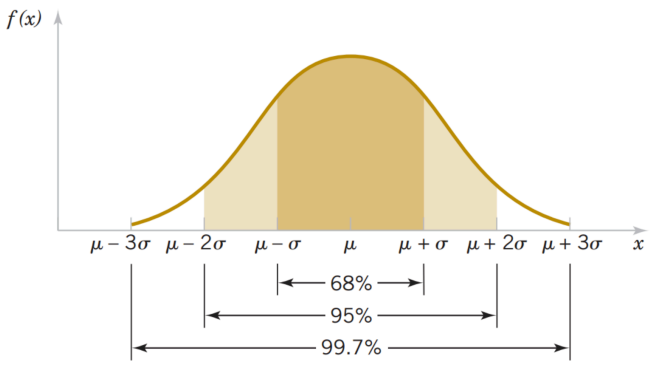

Figura 2 – Exemplo de distribuição normal.

Onde 95% aproximadamente dos valores de uma variável aleatória estão dentro de mais ou menos de um coeficiente de $K = 2$ desvios padrão da média e 99% aproximadamente dos valores de uma variável aleatória normal estão dentro de mais ou menos de um coeficiente de $K = 3$ desvios padrão de sua média (HOFFMANN, 2011).

Alguns testes são para identificar um único *outlier* e outros para vários. A seleção desses testes depende principalmente da quantidade e do tipo dos outliers, e do tipo de distribuição de dados (SEO, 2006).

> *"A escolha de um teste de discordância apropriado depende de: a) a distribuição, b) o conhecimento dos parâmetros de distribuição, c) o número de outliers esperados e d) o tipo de outliers esperados"(ACUNA; RODRIGUEZ, 2004, tradução nossa).* [4]

Os métodos de discordância têm duas principais desvantagens: primeiro, quase todos são para dados univariados, tornando-os inadequados para conjuntos de dados multidimensionais. Segundo, todos eles são baseados em distribuição normal e, na maioria das vezes, dados do mundo real são multivariados com distribuição desconhecida (ACUNA; RODRIGUEZ, 2004), como reforça (SEO, 2006) inferindo que ainda que os testes formais sejam bastante poderosos sob suposições estatísticas bem comportadas, a maioria das distribuições de dados podem ser desconhecidas ou podem não seguir distribuições específicas (SEO, 2006).

Outras limitações para detecção de *outliers* com testes formais, é que estão sujeitos aos efeitos *Masking* e *Swamping*. As seguintes definições fornecem uma compreensão desses efeitos (ACUNA; RODRIGUEZ, 2004):

- *Masking*: ocorre quando um conjunto de valores periféricos distorcem a média e as estimativas de covariância em relação a ele, e a distância resultante deste ponto periférico em relação à média é pequena. Diz-se que um *outlier* mascara um segundo *outlier*, se este *outlier* puder ser considerado como um *outlier* apenas por si próprio, mas não na presença do primeiro *outlier*, sendo assim, quando este é excluído o segundo aparece como um *outlier*.
- *Swamping*: ocorre quando um grupo de instâncias distantes distorcem a média e as estimativas de covariância em relação a ele, distanciando-se de outras instâncias não periféricas, fazendo com que a distância resultante dessas instancias até a média pareçam *outliers*. Pode-se dizer que um *outlier* anula um segundo

[4] *"The choice of an appropriate discordancy test depends on: a) the distribution, b) the knowledge of the distribution parameters, c) the number of expected outliers, and d) the type of expected outliers"* (ACUNA; RODRIGUEZ, 2004).

ponto também considerado *outlier*, apenas quando está sob a presença do primeiro ponto observado, ou seja, quando se exclui o primeiro, o segundo torna-se um *outlier*.

Os testes informais, geram um intervalo ou algum critério para a detecção de *outliers*. Para a definição destes intervalos de localização e escala, são empregados métodos de rotulagem. Então, quando um ponto é observado além desse intervalo ou critério, eles são considerados como *outlier*. Existem dois pretextos para usar um método de rotulagem, um deles é encontrar possíveis *outliers* como dispositivo de triagem antes de realizar um teste e o outro é encontrar, independente da distribuição dos dados, os valores extremos. Ainda que em geral seja simples de utilizar os métodos de rotulagem, alguns valores mesmo que fora do intervalo, podem ser identificados falsamente (SEO, 2006). De acordo com (ZHANG et al., 2007) os métodos não paramétricos identificam os *outliers* com base na medida da distância dimensional total entre os pontos, sendo assim todos os pontos distantes de seus próprios vizinhos na série de dados são considerados *outliers*.

> *"Em métodos não-paramétricos para detecção de outliers, o modelo de "dados normais" é aprendido a partir dos dados de entrada, ao invés de assumir um a priori"(HAN et al., 2012, tradução nossa).* [5]

Quando comparamos ambos os métodos, os não paramétricos, são mais flexíveis e autônomos devido ao fato de não necessitarem de conhecimento de distribuição de dados. O problema de utilizar essa abordagem está na sua complexidade dispendiosa de tempo, principalmente quando se trata de um conjunto de dados de alta dimensão (ZHANG et al., 2007). A maioria dos métodos não paramétricos não assume um modelo completamente livre de parâmetros frequentemente assumindo que a quantidade e a natureza dos parâmetros são flexíveis e não são fixados antecipadamente (HAN et al., 2012).

2.4 Testes de Normalidade

Para avaliar se as séries de dados fornecidos pelas empresas se aproximavam à uma distribuição normal foi utilizado os testes de assimetria (*Skewness*), Curtose (*Kurtosis*), Shapiro e Kolmogorov - Smirnov.

[5] *"In nonparametric methods for outlier detection, the model of "normal data" is learned from the input data, rather than assuming one a priori"(HAN et al., 2012).*

2.4.1 Assimetria

É uma medida que informa o grau de afastamento que uma distribuição apresenta do seu eixo de simetria (CORREA, 2003). Esta medida pode ser encontrada através do Coeficiente do momento de assimetria. O valor da assimetria pode ser positivo ou negativo, ou até mesmo indefinido. Se a assimetria for 0, os dados são perfeitamente simétricos, embora seja bastante improvável para dados reais. De modo geral pode-se inferir que (GOODDATA, 201-?):

- Se a medida de assimetria for menor que -1 ou maior que 1, a distribuição é altamente distorcida;
- Se estiver entre -1 e $-0,5$ ou entre $0,5$ e 1, a distribuição será moderadamente inclinada;
- Se estiver entre $-0,5$ e $0,5$, a distribuição é aproximadamente simétrica.

Exemplo de como se calcula este teste:

2.4.2 Curtose

Curtose diz a altura e nitidez do ponto central em relação ao de uma curva normal, é o grau de achatamento da distribuição (CORREA, 2003). Para o cálculo do grau de curtose de uma distribuição utiliza-se o coeficiente do momento de curtose (K). De modo geral pode-se inferir que:

1. Se $K = 3$, a distribuição é mesocúrtica;
2. Se $K > 3$, a distribuição é leptocúrtica;
3. $k < 3$, a distribuição é platicúrtica.

O 3 é muitas vezes explicado como uma correção para fazer a curtose da distribuição normal igual a zero, a curtose é aproximadamente 3 para uma distribuição normal (CORREA, 2003).

2.4.3 Shapiro-Wilk

Segundo (LOPES et al., 2013) esse teste pode ser utilizado em amostras de qualquer tamanho. Para a aplicação deste teste é necessário calcular:

$$S_h = \frac{b2}{s2}$$

, para isso precisa-se seguir os passos abaixo (ACTION, 201-?):

1. Ordenar a série;
2. Calcular $s2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$;
3. Calcular o b:
   - Se o n é par: $b = \sum_{i=1}^{n/2} a_{n-i+1} \times (x_{n-i+1} - x_i)$;
   - Se o n é ímpar: $b = \sum_{i=1}^{(n+1)/2} a_{n-i+1} \times (x_{n-i+1} - x_1)$.

2.4.4 Kolmogorov-Smirnov

Segundo (SCUDINO, 2008) esse teste avalia o grau de concordância entre a distribuição de um conjunto de valores amostrais e determina a distribuição teórica específica. Testa se os valores amostrais podem ser considerados como vindos de uma população com uma suposta distribuição teórica. Ele compara a distribuição de frequências acumulada sob hipótese nula (H_0) com a distribuição de frequência acumuladas da amostra da seguinte forma:

- Primeiramente este teste se utiliza de duas hipóteses:
  - H_0 para hipotese nula, a amostra é proveniente de uma distribuição aproximadamente normal;
  - H_1 para hipotese alternativa, a amostra não é uma distribuição normal.
- Dado $F_0(X)$ que representa a distribuição teórica acumulada sob H_0, e $S_n(X)$ que representa a distribuição de frequência dos valores amostrais de uma amostra aleatória de N observações. Como H_0 supõe que a amostra foi obtida da distribuição $F_0(X)$, então para cada valor de X em que $S_n(X_i)$ esteja próximo de $F_0(X_i)$, espera-se que as diferenças entre eles sejam pequenas, focalizando o desvio máximo (max): ($D = max|F_0(X_i) - S_n(X_i)|$).

Com isto, a regra de decisão para este teste é que se P-value, for menor que o nível de significância k, os dados apresentam distribuição normal ($max|F_0(X) - S_n(X)| < K$), caso contrário a hipotese H_0 é refutada e a amostra não provém de uma distribuição normal.

O p-value é o parâmetro valor de prova que pode ser interpretado como a medida do grau de concordância entre os dados e a hipótese nula (H0), sendo H0 correspondente à distribuição normal. Quanto menor for o P-value, menor é a consistência entre os dados e a hipótese nula. Então, a regra de decisão adotada para saber se a distribuição é Normal ou não é rejeitar H0: (i) se $P - value \leq$, rejeita-se H0, ou seja, não pode-se admitir que o conjunto de dados em questão tenha distribuição normal;

(ii) se $P-value >$, não se rejeita H0, ou seja, a distribuição Normal é uma distribuição possível para o conjunto de dados em questão (LOPES et al., 2013).

2.5 Técnicas Específicas de Identificação de *Outliers*

Existem várias técnicas para identificação de *outliers* presentes na literatura, mas será abordado nesta seção quatro métodos específicos de rotulagem para dados univariados, que serão os métodos utilizados nesta pesquisa.

2.5.1 Método Desvio Padrão - DP

É um método de rotulagem, que filtra *outliers* usando intervalos que se baseiam nos valores de desvio padrão e da média de uma série de dados (SEO, 2006).

Segundo (ACUNA; RODRIGUEZ, 2004) dado um conjunto de dados X_n de n variáveis x , seja $\bar{X_n}$ a média e s o desvio padrão da amostra de dados. Neste método uma observação x é considerada como um outlier se estiver fora do intervalo de:

$$(\bar{X_n} - Ks, \bar{X_n} + Ks),$$

onde o valor de K é geralmente tomado como 2 ou 3, a justificativa para isto se dá ao fato de que assumindo-se uma distribuição normal, espera-se ter 95% e 99% dos dados no intervalo centralizado na média, com um semi-comprimento igual a 2 e 3 de desvio padrão, respectivamente. O problema com esses critérios é que a distribuição normal dos dados é algo que frequentemente não ocorre. Além disso, a média e o desvio padrão são altamente sensíveis a *outliers* (ACUNA; RODRIGUEZ, 2004). Assumindo a série $X_n = [4.2, 5.4, 4.7, 4.7, 4.8, 4.9, 5, 5, 5.1, 5.2, 5.7, 5.8, 15, 16]$, para exemplificar este método, onde $\bar{X_n} = 6.535714$ e $s = 3.6858041$, para $K = 2$, foi identificado os valores 15 e 16 como *outliers* e para $K = 3$ não foi identificado nenhum *outlier*.

2.5.2 Método Boxplot

Introduzido por John Tukey, o *Boxplot* é uma exibição gráfica para exibir informações sobre dados contínuos univariados, como a mediana, quartil inferior (Q1), quartil superior (Q3), extremo inferior e extremo superior de um conjunto de dados. Para este método é usado o intervalo interquartil (IQR) que é a distância entre os quartis $Q1$ e $Q3$: $IQR = Q3 - Q1$ (SEO, 2006).

Neste método são distinguidos dois tipos de outliers: *mild outliers* e *extreme outliers* :

1. Para os *mild outliers*, considerados também como *outliers* suaves ou cercas internas, o intervalo localiza-se a uma distância K de $1,5\ IQR$ abaixo de $Q1$ e acima de $Q3$: $[Q1 - 1.5\times IQR, Q3 + 1.5\times IQR]$

2. Para os *extreme outliers*, considerados também como *outliers* extremos ou cercas externas, o intervalo localiza-se a uma distância K de $3\ IQR$ abaixo de $Q1$ e 3 acima de $Q3$: $[Q1 - 3\times IQR, Q3 + 3\times IQR]$

Uma observação x é declarada como um possível *outlier* se estiver entre as cercas interna e externa, sendo considerado *outlier* se estiver fora desses intervalos (SEO, 2006).

Segundo (ACUNA; RODRIGUEZ, 2004), os números 1.5 e 3 são escolhidos por comparação com uma distribuição normal. Já (SEO, 2006) diz que não há base estatística para o motivo pelo qual Tukey usa os intervalos 1.5 e 3 sobre o IQR.

Assumiremos a série $X_n = [4.2, 5.4, 4.7, 4.7, 4.8, 4.9, 5, 5, 5.1, 5.2, 5.7, 5.8, 15, 16]$, para exemplificar este método. Com $q1 = 4.8249$, $q3 = 5.625$ e $IQR = 0.8000000000000007$. Para a constante $K = 1.5$ foram identificados os valores $4.2, 15$ e 16 como *outliers* e para $K = 3$ foram identificados apenas 15 e 16.

2.5.3 Método Boxplot Ajustado

O método de Boxplot proposto por Tukey, não considera a assimetria dos dados, então Vanderviere e Huber , introduziram o Boxplot ajustado, que leva em conta o medcouple (MC), uma medida robusta de assimetria para uma distribuição de dados distorcida, para que em seguida e de acordo com o valor obtido para esta medida, se possa calcular o limite inferior ($Q1$) e o limite superior ($Q3$) do intervalo (SEO, 2006).

Para a aplicação deste método, dada uma distribuição $X_n = \{x_1, x_2, ..., x_n\}$, parte-se da premissa que este é um conjunto amostral a partir de uma distribuição univariada, ele será ordenado como $x_1 \leq x_2 \leq ... \leq x_n$, o MC desses dados é definido como:

$$MC(x_1, ..., x_n) = med\frac{(x_j - med_k) - (med_k - x_i)}{x_j - x_i},$$

onde med_k é a mediana de X_n, i e j satisfaz $x_i \leq med_k \leq x_j$, e $x_i \neq x_j$ (SEO, 2006).

O intervalo de Boxplot ajustado é dado como segue:

$$[L, U] = [Q1 - 1.5\times exp(-3.5MC)\times IQR, Q3 + 1.5\times exp(4MC) * IQR], seMC \geq 0$$

$$[L, U] = [Q1 - 1.5\times exp(-4MC)\times IQR, Q3 + 1.5\times exp(3.5MC) * IQR], seMC \leq 0,$$

onde L é a cerca inferior e U é a cerca superior do intervalo. As observações que estiverem fora do intervalo são considerados *outlier*. O valor do MC varia entre -1 e 1.

Se $MC = 0$, os dados são simétricos e o Boxplot ajustado torna-se o Boxplot de Tukey. Se MC > 0, os dados têm uma distribuição assimétrica à direita, enquanto que se MC < 0, os dados têm uma distribuição assimétrica à esquerda (SEO, 2006 apud BRYS; ROUSSEEUW, 2005).

Para exemplo, temos a série $X_n = [4.2, 5.4, 4.7, 4.7, 4.8, 4.9, 5, 5, 5.1, 5.2, 5.7, 5.8, 15, 16]$, com $q1 = 4.8249$, $q3 = 5.625$ e $IQR = 0.8000000000000007$. Aplicando este método os valores 4.2, 15 e 16 foram identificados como *outliers*.

2.5.4 Median Absolute Deviation - MAD_e

Este método não é afetado pela presença de valores extremos dos conjuntos de dados. Essa abordagem é semelhante ao método de desvio padrão, no entanto, a mediana e o MAD, uma medida robusta da dispersão de um conjunto de dados, são empregado neste método ao invés da média e do desvio padrão (SEO, 2006). Segue abaixo os passos para a aplicação deste método:

1. Calcula-se primeiramente a mediana dos dados;
2. O segundo passo é calcular o desvio absoluto de cada elemento na série de dados em relação a mediana: $D_i = |x_i - \tilde{x}|$, onde D_i é o desvio absoluto, x_i é o elemento dado e $\tilde{x}$ o valor da mediana;
3. Neste passo, será calculado a mediana dos desvios absolutos calculados anteriormente com isso teremos o MAD, que é um valor estimador, similar ao desvio padrão, mas possui um ponto de ruptura[6];

A partir disso pode-se identificar *outliers* da seguinte forma:

$$MAD_e = k \times MAD,$$

onde a constante k é uma um fator de consistência que torna o MAD não-viesado na distribuição normal. Se você sabe que a distribuição subjacente é normal, a constante de consistência K deve ser definida como $1,4826$, do contrário pode-se definir 2 ou 3 como valores de ponto de corte, como afirma (SEO, 2006) e (ROSENMAI, 2013).

Uma problemática trazida por este método é que se mais de 50% dos dados na série tiverem valores idênticos, o MAD será igual a zero. Todos os pontos em seu conjunto de dados, exceto aqueles que são iguais à mediana, serão marcados como valores discrepantes, independentemente do nível em que você definiu o limite de valor *outlier* (ROSENMAI, 2013).

[6] O ponto de ruptura ou como chamado em inglês, *breakdown*, de um ponto estimador, pode ser definido geralmente como a porcentagem mais alta dos dados que podem variar sem distorcer o estimador.

Aplicando este método na série $X_n = [1, 2, 3, 3, 4, 4, 4, 5, 5.5, 6, 6, 6.5, 7, 7, 7.5, 8, 9, 12, 52, 90]$. Foi obtido os desvios absolutos $D_i = [5, 4, 3, 3, 2, 2, 2, 1, 0.5, 0, 0, 0.5, 1, 1, 1.5, 2, 3, 6, 46, 84]$, o $mad = 2.0$, o $mad_e = 2.9652$ e a $mediana = 6.0$. Para $K = 2$, foram detectados os valores$12, 52$ e 90 como *outliers* e para $K = 3$ apenas os valores 52 e 90 foram apontados como *outliers*.

2.6 Trabalhos Relacionados

Ao longo desta pesquisa foi encontrado alguns trabalhos que comparam e propõem diversas técnicas para a detecção de *outliers*, um deles foi o artigo de (SEO, 2006) que faz um estudo comparativo entre métodos de discordância e métodos rotulagens, trazendo alguns testes dentro de cada uma dessas categorias e apontando em quais situações cada um deles podem ser utilizados, fazendo uso de exemplos e simulações, exibindo que cada método tem diferentes medidas para detectar *outliers* e mostra diferentes comportamentos de acordo com a assimetria e o tamanho da amostra dos dados. Outros artigos que estão nesta mesma linha de pesquisa são os dos autores (ACUNA; RODRIGUEZ, 2004), (ZHANG et al., 2007), (BEN-GAL, 2005) e (COUSINEAU; CHARTIER, 2010).

Quando trazemos o artigo de (ANASTACIO, 2004) pode-se fazer uma analise sobre *outliers* aplicados para a análise de dados contábeis, para que se possa extrair resultados adequados para que um analista ou auditor possa detectar, de forma genérica, indícios de erros, fraudes ou anomalias, por exemplo.

Buscando compreender sobre a detecção de outliers em dados distorcidos e sobre tipos de dados foram utilizados os artigos de (HUBERT; VEEKEN, 2007), (RAHMAN et al., 2014), (ZHANG et al., 2007) (LOO, 2010).

3 Materiais e Métodos

O procedimento realizado nessa pesquisa teve como objetivo analisar a adequação das diferentes abordagens de identificação de *outliers* ao contexto de lançamentos financeiros das empresas. O trabalho foi conduzido em 6 etapas:

- Etapa 1: Revisão da literatura sobre *outliers*;
- Etapa 2: Levantamento dos dados de lançamentos financeiros;
- Etapa 3: Análise exploratória dos dados;
- Etapa 4: Escolha das categorias para análise;
- Etapa 5: Aplicação das técnicas de identificação de *outliers* selecionadas;
- Etapa 6: Análise comparativa dos resultados obtidos.

3.1 Levantamento dos Dados de Lançamentos Financeiros

Nesta fase, foram coletados dados de lançamentos financeiros de contas contábeis, estas representam os custos com água, salário dos funcionários, energia, internet, dentre outros, neste trabalho cada conta contábil será uma categoria, estas são de períodos variados advindos de vinte empresas reais, cujas identidades serão mantidas em sigilo por se tratar de informações sensíveis. Para cada lançamento financeiro destas empresas, foram obtidos os seguintes dados: data do lançamento, descrição, categoria e valor, a FIGURA 3 exemplifica os dados brutos.

	A	B	C	D	E
1		Date	Value	Descricao	Categoria
2	0	2016-10-03	433.33	Metade do 13º Thia	13º Salário
3	1	2016-12-12	525	13º Catharina	13º Salário
4	2	2016-12-12	1319	13º Bruna	13º Salário
5	3	2016-12-12	1016	13º João	13º Salário
6	4	2016-12-12	168	13º Wendell	13º Salário
7	5	2016-12-13	364	13º Salário Thiago	13º Salário
8	6	2017-05-03	1361.59	Metade do 13º João	13º Salário
9	7	2017-06-21	1825.77	Metade 13º Bruna	13º Salário
10	8	2017-11-27	779.68	13º Roberta	13º Salário
11	9	2017-12-14	1415	13º Salário Bruna	13º Salário
12	10	2017-12-14	1073	13º Salário João	13º Salário
13	11	2017-12-14	533	13º Salário Wendell	13º Salário
14	12	2017-12-14	639	Cheque: 13º roberta	13º Salário
15	13	2016-02-15	101	Compesa - Sala torre	Água
16	14	2017-09-07	119.6	Mlabs - sistema de g	Água
17	15	2017-09-12	109.17	Combustível	Água
18	16	2016-02-11	761.73	Condomínio	Aluguel
19	17	2016-02-11	2800	Aluguel Fevereiro 20	Aluguel

Figura 3 – Exemplo dos dados brutos

3.2 Análise Exploratória dos Dados

Os dados fornecidos pelas empresas foram redistribuídos em uma nova planilha e divididos em três grupos: O primeiro grupo possui as categorias que contém uma quantidade de 6 a 10 lançamentos financeiros, o segundo grupo contém as categorias com 11 a 20 lançamentos, e no terceiro as categorias com mais de 20 lançamentos. Esses grupos foram divididos desta forma, para facilitar a aplicação dos testes de normalidade e posteriormente a análise, pois alguns testes, como por exemplo o de Curtose, são indicados para séries que possuem mais de 20 itens. As séries das categorias que possuíam menos de 6 lançamentos foram desconsideradas deste estudo.

Sobre estas informações foi efetuada a análise exploratória para identificar se as séries fornecidas seguiam uma distribuição teórica. Por exemplo, se seguiam a distribuição normal ou se possuíam assimetria. Isto foi importante para esta pesquisa, pois faz parte da premissa deste trabalho, pois os métodos de identificação de *outliers* foram selecionados com base nisto.

Devido à constatação da natureza bastante variada das distribuições dos dados amostrais, tanto em nível da assimetria quanto Curtose, foi tomada a decisão neste projeto de avaliar a votação conjunta de quatro testes de normalidade visto que quando tomados individualmente, eles fornecem votos distintos.

Para isto, foi utilizada a linguagem de programação Python e as bibliotecas Pandas, Numpy e a Scipy, com funções nativas para esses testes, o código de aplicação pode ser melhor observado no APÊNDICE A.

Foram utilizados os seguintes critérios para cada testes de normalidade aplicado:

1. Para o teste de Curtose foi aplicado a função kurtosis que retorna o coeficiente de curtose e kurtosistest, ambas nativas da biblioteca Scipy. A kurtosistest é indicada para séries de dados maiores de 20 itens e além do valor do coeficiente de Curtose, ele também retorna o P-value (SCIPY.ORG, 2018a);

2. Para o teste de assimetria, foi aplicada a função Skew, esta função retorna o coeficiente de assimetria. Foi utilizado o intervalo $-0,5$ e $0,5$, para determinar se uma série é ou não uma distribuição normal, ou seja se o coeficiente de assimetria estiver entre este intervalo, a distribuição é aproximadamente simétrica (SCIPY.ORG, 2018c). Também foi utilizado a função Skewtest que além do coeficiente de assimetria também retorna o P-value. Essa função é indicada para séries com no mínimo 8 itens,(SCIPY.ORG, 2018d).

3. Para o teste de Shapiro-Wilk foi utilizada a função Shapiro, da biblioteca scipy. Esta função retorna o valor para o coeficiente de Shapiro e o P-value. Nesta

pesquisa vamos utilizar o $p - value > 0.05$ para indicar a falta de normalidade (SCIPY.ORG, 2018b).

4. Para o teste de Kolmogorov-Smirnov foi utilizado a função Ktest, que retorna o coeficiente estatístico do teste e o P-value. Nesta pesquisa foi utilizado o nível de significância igual a 0.05 (SCIPY.ORG, 2014).

Para toda categoria presente em cada empresa foram analisados os resultados dos testes de normalidade individualmente e também combinados, a TABELA 1 demonstra as combinações utilizadas.

Categoria	Imposto FGTS
Quantidade de lançamentos	**37**
Valor Kurtosis	**0,81**
P-value Kurtosis	**0,43**
Normal para Kurtosis?	**1**
Valor Shapiro	**0,97**
P-value Shapiro	**0,31**
Normal para Shapiro?	**1**
Valor Skweness	**0,76**
P-value Skweness	**0,76**
Normal para Skweness?	**1**
Valor Kolmogorov	**1**
P-Value Kolmogorov	**0**
Normal para Kolmogorov?	**0**
Passou em todos os testes?	**0**
Passou em nenhum dos testes?	**0**
Apenas Skweness V	**0**
Apenas Kurtosis V	**0**
Apenas Shapiro V	**0**
Apenas Kolmogorov V	**0**
Skweness e kurtosis V	**0**
Kurtosis e Shapiro V	**1**
Kolmogorov e Shapiro V	**0**
Kolmogorov e Skewness V	**0**
Kurtosis e Skweness V	**1**
Kolmogorov e Kurtosis V	**0**
Skweness e Shapiro V	**1**
Normal | Empate	**Normal**

Tabela 1 – Ilustração das combinações utilizadas com os testes de normalidade do grupo 3

A categoria utilizada no exemplo acima (Imposto FGTS) mostra como os dados foram organizados para essa análise. Foi obtida a quantidade de lançamentos da categoria, os valores retornados pelos testes, o P-value, uma *flag* indicando se as séries foram consideradas normais de acordo com cada teste. Também definiu-se uma

flag para indicar se todos os testes indicaram uma distribuição normal, se nenhum deles apresentou normalidade, e as combinações dos testes dois a dois. Uma *flag* de empate também foi definida, para os casos em que exatamente dois testes indicaram normalidade. Além disso, para esta pesquisa considerou-se que a série apresentava distribuição normal se pelo menos três testes indicassem normalidade (Normal).

3.3 Escolha das Categorias para Análise

Para facilitar a análise dos resultados nesta etapa, foi realizada uma seleção de categorias e divisão dos dados. Primeiramente, devido à restrição relacionada à quantidade de dados necessária para aplicação dos testes de normalidade, o grupo que possuía séries com 6 a 10 lançamentos (Grupo 1), foi desconsiderado neste estudo.

Também houve um agrupamento dos dados de acordo com a simetria da sua distribuição devido à natureza bastante diversificada das amostras levantadas que corresponde à realidade de lançamentos financeiros de diversas empresas e, principalmente para categorizar tais distribuições quanto ao nível de assimetria. O objetivo desse agrupamento é poder analisar o comportamento dos métodos em séries com diferentes coeficientes de assimetria. Segundo (SEO, 2006) se o valor da assimetria é negativo, a distribuição dos dados será inclinada para a esquerda e, se o valor da assimetria for positivo, a distribuição dos dados será inclinada para a direita. Qualquer dado simétrico tem um valor zero de assimetria. Com isto ambos os grupos 2 e 3 foram subdivididos em três subgrupos, organizados de acordo com o coeficiente de assimetria, como segue abaixo:

- Grupo 2.1 e grupo 3.1 para alta assimetria, ou seja, X < -1 ou X > 1.0;
- Grupo 2.2 e grupo 3.2 para média assimetria, ou seja, -1 <= X < -0.5 ou 0.5 < X <= 1.0;
- Grupo 2.3 e grupo 3.3 para baixa assimetria, ou seja, -0.5 <= x <= 0.5;

Então, os dados foram rearranjados para facilitar a análise na etapa posterior, ficando conforme exibida na FIGURA 4. Observe que para cada categoria calculou-se o coeficiente de assimetria e logo abaixo foram colocados os lançamentos financeiros pertencente à categoria.

GRUPO 3.3				
Pro-labore (	Serviço de terceiro (	Imposto: FGTS	ENERGIA CELPE	Energia (
0	0	0	0,147526778	-0,181
Valor	Valor	Valor	Valor	Valor
4685	30	512	717,55	1089
4685	30	512	620,92	1150,8
4685	30	512	601,88	1252
4685	30	512	572,9	762,3
4685	30	512	523,31	409,19
4685	30	512	632,89	856,7
4685	30	512	633,94	1103,1
4685	30	512	711,18	1216,4
4685	30	512	941,15	1302,1
4685	30	512	948,08	1426,3
4685	30	512	543,66	1149,9
4685	30	512	1069,48	1003,7
4685	30	512	976,36	1144,1
4685	30	512	975,88	1435,7
4685	30	512	1011,42	1328,1
4685	30	512	305,24	1431,7
4685	30	512	1062,87	950,33
4685	30	512	183,79	1169,5
4685	30	512	418,31	1123,3
4685	30	512	337,75	1057,9
4685	30	512	438,18	1129,7
4685	30	512	86,39	1336,8
4685	30	512	71,17	1368,1
4685	30	512	529,01	1274,2
4685	30	512	614,01	1422,1
4685	30	512	417,74	1495,8
	30	512	86,77	1705,2
	30	512	636,8	1198,2
	30	512	617,15	1116,7
	30	512	488,31	508,33
	30	512	545,31	164,08
	30	512	493,71	460,95
	30	512	563,65	245,47
	30	512	496,52	1070,8
	30	512	647,41	466,88
	30	512	494,4	153,44
	30	512	517,86	181,08
	30	512	544,99	1152
	30	512	568,41	877,39

Figura 4 – Exemplo da planilha com os dados reorganizados do grupos 3 para baixa assimetria.

3.4 Experimentos Realizados

3.4.1 Aplicação das Técnicas de Identificação de *Outliers* Selecionadas

Devido a variedade de técnicas presentes na literatura, foram selecionados para análise quatro métodos específicos de rotulagem para dados univariados. Neste ponto, foi aplicado os métodos para identificação de *outliers*, individualmente, nos lançamentos previamente selecionados, foram estes: Desvio padrão, M.A.D, Boxplot, e Boxplot ajustado, utilizando Python (APÊNDICE B). Buscava-se com isso identificar valores que fossem discrepante com o restante da série e como cada teste se comportava individualmente em cada série, para comparação e análise posteriormente.

3.4.2 Análise Comparativa dos Resultados

Neste passo foi realizada uma análise qualitativa dos resultados. O objetivo foi avaliar as congruências e divergências dos métodos na identificação de *outliers*, assim como comparar os *outliers* identificados pelos métodos com a identificação feita por especialistas. Para isso, as categorias foram divididas em dois grupos, o primeiro continha apenas as séries onde os métodos de identificação de *outlier* indicaram exatamente os mesmos itens como *outliers* e o segundo que continha apenas as séries onde estes métodos indicaram diferentes lançamentos como *outliers*. O objetivo dessa

divisão foi focar as análises nas categorias onde os métodos divergiam.

A partir disso, foram selecionadas aleatoriamente cinquenta séries, sendo quatro séries das categorias em que os métodos apresentaram resultados iguais e as demais do outro grupo. Essa redução da quantidade de categorias para análise foi necessária para que os especialistas pudessem realizar a identificação dos *outliers* manualmente em cada categoria.

Para comparar o resultado de identificação de *outliers* pelos métodos entre si, e com a identificação feita pelos especialistas, utilizou-se o método Kappa de Cohen. Este método resulta em um valor que expressa o nível de concordância entre dois avaliadores (juízes) em um problema de classificação (LANDIS; KOCH, 1977), ou seja o Kappa é uma medida de concordância interobservador e mede o grau de concordância além do que seria esperado tão somente pelo acaso.

Para se caracterizar melhor os outliers, principalmente no contexto de uma avaliação por especialistas, é necessário ter os lançamentos caracterizado ou classificado várias vezes, por exemplo, por mais de um juiz, pois eles irão divergir entre si.

Para avaliar se a concordância é aceitável a hipótese testada é se o Kappa=0 indicando concordância nula ou se Kappa>0. No caso de rejeição da hipótese nula aponta que a medida de concordância é significantemente maior que zero, não necessariamente significando que há concordância alta ou baixa, mas ainda assim indicando alguma concordância. Cabe uma avaliação para atestar se a medida obtida é satisfatória.

Para manter uma nomenclatura consistente ao descrever o nível de concordância entre os juízes de acordo com as estatísticas kappa, este estudo foi baseado nas variações de perspectivas abordadas por (LANDIS; KOCH, 1977), que está ilustrado na TABELA 2.

Valor para o Kappa	**Perspectiva**
<0	Sem concordância
0.0 à 0.19	Concordância pobre
0.20 à 0.39	Concordância justa
0.40 à 0.59	Concordância moderada
0.60 à 0.79	Concordância substancial
0.80 à 1.00	Concordância quase perfeita

Tabela 2 – Significação dos níveis de concordância

Os resultado entre os métodos de detecção de *outliers* para cada categoria foram comparados entre si. Observe na FIGURA 5 que o resultado de cada método para cada lançamento das categorias foi marcado com s, que significa que o item é um *outlier*, ou n, caso contrário.

	A	B	C	D	E	F	G	H
1	Dados	Boxplot (1.5)	Boxplot (3)	Boxplot ajustado	MAD (2)	MAD (3)	Desvio padrão (2)	Desvio Padrão (3)
48	OUTROS INVESTIMENTOS							
49	235	n	n	n	n	n	n	n
50	371,22	n	n	n	n	n	n	n
51	99	n	n	s	n	n	n	n
52	189	n	n	n	n	n	n	n
53	600	n	n	n	n	n	n	n
54	337,95	n	n	n	n	n	n	n
55	484,5	n	n	n	n	n	n	n
56	57,86	n	n	s	n	n	n	n
57	260	n	n	n	n	n	n	n
58	591,87	n	n	n	n	n	n	n
59	260	n	n	n	n	n	n	n
60	60000	s	s	s	s	s	s	s
61	788,48	n	n	n	n	n	n	n
62	260	n	n	n	n	n	n	n
63	142	n	n	n	n	n	n	n
64	179,91	n	n	n	n	n	n	n
65	1000	n	n	n	s	n	n	n
66	1000	n	n	n	s	n	n	n
67	INOVA Cartório							
68	12,5	n	n	n	n	n	n	n
69	87	n	n	n	n	n	n	n
70	36	n	n	n	n	n	n	n
71	306	n	n	n	s	s	n	n
72	60	n	n	n	n	n	n	n
73	144	n	n	n	s	s	n	n
74	12,5	n	n	n	n	n	n	n
75	37,5	n	n	n	n	n	n	n
76	10,8	n	n	n	n	n	n	n
77	3,6	n	n	n	n	n	n	n
78	32,5	n	n	n	n	n	n	n
79	9100	s	s	s	s	s	s	s
80	3250,05	s	s	n	s	s	n	n
81	OUTROS impostos mercadorias e serviços							
82	192	n	n	n	n	n	n	n
83	1555,34	n	n	n	n	n	n	n
84	373,79	n	n	n	n	n	n	n
85	624,12	n	n	n	n	n	n	n
86	1251,89	n	n	n	n	n	n	n
87	1365,82	n	n	n	n	n	n	n
88	509,74	n	n	n	n	n	n	n
89	1248,66	n	n	n	n	n	n	n
90	5000	s	n	n	s	s	n	n

Figura 5 – Exemplo da planilha que continha o resultado dos métodos de identificação de outlier.

Então, criou-se uma coluna a mais que representa o resultado da votação entre os métodos, chamada Votação-método (FIGURA 6). Essa coluna foi preenchida com o valor indicado por mais de 50% dos métodos. Então, os valores das colunas de cada método individualmente foram comparados com o valor da Votação-método utilizando o Kappa.

	A	B	C	D	E	F	G	H	I	J
1	Dados	Boxplot (1.5)	Boxplot (3)	Boxplot ajustado	MAD (2)	MAD (3)	Desvio padrão (2)	Desvio Padrão (3)	Boxplot + Skeness	Votação-método
2	Imposto: PIS	s	s	s	s	s	s	n	s	s
3	63,31	s	s	s	s	s	s	n	s	s
4	37,07	n	n	n	n	n	n	n	n	n
5	532,76	n	n	n	n	n	n	n	n	n
6	532,76	n	n	n	n	n	n	n	n	n
7	532,76	n	n	n	n	n	n	n	n	n
8	532,76	n	n	n	n	n	n	n	n	n
9	532,76	n	n	n	n	n	n	n	n	n
10	532,76	n	n	n	n	n	n	n	n	n
11	532,76	n	n	n	n	n	n	n	n	n
12	532,76	n	n	n	n	n	n	n	n	n
13	532,76	n	n	n	n	n	n	n	n	n
14	532,76	n	n	n	n	n	n	n	n	n
15	532,76	n	n	n	n	n	n	n	n	n
16	532,76	n	n	n	n	n	n	n	n	n

Figura 6 – Resultado final para a comparação entre os métodos de identificação de outlier.

As categorias também foram submetidas à uma análise por três avaliadores

humanos, um gestor de projetos, um gestor financeiro e um pesquisador, que apontaram os itens que julgaram *outliers*. Por não existir uma definição precisa do que é ou não um outlier no domínio estudado foi preciso que especialistas dessem sua opinião sobre os lançamentos das séries para que estas pudessem ser comparadas com os resultados dos métodos de identificação selecionados para esta pesquisa. Esta fase seguiu os seguintes critérios:

1. Critério 1: Se mais de 50% dos especialistas marcarem um valor como *outlier*, este valor será considerado um *outlier* e com isto resultou em uma coluna Votação-especialista1 como observa-se na FIGURA 7.

2. Critério 2: Se pelo menos um dos especialistas marcar um valor como *outlier*, este valor será considerado um *outlier* e com isto resultou em uma coluna Votação-especialista2 como observa-se na FIGURA 7.

	A	B	C	D	E	F
1	**Dados**	**Avaliador 1**	**Avaliador 2**	**Avaliador 3**	**Votação_Especialista 1**	**Votação_Especialista 2**
2	**Imposto (PIS)**					
3	63,31	s	s	s	s	s
4	37,07	s	s	s	s	s
5	532,76	n	n	n	n	n
6	532,76	n	n	n	n	n
7	532,76	n	n	n	n	n
8	532,76	n	n	n	n	n
9	532,76	n	n	n	n	n
10	532,76	n	n	n	n	n
11	532,76	n	n	n	n	n
12	532,76	n	n	n	n	n
13	532,76	n	n	n	n	n
14	532,76	n	n	n	n	n
15	532,76	n	n	n	n	n
16	532,76	n	n	n	n	n
17	**Software tratamento de ponto**					
18	39,9	n	n	n	n	n
19	39,9	n	n	n	n	n
20	195	s	s	s	s	s
21	39,9	n	n	n	n	n
22	39,9	n	n	n	n	n
23	39,9	n	n	n	n	n
24	39,9	n	n	n	n	n
25	39,9	n	n	n	n	n
26	39,9	n	n	n	n	n
27	39,9	n	n	n	n	n
28	39,9	n	n	n	n	n
29	39,9	n	n	n	n	n
30	39,9	n	n	n	n	n
31	39,9	n	n	n	n	n
32	39,9	n	n	n	n	n
33	39,9	n	n	n	n	n
34	260	n	n	n	n	n
35	274	s	n	n	n	s
36	260	n	n	n	n	n
37	260	n	n	n	n	n
38	260	n	n	n	n	n
39	260	n	n	n	n	n
40	260	n	n	n	n	n

Figura 7 – Exemplo da planilha com as séries contendo o resultado dos avaliadores e resultado final da comparação entre os votos deles.

Após isto o método Kappa de Cohen foi aplicado das seguintes formas:

- Entre cada avaliador e a coluna Votação-especialista1;
- Entre cada avaliador e a coluna Votação-especialista2;
- Entre as colunas de Votação-método e Votação-especialista1;

- Entre as colunas de votação-método e Votação-especialista2;
- Entre as colunas de votação-método e Votação-especialista2;
- Entre cada método automático e Votação-especialista1;
- Entre cada método automático e Votação-especialista2;

4 Resultados e Discussões

Primeiramente serão apresentadas informações descritivas dos dados utilizados nesse trabalho. Observe na TABELA 3 a quantidade das categorias de cada empresa, a quantidade geral de lançamentos por empresa, a quantidade máxima e a quantidade mínima de itens lançados por empresa. Com esses valores pôde-se encontrar para a quantidade de categorias, o desvio padrão e média da quantidade de lançamentos presentes nas categorias de cada empresa.

Empresa	**Quant. de categorias**	**Quant. de lançamentos**	**Média de lançamentos**	**Desvio padrão da quant. de itens**	**Quant. de itens da categoria com mais itens**	**Quant. de itens da categoria com menos itens**
Empresa 1	30	1103	36,97	28,42	147	10
Empresa 2	31	3331	107,46	162,85	696	13
Empresa 3	15	924	61,6	43,67	160	12
Empresa 4	51	2939	57,63	95,01	473	10
Empresa 5	5	287	57,4	49,82	117	13
Empresa 6	33	6331	191,85	309,70	1269	14
Empresa 7	25	1350	54	67,52	258	10
Empresa 8	13	309	23,77	14,02	56	13
Empresa 9	48	9958	207,46	531,33	3458	13
Empresa 10	13	583	44,85	73,92	285	10
Empresa 11	18	1938	107,67	99,14	309	10
Empresa 12	31	616	19,88	13,96	66	10
Empresa 13	81	9433	116,46	244,58	1470	10
Empresa 14	87	10245	117,76	123,78	613	12
Empresa 15	6	274	45,67	42,57	126	10
Empresa 16	54	3721	68,91	128,28	857	11
Empresa 17	12	353	29,42	29,31	112	11
Empresa 18	39	4163	106,75	113,78	496	14
Empresa 19	34	1555	45,74	84,98	495	10
Empresa 20	34	5692	167,42	295,18	1452	12

Tabela 3 – Ilustração das séries com valores maiores ou iguais a 10.

As empresas nem sempre apresentavam a mesma quantidade de categorias e estas podiam se apresentar em tipos diferentes entre as empresas. Os lançamentos variavam entre si e em alguns casos apresentavam valores repetidos, com nenhuma variância ou pouca variância.

Na análise exploratória, com a aplicação dos testes de normalidade foram en-

contrados os seguintes resultados para cada grupo:

- **Grupo 2:** Apresentou 149 categorias, das quais:

Testes de normalidade	Teste de normalidade verdadeiro
Curtose	55,70%
Assimetria	27,51%
Shapiro	23,48%
Kolmogorov	0%

Tabela 4 – Resultados dos testes de normalidade para grupo 2

Todos os testes verdadeiros	0%
Não passaram em pelo menos um dos testes	16,77%
Empate	22.14%
Distribuições normal	0%
Não passaram em pelo menos dois testes	77.85%

Tabela 5 – Resultados gerais para grupo 2

- **Grupo 3:** Apresentou 491 categorias, das quais:

Testes de normalidade	Teste de normalidade verdadeiro
Curtose	23,89%
Assimetria	20,36%
Shapiro	2,85%
Kolmogorov	0%

Tabela 6 – Resultados dos testes de normalidade para grupo 3

Todos os testes verdadeiros	0%
Não passaram em pelo menos um dos testes	68,22%
Empate	7,12%
Distribuições normal	2,64%
Não passaram em pelo menos dois testes	90,22%

Tabela 7 – Resultados gerais para grupo 3

Nesta etapa, pode-se perceber nos dados apontados nas TABELAS 4, 5, 6 e 7, que com a aplicação dos testes de normalidade foi verificado que a maioria das séries não seguem uma distribuição normal.

Além disso, pode ser também observado na FIGURA 8 os resultados para assimetria. Em maioria estes demonstraram que a quantidade de séries que atendiam à uma distribuição normal, foram pouco significativas.

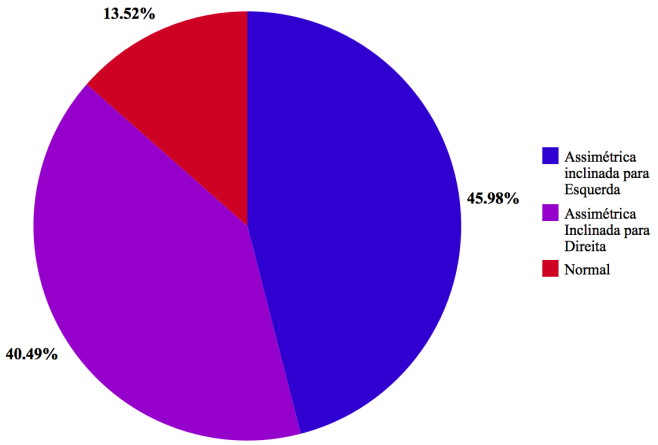

Figura 8 – Resultados para os valores de assimetria

Devido à constatação de que a maioria das categorias possuíam lançamentos com valores que não obedeciam à distribuição normal, foram selecionado quatro testes que não possuíam esta premissa. Com isto na etapa quatro, com a aplicação dos métodos de identificação de *outliers* que foram selecionados, foi observado um total de 40 categorias nas quais os resultados dos testes foram iguais. Observou-se uma particularidade nessas categorias, boa parte dos lançamentos possuíam valores iguais, apresentando poucos valores diferentes na série. Por outro lado, em 106 categorias o resultado de identificação de *outliers* foi diferente entre os testes.

Na etapa seis, devido à grande quantidade de categorias para analisar, foi feita uma seleção amostral de 50 categorias, conforme explicado anteriormente na seção de Materiais e Métodos. Nesta pesquisa pretendeu-se responder questões estatísticas relativas a comparação entre os resultados obtidos pela aplicação das diferentes técnicas para detecção de *outliers* com a avaliação por especialistas. Ou seja, se haveria algum método automático que atenderia às expectativas dos especialistas ou se eles apresentam diferenças muito discrepantes entre si, pois não há um método ou uma definição clara que garanta constância na detecção de *outliers* automaticamente. Com isto foram realizadas algumas comparações a fim de identificar quais métodos mais se aproximavam da Votação-método. Na primeira foi aplicado o Kappa de Cohen entre os métodos aplicados com a Avaliação-método. Abaixo, NA FIGURA 9 pode-se examinar os resultados obtidos com esta comparação.

Os níveis de concordância obtidos acima permitiu observar que a comparação do método Boxplot com o coeficiente $k = 3$ e votação-método foi o único que obteve a perspectiva quase perfeita entre as combinações dos métodos com Votação-método, em segundo lugar os métodos Boxplot com o coeficiente $k = 1.5$, MAD com o coeficiente $k = 3$ combinados com votação-método obtiveram perspectiva substancial, e em

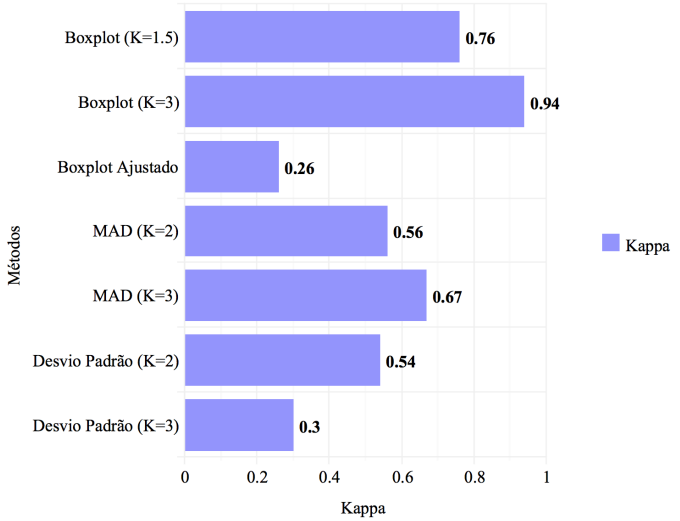

Figura 9 – Resultados do Kappa entre o juízes: Métodos e Votação-métodos

terceiro lugar o método MAD com coeficiente k=2 e votação-método com concordância moderada, o que significa dizer que o método Boxplot com o coeficiente $k = 3$ melhor atendeu as expectativas dos especialistas, diante das comparações entre os métodos automáticos e votação-método.

Nesta segunda fase de comparações foi aplicado o Kappa de Cohen entre os avaliadores e Votação-Especialistas tanto para o critério 1, quanto para o critério 2. Abaixo, Na FIGURA 10 pode-se examinar os resultados obtidos com esta comparação. Com isto pode-se avaliar que o avaliador que melhor concordou com votação-especialista utilizando o primeiro critério, foi o Avaliador 3, com um valor para o Kappa de 0.64, obtendo perspectiva substancial. Já o avaliador que melhor concordou com votação-especialista utilizando o segundo critério foi o Avaliador 2, onde obteve perspectiva substancial também, mas com um valor mais elevado para o Kappa (0.79).

Nesta terceira da fase de comparações foi aplicado o Kappa de Cohen entre os métodos e Votação-Especialistas, também para os dois critérios. Abaixo, Na FIGURA 11 pode-se examinar os resultados obtidos com esta comparação.

Com os valores representados acima pode-se avaliar que o método que melhor concordou com votação-especialista utilizando o primeiro critério, foi o método Desvio Padrão com o coeficiente $K = 3$, com um valor para o Kappa de 0.28 obtendo perspectiva justa. Já o método que melhor concordou com votação-especialista utilizando o segundo critério foi o Boxplot com coeficiente $K = 1.5$, onde obteve perspectiva moderada, com um valor para o Kappa de 0.52. Estas comparações foram a que melhor concordaram com Votação-especialista, isto significa que o nível de concordância para este caso é comedido, sendo assim, de acordo com os especialistas se mostraram razoáveis para determinar o que seria um *outlier*.

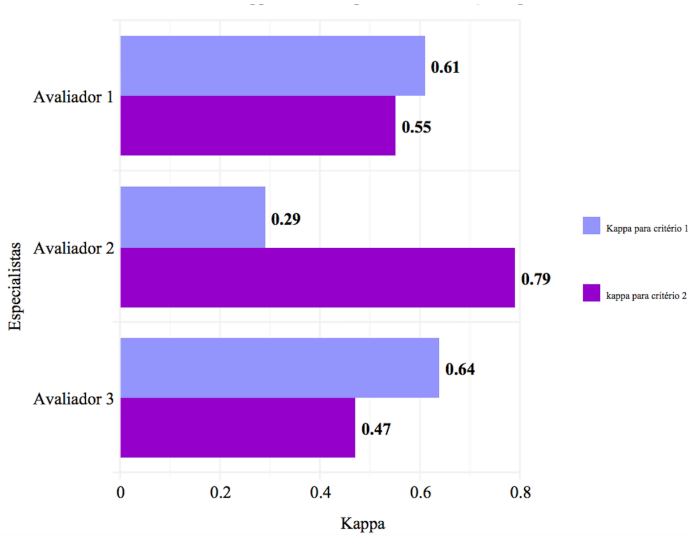

Figura 10 – Resultados do Kappa entre os juízes: Especialistas e Votação-Especialistas

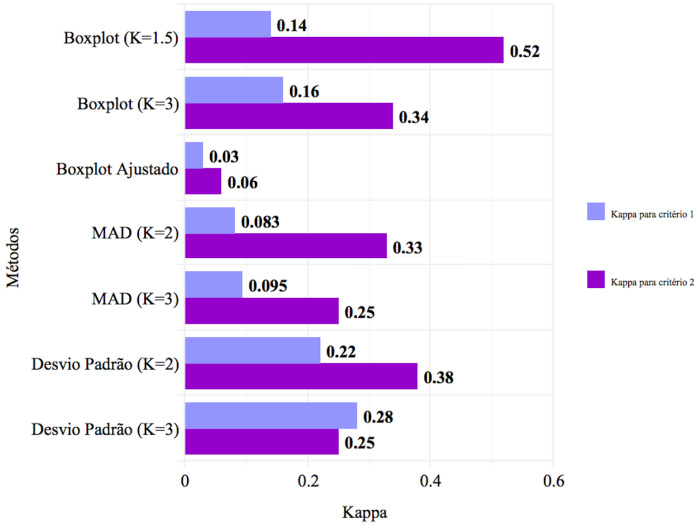

Figura 11 – Resultados do Kappa entre o juízes: Métodos e Votação-Especialistas

Em mais uma etapa de comparações, foi aplicado o Kappa de Cohen entre Votação-Especialistas de ambos os critérios e Votação-método. Abaixo, Na FIGURA 12 pode-se examinar os resultados obtidos com esta comparação.

De acordo com o resultados acima a comparação entre Votação-especialistas (critério 1) com votação-método resultou em perspectiva pobre, ou seja eles não tiveram um bom índice de concordância kappa, o que nos mostra que o resultado a partir da deliberação dos métodos automáticos todos combinados não são eficazes para identificar os outliers da forma que os especialistas esperam para este primeiro critério. Já na comparação entre Votação-especialistas (critério 2) com votação-método resultou em perspectiva justa, ou seja eles não tiveram um índice de concordância kappa muito relativo, o que nos dá espaço para estudar melhores formas de identificar

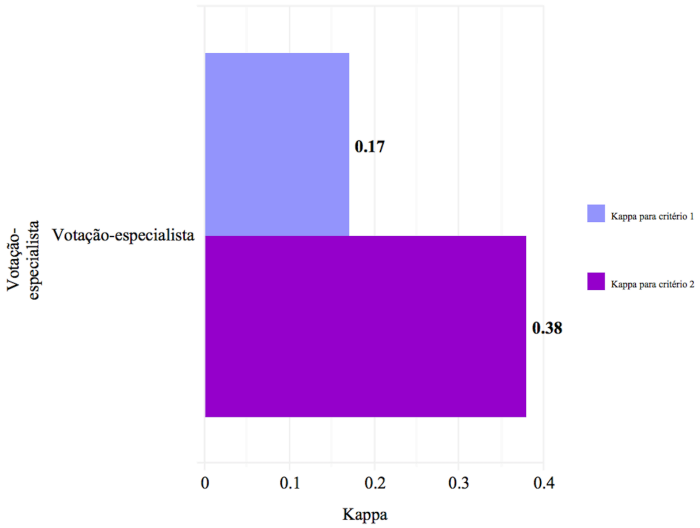

Figura 12 – Resultados do Kappa entre o juízes: Métodos e Votação-Especialistas (critério 1); Métodos e Votação-Especialistas (critério 2)

outliers.

5 Conclusão

O principal objetivo desta pesquisa foi comparar algumas técnicas de identificação de *outliers* para encontrar a que melhor se adequa na identificação de anomalias em lançamentos financeiros de empresas. Para que isto fosse alcançado, foi preciso coletar informações, fazer análise exploratória dos dados, escolher e aplicar métodos de identificação de *outliers* e, a partir daí, fazer análises comparativas que pudessem avaliar este cenário como um todo.

Com os resultados demonstrados na seção anterior, pode-se observar que ao realizar os teste de normalidade, verificou-se que os dados não seguiam uma distribuição normal e possuíam assimetria na distribuição. Devido a isso, essas características foram levadas em consideração na escolha das técnicas de identificação de *outliers* comparadas.

Na etapa seis, com a aplicação do Kappa sobre os resultados dos métodos de detecção de *outliers* e a avaliação dos especialistas foi possível afirmar que:

- O melhor dos métodos automáticos foi o Desvio padrão com coeficiente k=3, apresentando nível de concordância de 0.42 quando comparado com votação-especialista,este teve uma perspectiva de concordância maior que a comparação entre Votação-método e Votação-especialista
- A avaliação individual de cada especialista foram submetidas à comparação com votos-especialistas obtendo-se assim níveis de concordância substancial e alto.
- Na comparação de cada método com o voto-especialista, em geral foi obtido seis resultados de baixa concordância e atingiu-se apenas um resultado de concordância moderada, o que não condiz com o que os especialistas esperam.
- A concordância entre votação-método e votação-especialistas foi de baixa concordância (Kappa = 0.23), ou seja, a combinação de todos métodos automáticos também não condiz com o que os especialistas esperam.

Em suma, se pode concluir, dos resultados experimentais obtidos que em geral estas comparações demostraram uma diferença substancial entre essas duas formas de avaliação. Portanto, ainda há espaço para estudar melhores formas de identificar outliers e chegar mais próximo do que os especialistas esperam.

Algumas limitações desse trabalho foram:

- Os dados obtidos das empresas são foram amostrados de modo que permitam generalização para todos os tipos de empresa, pois foram coletados apenas das empresas que utilizam uma determinada ferramenta de gestão empresarial.
- Não foi possível definir de maneira exata o que é um *outlier*. Portanto, foi necessário a avaliação de especialistas sobre o que seria um valor incomum nas categorias estudadas.
- Na etapa 6 dessa pesquisa, optou-se por selecionar categorias onde os métodos de identificação de outliers divergiam, em vez de selecionar as categorias aleatoriamente. Essa escolha pode ter influenciado nos resultados obtidos, levando a uma menor concordância entre os métodos automáticos e os especialistas, pois as categorias descartadas podem ser consideradas mais ”fáceis”de avaliar.

5.1 Trabalhos Futuros

Como possíveis trabalhos futuros, pode-se apontar:

- Combinar os técnicas que melhor se adequaram aos dados no contexto atual deste trabalho; e fazer uma comparação mais aprofundada entre eles, e entre elas e a avaliação humana. Tal combinação servirá de base para se estabelecer em quais circunstâncias a combinação de tais métodos fornecem melhores resultados.
- Realizar um estudo específico sobre lançamentos que sigam uma distribuição normal e aplicar métodos específicos que tenham como premissa dados distribuídos normalmente.
- Propor e desenvolver um novo método híbrido que levem em conta as pontos fortes e pontos fracos de diversas técnicas de detecção de outliers. Tal abordagem tem o grande potencial de mitigar principalmente os efeitos de mascaramento e de inundação, por exemplo.
- Avaliação experimental do novo método considerando o nível de confiabilidade das técnicas quanto a sua acurácia e outras métricas de interesse.
- Expandir as avaliações mencionadas acima para uma maior quantidade de amostras, objetivando então obter resultados com maior significância estatística.

Referências

ACTION, P. *Teste de Shapiro-Wilk*. 201–? Disponível em: <http://www.portalaction.com.br/inferencia/64-teste-de-shapiro-wilk>. Citado na página 21.

ACUNA, E.; RODRIGUEZ, C. On detection of outliers and their effect in supervised classification. Porto Rico, 2004. Disponível em: <http://academic.uprm.edu/eacuna/paperout.pdf>. Citado 4 vezes nas páginas 19, 23, 24 e 26.

ANASTACIO, A. C. Análise das demonstrações contabéis e sua importância na verificação econômio-financeira das empresas. Florianópolis, Dezembro 2004. Disponível em: <http://tcc.bu.ufsc.br/Contabeis295565.pdf>. Acesso em: 02 Julho de 2018. Citado 2 vezes nas páginas 15 e 26.

BAZZOTTI, C.; GARCIA, E. A importância do sistema de informação gerencial na gestão empresarial para tomada de decisões. Ciências Sociais Aplicadas em Revista, v. 6, n. 11, 2006. Disponível em: <http://saber.unioeste.br/index.php/csaemrevista/article/view/368/279>. Acesso em: 02 Julho de 2018. Citado na página 14.

BEN-GAL, I. Outlier detection. In: MAIMON, O.; ROCKACH, L. (Ed.). *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. 2nd. ed. [S.l.]: Kluwer Academic Publishers, 2005. cap. 1. Citado 2 vezes nas páginas 15 e 26.

BRYS, G.; ROUSSEEUW, P. A robustification of independent component analysis. Journal of Chemometrics, 2005. Citado na página 25.

CHENG Ângela; MENDES, M. M. A importância e a responsabilidade da gestão financeira na empresa. Paraguai, Outubro 1989. Disponível em: <http://www.scielo.br/pdf/cest/n1/n1a02.pdf>. Acesso em: 29 Junho de 2018. Citado na página 14.

CORREA, S. M. B. B. *Probabilidade e Estatística*. 2ª edição. ed. Belo Horizonte: PUC Minas Virtual, 2003. 116 p. Citado na página 21.

COUSINEAU, D.; CHARTIER, S. Outliers detection and treatment: a review. International Journal of Psychological Research, v. 3, n. 1, p. 58–67, 2010. Citado 2 vezes nas páginas 16 e 26.

GOODDATA. *Testes de normalidade - Assimetria e curtose*. 201–? Disponível em: <https://help.gooddata.com/display/doc/Normality+Testing+-+Skewness+and+Kurtosis>. Acesso em: 19 Julho de 2018. Citado na página 21.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. [S.l.]: Morgan Kaufmann, 2012. 543–584 p. Citado 4 vezes nas páginas 15, 16, 17 e 20.

HAWKINS, D. M. *Identification of outliers*. London: Chapman and Hall, 1980. Citado na página 16.

HOFFMANN, R. *Estatistica para Economistas*. 4rd rev e ampl.. ed. São Paulo: Cengage Learning, 2011. Citado na página 19.

HUBERT, M.; VEEKEN, S. V. der. Outlier detection for skewed data. Dezembro 2007. Citado na página 26.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *biometrics*, JSTOR, p. 159–174, 1977. Citado na página 32.

LOO, M. P. J. V. der. Distribution based outlier detection in univariate data. Statistics Netherlands, 2010. Citado na página 26.

LOPES, M.; BRANCO, V. C.; SOARES, J. B. Utilização dos testes estatísticos de kolmogorov-smirnov e shapiro-wilk para verificação da normalidade para materiais de pavimentação. v. 21, p. 59–66, Junho 2013. Citado 2 vezes nas páginas 21 e 23.

PAES Ângela T. Por dentro da estatística: Análise univariada e multivariada. einstein: Educ Contin Saúde, 2018. Disponível em: <http://apps.einstein.br/revista/arquivos/PDF/1595-EC_v8n1p1-2.pdf>. Citado na página 17.

RAHMAN, S. K.; SATHIK, M.; KANNAN, K. S. A novel approach for univariate outlier. Fevereiro 2014. Citado na página 26.

ROSENMAI, P. *Using the Median Absolute Deviation to Find Outliers*. 2013. Disponível em: <http://eurekastatistics.com/using-the-median-absolute-deviation-to-find-outliers/>. Acesso em: 11 Julho de 2018. Citado na página 25.

SCIPY.ORG. *scipy.stats.kstest*. 2014. Disponível em: <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html>. Citado na página 29.

SCIPY.ORG. *scipy.stats.kurtosistest*. 2018. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosistest.html#scipy-stats-kurtosistest>. Citado na página 28.

SCIPY.ORG. *scipy.stats.shapiro*. 2018. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html#scipy.stats.shapiro>. Citado na página 29.

SCIPY.ORG. *scipy.stats.skew*. 2018. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html>. Citado na página 28.

SCIPY.ORG. *scipy.stats.skewtest*. 2018. Disponível em: <http://scipy.github.io/devdocs/generated/scipy.stats.skewtest.html?highlight=skewness>. Citado na página 28.

SCUDINO, P. A. *A Utilização de Alguns Testes Estatísticos para Análise da Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro*. 2008. Disponível em: <http://www.ufrrj.br/abelhanatureza/paginas/docs_estado/Estudomercado_mel.pdf>. Citado na página 22.

SEBRAE. *Gestão Financeira*. 201–? Disponível em: <http://www.sebraepr.com.br/PortalSebrae/artigos/Gest%C3%A3o-Financeira>. Acesso em: 26 Junho de 2018. Citado na página 14.

SEBRAE. *Controles financeiros são essenciais para a gestão do capital de giro*. 2016. Disponível em: <http://www.sebrae.com.br/sites/PortalSebrae/artigos/controles-financeiros-sao-essenciais-para-a-gestao-do-capital-de-giro,df395415e6433410VgnVCM1000003b74010aRCRD>. Acesso em: 29 Junho de 2018. Citado na página 14.

SEBRAE. *Atividade financeira: Como fazer uma análise financeira?* 2017. Disponível em: <http://www.sebrae.com.br/sites/PortalSebrae/ufs/pr/artigos/como-fazer-uma-analise-financeira,d6b1288acc58d510VgnVCM1000004c00210aRCRD>. Acesso em: 02 Junho de 2018. Citado na página 15.

SEBRAE. *Fluxo de caixa: o que é e como implantar*. 2018. Disponível em: <http://www.sebrae.com.br/sites/PortalSebrae/artigos/fluxo-de-caixa-o-que-e-e-como-implantar,b29e438af1c92410VgnVCM100000b272010aRCRD>. Acesso em: 29 Junho de 2018. Citado na página 14.

SEO, S. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. Dissertação (Master of Science) — University of Pittsburgh, Pennsylvania, Abril 2006. Citado 8 vezes nas páginas 18, 19, 20, 23, 24, 25, 26 e 30.

ZHANG, Y.; MERATNIA, N.; HAVINGA, P. A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. Enschede, Novembro 2007. Citado 6 vezes nas páginas 15, 16, 17, 18, 20 e 26.

Apêndices

APÊNDICE A – Código Fonte Utilizado nos Testes de Normalidade

A.1 Bibliotecas utilizadas

```
import pandas as pd
import xlsxwriter as xls
import numpy as np
from scipy.stats import kurtosis, shapiro, skew, kstest, iqr, stats
import pyexcel as p
```

A.2 Código para Aplicação dos Testes de Normalidade (Exemplificando Grupo 1)

```
for key, df in dfs.items():
  for categoria in df.Categoria.unique():
    if((len(df[df.Categoria==categoria])>=6) and (len(df[df.Categoria==
      #funcao(aba_grupo1, num_linhas_1, num_col, df[df.Categoria == ca
       KurtosisVar=Kurtosis(df[df.Categoria==categoria].Value)
       ShapiroVar=Shapiro(df[df.Categoria==categoria].Value)
       SkewnessVar=Skewness(df[df.Categoria==categoria].Value)
       KolmogorovVar=Kolmogorov(df[df.Categoria==categoria].Value)
```

Para melhor entendimento pode-se consultar o repositório: https://github.com/concrete-raissa-brizeno/OutliersIdentifier

A.3 Funções Utilizadas nos Testes de Normalidade

```
#Função para Curtose
def Kurtosis(df):
    if kurtosis(df, fisher=True) > 0:
        return 1
    else:
        return 0
#Função para shapiro
def Shapiro(df):
    # p-value
```

```
    if (shapiro(df)[1]) > 0.05:
        return 1
    else:
        return 0

#Função para assimetria
def Skewness(df):
    if (skew(df) >= -0.5) and (skew(df) <= 0.5):
        return 1
    else:
        return 0

#Função para Kolmogorov
def Kolmogorov(df):
    #print(kstest(df, 'norm'))
    if kstest(df, 'norm')[1] > 0.05:
        return 1
    else:
        return 0

#funções apenas para o grupo 3

#Função para Curtose (retorna tbm p-value)
def KurtosisTest(df):
    if kurtosistest(df)[1] > 0.05:
        return 1
    else:
        return 0

#Função para assimetria (retorna tbm p-value)
def SkewTest(df):
    if skewtest(df)[1] > 0.05:
        return 1
    else:
        return 0
```

Para melhor entendimento pode-se consultar o repositório: https://github.com/concrete-raissa-brizeno/OutliersIdentifier

APÊNDICE B – Código Fonte Utilizados nos Testes para Identificação de *Outliers*

B.1 Bibliotecas Utilizadas

```
import numpy as np
import pandas as pd
import xlsxwriter as xls
import xlrd as xlr
from statsmodels.stats.stattools import medcouple
import math
from openpyxl import Workbook
from openpyxl.compat import range
from openpyxl.utils import get_column_letter
from openpyxl import load_workbook
```

B.2 Funções para os Métodos de Identificar *Outliers*

```
#Desvio padrão com coeficiente 2
array_outliers = []
def desvio_padrao_test2(valores):
    sd = np.std(valores)
    media = np.mean(valores)
    outliers = []
    for x in valores:
        if x > (media + 2*sd):
            (outliers.append(x))
        if x < (media - 2*sd):
            (outliers.append(x))
    print(outliers)
    array_outliers.append(outliers)

#Desvio padrão com coeficiente 3
array_outliers = []
def desvio_padrao_test3(valores):
    sd = np.std(valores)
```

```
    media = np.mean(valores)
    outliers = []
    for x in valores:
        if x > (media + 3*sd):
            outliers.append(x)
        if x < (media - 3*sd):
            outliers.append(x)
    array_outliers.append(outliers)

#Boxplot com coeficiente 1.5
#função de boxplot pra 1.5
array_outliers = []
def boxplot_test1(valores):
    outliers = []
    q1= np.percentile(valores, 25)
    q3= np.percentile(valores,75)
    iqr = q3-q1
    for x in valores:
        if x > (q3 + 1.5*iqr):
            outliers.append(x)
        if x < (q3 - 1.5*iqr):
            outliers.append(x)
    array_outliers.append(outliers)

#Boxplot com coeficiente 3
def boxplot_test2(valores):
    outliers = []
    q1= np.percentile(valores, 25)
    q3= np.percentile(valores,75)
    iqr = q3-q1
    for x in valores:
        if x > (q3 + 3*iqr):
            outliers.append(x)
        if x < (q3 - 3*iqr):
            outliers.append(x)
    array_outliers.append(outliers)
array_outliers=[]

#Boxplot Ajustado
```

```
def boxplot_ajustado(valores):
    mc = medcouple(valores)
    Q1= np.percentile(valores, 25)
    Q3= np.percentile(valores,75)
    IQR = Q3-Q1
    outliers = []
    for x in valores:
        if mc>=0:
            if x < Q1-1.5*(math.exp(-3.5*mc))*IQR:
                outliers.append(x)
            if x > Q3+1.5*(math.exp(4*mc))*IQR:
                outliers.append(x)
        elif mc<=0:
            if x < Q1-1.5*(math.exp(-4*mc))*IQR:
                outliers.append(x)
            if x > Q1-1.5*(math.exp(3.5*mc))*IQR:
                outliers.append(x)
    array_outliers.append(outliers)

#Função para MAD com coeficiente 3
def mad_test3(valores):
    valores = np.array(valores)
    mediana = np.median(valores)
    desvios_absolutos = np.abs(valores - mediana)
    mad = np.median(desvios_absolutos)
    mad_e = 1.4826*mad
    c = (valores[((desvios_absolutos/mad_e) > 3)])
    return list(c)

#Função para MAD com coeficiente 2
def mad_test2(valores):
    valores = np.array(valores)
    mediana = np.median(valores)
    desvios_absolutos = np.abs(valores - mediana)
    mad = np.median(desvios_absolutos)
    mad_e = 1.4826*mad
    c = (valores[((desvios_absolutos/mad_e) > 2)])
    return list(c)
```