UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Uma Aplicação de Técnicas de Detecção de *Outliers* Multivariados para Identificação de Fraude Monetária

Alex Junio Souza da Silva

Ouro Preto-MG
2024

Alex Junio Souza da Silva

Uma Aplicação de Técnicas de Detecção de *Outliers* Multivariados para Identificação de Fraude Monetária

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador: Josino José Barbosa

Coorientador: Anderson Ribeiro Duarte

Ouro Preto

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S586u Silva, Alex Junio Souza da.
Uma aplicação de técnicas de detecção de outliers multivariados para identificação de fraude monetária. [manuscrito] / Alex Junio Souza da Silva. - 2024.
26 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Josino José Barbosa.
Coorientador: Prof. Dr. Anderson Ribeiro Duarte.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Estatística .

1. Papel moeda. 2. Fraude monetária. 3. DDCAM. 4. Outliers multivariados. I. Barbosa, Josino José. II. Duarte, Anderson Ribeiro. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 31

Bibliotecário(a) Responsável: Luciana De Oliveira - SIAPE: 1.937.800

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Alex Junio Souza da Silva

Uma aplicação de técnicas de detecção de *outliers m*ultivariados para identificação de fraude monetária

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em 22 de fevereiro de 2024

Membros da banca

Dr. Josino José Barbosa - Orientador (Universidade Federal de Ouro Preto)
Dr. Anderson Ribeiro Duarte - Coorientador (Universidade Federal de Ouro Preto)
Dr. Helgem de Souza Ribeiro Martins (Universidade Federal de Ouro Preto)
Dr. Fernando Luiz Pereira de Oliveira (Universidade Federal de Ouro Preto)

Prof. Dr. Josino José Barbosa, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 22/02/2024

Documento assinado eletronicamente por **Josino José Barbosa**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/03/2024, às 11:19, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0676841** e o código CRC **3175D105**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

Agradecimentos

Primeiramente, agradeço a Deus e a minha mãe por ser minha maior inspiração e fonte de força, motivando-me diariamente a buscar mais. A minha namorada, expresso profunda gratidão por estar sempre ao meu lado, oferecendo um apoio inestimável ao longo desta jornada, assim como aos meus irmãos.

Quero estender meus agradecimentos aos meus amigos. Na universidade, agradeço aos colegas que compartilharam comigo essa trajetória acadêmica e, fora da universidade, expresso minha gratidão aos amigos que fiz ao longo da vida. A cada um deles, meu sincero agradecimento por contribuírem de maneira única para a minha formação.

Em especial, agradeço aos professores Josino e Anderson pela orientação, apoio, confiança e paciência, que foram fundamentais para o meu desenvolvimento acadêmico e pessoal. À Universidade Federal de Ouro Preto, expresso minha sincera gratidão pelo suporte oferecido e pela oportunidade de crescimento.

Agradeço a todos que, de maneira direta ou indireta, contribuíram para a minha formação.

Meu muito obrigado a cada um de vocês!

Resumo

As metodologias de detecção de *outliers* multivariados possuem uma extensa gama de aplicabilidade. Novas metodologias são discutidas na fronteira do conhecimento dada sua grande aplicabilidade. Dentre as mais variadas aplicações, a identificação de fraudes financeiras e monetárias despertam grande interesse. Esse estudo apresenta algumas frentes de investigação apoiadas no assunto. Metodologias atuais de detecção de *outliers* multivariados são discutidas, tratam-se de metodologias que transcendem o nível usual de conhecimento em nível de graduação mesmo para estudantes de Estatística. Dados de fraudes monetárias ligadas à emissão de papel moeda falsificado são discutidos. Constatações e interpretações estatísticas associadas à interpretação de resultados das metodologias de detecção de *outliers* são apresentadas. Estas ferramentas estatísticas são de grande valia para tentar elucidar essas fraudes. Os valores *outliers* são elementos usualmente incomuns ao conjunto de dados, em geral valores extremos quanto a ordem de grandeza das variáveis. Diversos métodos para detecção de *outliers* já são bastante difundidos na literatura, mas as investigações em *outliers* multivariados ainda estão em pleno estudo. Metodologias com este propósito são apresentadas na revisão de literatura, em particular, a metodologia *Data-driven Cluster Analysis Method* (DDCAM) que é apresentada e aplicada aos dados em estudo.

Palavras-chave: Papel moeda, Fraude monetária, DDCAM, *outliers* multivariados.

Abstract

Multivariate outlier detection methodologies have a wide range of applicability. Given their broad applicability, new methods are discussed at the frontier of knowledge. Identifying financial and monetary fraud is very interesting among the most varied applications. This study presents some research fronts supported by the subject. Current methodologies for detecting multivariate outliers are discussed; these are methodologies that transcend the usual level of knowledge at the undergraduate level, even for Statistics students. Data on monetary fraud linked to the issuance of counterfeit paper-money are discussed. Statistical findings and interpretations are associated with interpreting results from outlier detection methodologies. These statistical tools are of great value in trying to elucidate these frauds. Outliers values are usually unusual elements in the data set, generally extreme values in terms of the order of magnitude of the variables. Several methods for detecting outliers are already widespread in the literature, but investigations into multivariate outliers are still under study. The literature review presents methods for this purpose, particularly the methodology Data-driven Cluster Analysis Method (DDCAM), which is presented and applied to the data under study.

Keywords: Paper-money, Monetary fraud, DDCAM, multivariate outliers.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Em problemas de interesse prático que envolvem conjuntos de dados, em algumas situações específicas, a presença de observações demasiadamente discrepantes com respeito às outras observações é verificada. As observações que atendem à essa condição, usualmente são denominados como valores *outliers*. Barbosa, Duarte e Martins (2020) [2] afirmam que um *outlier* é uma observação que destoa do padrão usual dos demais elementos do conjunto de dados.

A detecção dessas observações incomuns pode ter um valor muito importante na análise de dados. Esses elementos discrepantes podem contribuir significativamente para obter informações não observadas em uma análise estatística puramente simples. A aplicação de modelos estatísticos convencionais a conjuntos de dados que contenham valores discrepantes pode levar a resultados inadequados e decisões incorretas. Em muitas situações práticas, os próprios *outliers* são de interesse especial e a sua identificação pode ser o objetivo central da análise.

Os *outliers* são valores atípicos em conjuntos de dados, que divergem significativamente da maioria das observações. Sua existência é relevante por diversos motivos: podem indicar erros nos dados, distorcer análises estatísticas, influenciar modelos e impactar decisões. No entanto, vale ressaltar que *outliers* também podem conter informações valiosas e únicas. A identificação e interpretação cuidadosa de valores *outliers* podem revelar *insights* valiosos e contribuir para a descoberta de eventos excepcionais ou tendências ocultas.

Em termos estatísticos, um *outlier* é uma observação que se encontra significativamente distante do restante do conjunto de dados. Essa discrepância é geralmente definida com base em critérios estatísticos, como uma diferença em relação à média, mediana ou alguma outra medida do conjunto de dados. Os *outliers* são valores atípicos que podem indicar erros de medição, variações extremas nos dados ou situações excepcionais.

Neste contexto é possível abordar dois tipos de observações *outliers*, as univariadas e as multivariadas que se distinguem pela forma como são identificadas e interpretadas em relação aos dados. Os *outliers* univariados ocorrem em espaço unidimensional e são detectados por meio de métodos estatísticos simples que não consideram as correlações entre variáveis. Por outro lado, os *outliers* multivariados são observações atípicas identificados por meio da consideração do efeito das interações entre várias variáveis. Diante disso, exigem métodos estatísticos mais sofisticados. Os *outliers* multivariados são particularmente úteis para identificar anomalias que não seriam evidentes ao exa-

minar cada variável isoladamente. Além disso, são valiosos em aplicações em que as interações entre variáveis desempenham um papel de relevância.

A detecção de *outliers* é uma técnica importante para diversas aplicações. Ela identifica valores discrepantes nos dados, que podem ser causados por erros de medição, falhas nos sistemas ou eventos anormais. Segundo Aggarwal (2017) [3], na maioria das aplicações, os dados são gerados por um ou mais processos distintos. Quando um desses processos se comporta de maneira incomum, pode gerar *outliers*. Portanto, um *outlier* pode conter informações úteis sobre as características anormais dos sistemas ou entidades que impactam no processo de geração de dados.

A presença de *outliers* em conjuntos de dados pode acarretar efeitos significativos e influenciar sobremaneira a interpretação e a confiabilidade das análises realizadas. Embora a identificação de *outliers* seja algumas vezes associada à necessidade de remoção destes dos conjuntos de dados, é fundamental reconhecer que esses valores atípicos podem conter informações valiosas para a análise em uma forma mais ampla. Em alguns casos, *outliers* podem indicar eventos excepcionais, revelar padrões não detectados anteriormente ou fornecer *insights* únicos. Particularmente no contexto de análise de dados do mercado financeiro a detecção de padrões incomuns tem um papel crucial na prevenção de fraudes e minimização de prejuízos.

1.1 Motivação

Os estudos associados com a detecção de *outliers* multivariados tem vasta aplicabilidade. A possibilidade de aprofundar estudos nesse tipo de metodologia já é, por si só, uma motivação significativa em um nível de cursos de graduação em Estatística. Além disso, a investigação do comportamento dessas metodologias em casos reais de dados sem a garantia prévia de normalidade são preponderantes para afiançar a qualidade dos métodos.

1.2 Objetivos

Este estudo tem objetivos de caráter científico e aplicado. As investigações sobre os métodos detecção de *outliers* multivariados tem interesse científico explícito em estudos de diversas áreas. Por outro lado, a aplicação em conjuntos de dados específicos e instâncias de teste mostram uma visão aplicada e corroborão a adequação das metodologias em estudo.

1.2.1 Objetivos Gerais

Discutir os aspectos de concepção e utilização de métodos detecção de *outliers* multivariados

1.2.2 Objetivos Específicos

i. utilizar a linguagem LaTeX, que é padrão na confecção de textos estatísticos em vários níveis de pesquisa;

ii. apresentar uma revisão bibliográfica que direciona para os assuntos da proposição e utilização de métodos detecção de *outliers* multivariados;

iii. estabelecer comparações de resultados entre metodologias por meio de um mesmo conjunto de dados aplicados.

Este texto é organizado da seguinte forma, inicialmente o capítulo introdutório menciona aspectos de pesquisa, bem como o delineamento prévio de objetivos de estudo. Em seguida, o segundo capítulo apresenta uma revisão conceitual e bibliografia sobre esse tema de pesquisa. O capítulo de Aspectos Metodológicos detalha os métodos detecção de *outliers* multivariados. O quarto capítulo apresenta de forma mais detalhada todo o conjunto de resultados alcançados. Por fim, o último capítulo apresenta as conclusões alcançadas com a utilização dessa investigação e também propostas de continuidade desse estudo.

2 Fundamentação Teórica

Existe um leque vasto de estratégias e metodologias voltadas para problemas focados na detecção e identificação de *outlier* multivariados. Este estudo não possui a pretensão de estabelecer uma revisão de literatura extensiva. Entretanto, uma revisão que contextualize e direcione para os estudos que serão abordados é preponderante.

A maior parte dos estudos metodológicos para detetcção de *outliers* multivariados foram desenvolvidos baseados na clássica distância de Mahalanobis. Rousseeuw e van Zomeren (1990) [4] apresentaram um método baseado no estimador robusto de volume mínimo de elipsóide (MVE). O estimador MVE pode ser obtido por meio do elipsóide de menor volume que cobre pelo menos k pontos da amostra, com $n/2 < k < n$.

Filmoser (2005) [5] e Filzmoser, Garrett e Reimann (2005) [6] introduziram um método baseado no estimador robusto do determinante de covariância mínima (MCD). O estimador MCD é obtido por um subconjunto que minimiza o determinante da matriz de covariâncias amostral. Este subconjunto tem tamanho determinado por uma janela h, com $n/2 < h < n$. A média desses h pontos é a estimativa de locação, e a estimativa de escala é proporcional à sua matriz de covariância. A medida da janela h determina a robustez do estimador.

O estudo proposto por Barbosa, Pereira e Oliveira (2018) [7] apresenta uma estratégia através de análise de *clusters* para detecção multivariada de *outliers*, nominada método por análise de *clusters* (CAM). Nessa abordagem foi utilizado um método de agrupamentos k-means para agrupar indivíduos semelhantes. Entretanto, uma deficiência marcante da metodologia reside na escolha ad-hoc para o número de clusters no procedimento k-means. A investigação porposta por Barbosa, Duarte e Martins (2020) [2] elucida esta deficiência com um conjunto experimental abrangente para diferentes valores de k.

Cerioli (2010) [8] apresentou o conjunto abrangente de testes multivariados de detecção de *outliers* baseados no estimador MCD. Além disso, o estudo apresenta uma aproximação para a distribuição robusta de distâncias e introduz o FSRMCD, um procedimento baseado na distância robusta de Mahalanobis. Finalmente, uma nova etapa de iteração é adicionada ao procedimento FSRMCD, gerando o procedimento MCD iterado e reponderado denominado IRMCD. FSRMCD e IRMCD são considerados mais poderosos em conjuntos contaminados do que testes semelhantes [9, 10]. O método IRMCD apresenta maior poder de detecção para amostras com maior nível de contaminação, enquanto o método FSRMCD está mais interessado em controlar falsos outliers.

Já os autores Jobe e Pokojovy (2015) [11] utilizaram os estimadores de regra

MCD reponderada de amostra finita (FSRMCD) para identificar uma largura de banda para estimar não parametricamente a densidade de probabilidade multivariada e, em seguida, identificar máximos locais por meio da aplicação do algoritmo *Modal Expectation Maximization* (MEM) (Li; Ray e Lindsay, 2007 [12]). Os dados de amostra associados ao mesmo máximo local são agrupados no mesmo cluster. Essa técnica de clustering é chamada de *Modal Association Clustering* (MAC). Jobe e Pokojovy (2015) [11] propuseram o método FSRMCD-MAC com esta estratégia de clusterização. Os métodos aqui apresentados constituem uma revisão básica da literatura sobre o assunto, os autores Patel, Kapoor, Sharma e Chakrabarti (2023) [13] apresentam uma revisão mais aprofundada da literatura geral sobre métodos de detecção de *outliers* multivariados em diversos contextos.

3 Abordagem do Problema e Aspectos Metodológicos

Esta investigação está particularmente direcionada para a aplicabilidade de uma técnica inovadora para o problema de detecção de valores *outliers* em dados multivariados. O método em questão é o *Data-driven Cluster Analysis Method* (DDCAM) [14]. O DDCAM parte de uma técnica anterior baseada em análise de agrupamentos o método nominado *Cluster Analysis Method* (CAM) [7]. O propósito metodológico desse material é descrever as metodologias CAM e DDCAM para posterior aplicação em um problema realista de utilização de metodologias de detecção de *outliers* multivariados. Incialmente, antes das descrições dos métodos CAM e DDCAM, é importante apresentar o método de agrupamento *K-means*. Posteriormente os métodos CAM e DDCAM serão descritos, além disso outros métodos com o mesmo propósito de detecção de *outliers* multivariados serão discutidos.

3.1 O Algoritmo *K-means*

O clássico algoritmo *K-means* se enquadra na tipologia de metodologias de aprendizado de máquina. Na hierarquia do aprendizado, destaca-se o aprendizado indutivo que é efetuado a partir de exemplos externos (coletados). O aprendizado indutivo se divide em aprendizado supervisionado e aprendizado não supervisionado.

Para problemas sem uma base de dados já rotulada é necessária a aplicação de técnicas não supervisionadas. No aprendizado não supervisionado o próprio conjunto de dados é utilizado com interesse de gerar aprendizado para otimização o funcionamento da técnica em uso. O algoritmo *K-means* é uma estratégia clássica de aprendizado não supervisionado. O algoritmo toma um conjunto original de dados e gera k agrupamentos especificos entre os dados em estudo.

O *K-means* é um algoritmo de agrupamento (*clustering*) amplamente utilizado em análise de dados e mineração de dados. É um algoritmo projetado para agrupar um conjunto de dados não rotulado em grupos distintos, em que cada grupo é denominado por *cluster*. O objetivo do algoritmo *K-means* é dividir os elementos do conjunto de dados em *clusters* de modo que os elementos dentro do mesmo *cluster* sejam mais semelhantes entre si do que com os pontos dos demais *clusters*.

Visto por outro ângulo, o agrupamento *K-means* é um método de separar elementos em torno de centros (centroídes). Em suma, é o efeito análoga da Química

denominado *clustering* que gera o efeito de particionar n elementos entre k grupos em que cada observação pertença ao grupo ao qual mais se aproxima da média do grupo. Isso resulta em uma divisão do espaço de dados em um Diagrama de Voronoi (veja a Figura 1.

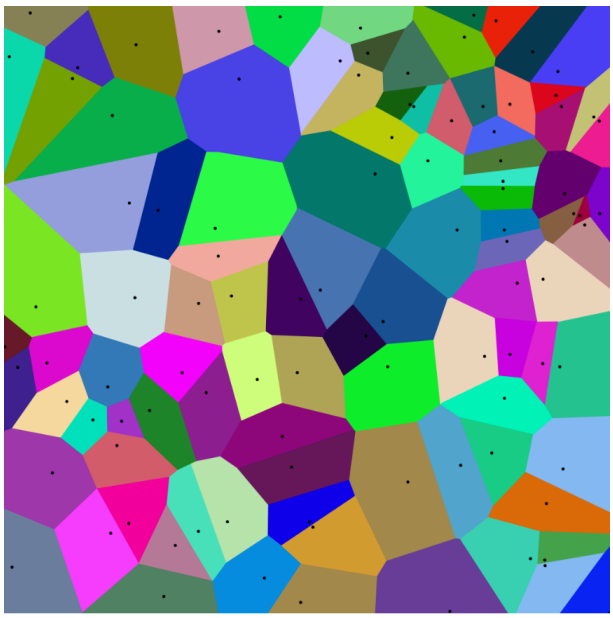


Figura 1 – Ilustração de um Diagrama de Voronoi hipotético.

Um Diagrama de Voronoi é um tipo de decomposição de um dado espaço em estudo, por exemplo, um espaço métrico, determinado pela distância entre objetos no espaço. Foi nomeado em decorrência dos trabalhos de Georgy Voronoi posteriormente, e também chamado de Tesselagem de Voronoi, ou um Mosaico de Dirichlet (em homenagem a Lejeune Dirichlet). Diagramas de Voronoi podem ser encontrados em diversos campos da ciência e tecnologia, até mesmo na arte, e tem inúmeras aplicações práticas e teóricas como pode ser verificado no estudo de Du; Faber e Gunzburger (1999) [15].

A proximidade entre elementos pode ser determinado por qualquer métrica de distância, ou seja, distância Euclidiana, distância de Manhattan, entre muitas outras. Diferentes métricas de distâncias conduzem para diferentes agrupamentos. O método *K-means* pode ser descrito por meio da execução de quatro etapas bem definidas:

- inicialização - o algoritmo parte da escolha de um valor K para a quantidade de clusters, em seguida, seleciona K pontos aleatórios (dentre os dados) como os centros iniciais dos *clusters* (centroides);
- atribuição - cada elemento dentre os dados é atribuído ao *cluster* cujo centroide é mais próximo com base na métrica de distância em uso;
- atualização - os centroides de cada *cluster* são recalculados como a média de todos os pontos atribuídos a esse *cluster*;

- repetição - os passos de atribuição e atualização são repetidos até que os centroides não mudem significativamente ou até que um critério de parada seja atingido (como um número máximo de iterações).

O algoritmo *K-means* é eficaz na identificação de estruturas de *clusters* em dados, é útil em diversas aplicações, como segmentação de clientes, análise de padrões em imagens, classificação de documentos entre muitas outras aplicações. No entanto, existem algumas limitações, como a sensibilidade à escolha inicial dos centroides e a necessidade de especificar o número de clusters (K) antecipadamente. Existem variações e extensões do algoritmo *K-means* que abordam algumas dessas limitações.

Particularmente para esta aplicação, a métrica de distância utilizada será a distância Euclidiana. A distância Euclidiana é uma métrica de distância comumente usada na análise de dados e no cálculo de distância entre pontos em espaços euclidianos. Ela mede a distância linear entre dois pontos em um espaço multidimensional. A fórmula para calcular a distância Euclidiana entre dois pontos, geralmente denotados como A e B, em um espaço N-dimensional, é a seguinte:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + ... + (z_2 - z_1)^2} \tag{3.1}$$

em que $x_1, y_1, \ldots, z_1$, são as coordenadas do primeiro ponto (A) e $x_2, y_2, \ldots, z_2$, são as coordenadas do segundo ponto (B).

Em resumo, a distância euclidiana calcula o comprimento da linha reta que liga dois pontos em um espaço multidimensional. Essa métrica é frequentemente usada em algoritmos de agrupamento, como o *K-means*, na classificação de objetos, na análise de similaridade e em diversas outras aplicações na análise de dados.

3.2 O Método CAM

O método de identificação de *outliers* multivariado, nominado CAM (*Cluster Analysis Method*), utiliza uma abordagem baseada em técnicas de agrupamento para detectar *outliers* em conjuntos de dados multivariados. O procedimento central do CAM emprega o método *K-means* para realizar a análise de agrupamentos.

Ao contrário do método tradicional *K-means*, o CAM determina o número de grupos (k) de forma diferente, por meio da regra $k = {}^{n}/_{10}$, em que n é o tamanho da amostra disponível. Isso implica que os indivíduos são divididos em grupos com base nessa regra específica. A escolha inicial dos centroides, pontos representativos de cada grupo, é feita selecionando k valores aleatórios. Para garantir a consistência dos resultados, a semente do processo aleatório *K-means* é fixada.

Uma vez formados os grupos, calcula-se o centroide de cada grupo, como o ponto médio dos indivíduos dentro desse grupo. Em seguida, é realizada a análise de *outliers* por meio do cálculo da distância euclidiana entre o centroide de cada grupo e a mediana dos dados. A distância euclidiana é uma medida de quão distante o centroide está da mediana.

O critério estabelecido para definir se um grupo de indivíduos é um grupo de *outliers* é baseado no desvio padrão amostral (s) das distâncias entre os centroides dos grupos e a mediana dos dados. Se a distância euclidiana entre o centroide de um grupo e a mediana dos dados for superior a 2,5 vezes o desvio padrão ($2{,}5 \times s$), esse grupo é considerado um grupo *outlier*. Assim, o CAM fornece uma metodologia que utiliza *K-means* de maneira específica e um critério baseado em distâncias para identificar eficientemente *outliers* multivariados em conjuntos de dados.

3.3 O Método DDCAM

O método DDCAM (acrônimo para *Data-driven Cluster Analysis Method*) parte de premissas iniciais associadas à aplicação de técnica predecessora CAM e acrescenta uma fator adaptativo para metodologia, que se ajusta de acordo com informações inerentes aos dados, daí a nomenclatura *Data-driven*. O DDCAM se destaca por integrar técnicas de análise de agrupamentos com um mecanismo adaptativo para a escolha do número de *clusters*, proporcionando uma detecção eficiente de *outliers* em conjuntos de dados complexos.

A metodologia do DDCAM é composta por vários estágios, que contribuem para a precisão na detecção de *outliers*. A seguir, são detalhados os principais passos de execução do DDCAM:

- estimação do valor δ - define-se δ como a proporção máxima de valores admissíveis como *outliers* no conjunto de dados, um estimador eficiente para δ (aqui dito $\hat{\delta}$) é obtido através da análise univariada do conjunto de dados, e considera médias e desvios padrão amostrais;
- processo de refinamento I - diferentes possíveis valores de k até um valor máximo (k_{max}) definido pela divisão da quantidade total de dados pelo logaritmo dessa quantidade, a validade dos agrupamentos é verificada através da restrição de que o menor grupo não deve ter mais que $\hat{\delta} \times n$ elementos;
- processo de refinamento II - refina ainda mais o conjunto de valores de k ao considerar a distância dos centroides à mediana dos dados e o número máximo de elementos em um agrupamento, os valores de k que não atendem simultaneamente a essas condições são descartados;

- busca pelo valor adequado k - o critério de informação Bayesiano (BIC) é utilizado para escolher o valor mais adequado para k dentro do conjunto refinado nos dois processos anteriores, o BIC é usado para penalizar modelos mais complexos, e evitar sobreajuste; o BIC auxilia a escolha do número ideal de clusters.

Em resumo, o DDCAM é uma abordagem sofisticada que, por meio de uma combinação de técnicas estatísticas e adaptativas, busca identificar de maneira precisa e eficaz os valores atípicos em conjuntos de dados complexos. A escolha adaptativa do número de *clusters* e a consideração da distância dos centroides à mediana dos dados são aspectos distintivos que contribuem para a robustez e eficiência do método na detecção de *outliers*.

Essas fases constituem a reparametrização do método CAM. Posteriormente, o procedimento CAM é executado com a parametrização definida nas quatro etapas anteriores.

3.4 O Método FSRMCD

O método FSRMCD (acrônimo para *Finite Sample Reweighted Minimum Covariance Determinant*) [8] surge como uma abordagem robusta para identificar observações atípicas em conjuntos de dados. Este método integra a robustez do clássico *Minimum Covariance Determinant* (MCD) com uma estratégia de reponderação, isso o torna eficaz em amostras pequenas e em situações em que a presença de *outliers* compromete a análise estatística. O FSRMCD aborda a questão do controle do tamanho, que se refere à capacidade do método de manter uma taxa de falsos positivos (rejeição incorreta da hipótese nula) em níveis aceitáveis. Esse controle é vital para garantir a confiabilidade das conclusões estatísticas. O método é projetado para manter o controle sobre o número de falsos *outliers*.

O MCD é conhecido por sua robustez na estimação da matriz de covariância, serve como fundamentação para o FSRMCD. Essa estimativa é menos sensível à presença de outliers do que a covariância tradicional, proporciona uma representação mais fiel da variabilidade nos dados. O diferencial do FSRMCD é a incorporação de uma estratégia de reponderação, projetada para lidar efetivamente com amostras pequenas. A reponderação visa reduzir o impacto dos *outliers*, e garantir que sua influência não seja amplificada durante a estimação da matriz de covariância. Isso é particularmente crucial em situações em que o número de observações é limitado. Os passos de execução do método FSRMCD são:

- estimação do MCD inicial - o método calcula o *Minimum Covariance Determinant* (MCD) por meio do algoritmo FAST-MCD de Rousseeuw e Van Driessen (1999)

[16];

- determinação do subconjunto MCD - com base no MCD inicial, o método determina um subconjunto robusto (h) que representa uma fração alta de observações centrais;
- reponderação MCD - o FSRMCD realiza uma reponderação dos dados, em que observações com maior desvio são atenuadas, isso reduz a influência de *outliers* na estimativa da matriz de covariância;
- simultaneidade ajustada - o método ajusta a estatística de teste por meio de uma distribuição Beta para garantir simultaneidade e permitir comparações significativas.

3.5 O Método IRMCD

O método IRMCD (acrônimo para *Iterated Reweighted Minimum Covariance Determinant*) [8] representa uma abordagem inovadora para esse desafio, integra a robustez do MCD com iterações adicionais e reponderação estratégica. Essa abordagem visa combinar a eficácia da detecção de *outliers* com a adaptabilidade a situações de contaminação nos dados. O IRMCD estende o MCD ao adicionar uma etapa adicional de iteração ao processo de detecção de *outliers*. Essa iteração visa melhorar a capacidade do método de identificar observações atípicas, especialmente em situações em que a contaminação pode não ser totalmente capturada na primeira iteração. Os passos de execução do método IRMCD são:

- estimação do MCD inicial - o método calcula o *Minimum Covariance Determinant* (MCD) por meio do algoritmo FAST-MCD de Rousseeuw e Van Driessen (1999) [16];
- iteração adicional - após a estimativa inicial, o método realiza uma iteração adicional para melhorar a sensibilidade à detecção de *outliers*;
- reponderação adicional - além da reponderação realizada no MCD, o IRMCD aplica uma estratégia de reponderação adicional durante a iteração, para ajustar a influência de observações atípicas;
- simultaneidade ajustada para testes individuais - o método ajusta as estatísticas de teste para cada hipótese individual com o uso de uma distribuição Beta escalada, e garantir o controle simultâneo durante testes múltiplos;
- aceitação ou rejeição da hipótese global - com base nas estatísticas ajustadas, o método decide aceitar ou rejeitar a hipótese global de ausência de outliers.

O IRMCD mantém um controle efetivo do tamanho, garante que a taxa de falsos positivos permaneça dentro de limites aceitáveis. Esse controle é crucial para preservar a integridade das análises estatísticas, especialmente em amostras pequenas. Os resultados de uso destacam a capacidade do IRMCD de melhorar a detecção de *outliers* em situações de contaminação intensa.

Com a exposição de toda esta metodologia, esse estudo complementa o objetivo de discussão de metodologias para detecção de *outliers* multivariados. Além disso, habilita para a análise de dados aplicados e comparações entre as metodologias que é o foco central do capítulo subsequente desse estudo.

4 Resultados Alcançados

Esta monografia aborda os resultados obtidos a partir da aplicação de técnicas de detecção de *outliers* multivariados em dados provenientes do setor bancário suíço (*Swiss bank*). O presente trabalho inicia-se com uma descrição da obtenção do dados, seguida por uma análise detalhada das variáveis em questão. Posteriormente, realiza-se uma abordagem analítica descritiva, com interesse em aprofundar na caracterização estatística dos dados. Por fim, são aplicados métodos específicos de detecção de *outliers* na base de dados, o que resulta em uma avaliação crítica e a interpretação dos resultados obtidos.

4.1 Dados

Os dados utilizados neste estudo foram obtidos por meio da utilização do pacote *mclust* [17]. Todo o processo de extração, tratamento e limpeza dos dados foi conduzido na linguagem R, com a utilização do ambiente de desenvolvimento *RStudio* versão 4.1.3. A base de dados bruta é composta por 200 observações, das quais 100 são classificadas como notas genuínas e as outras 100 como notas falsificadas. Para este estudo foram selecionadas todas notas genuinas e por meio de um sorteio aleatorio cinco notas falsificadas, com o objetivo de induzir a presença de *outliers*. Cada observação compõe-se de seis variáveis de interesse, conforme mostrado na Figura 2.

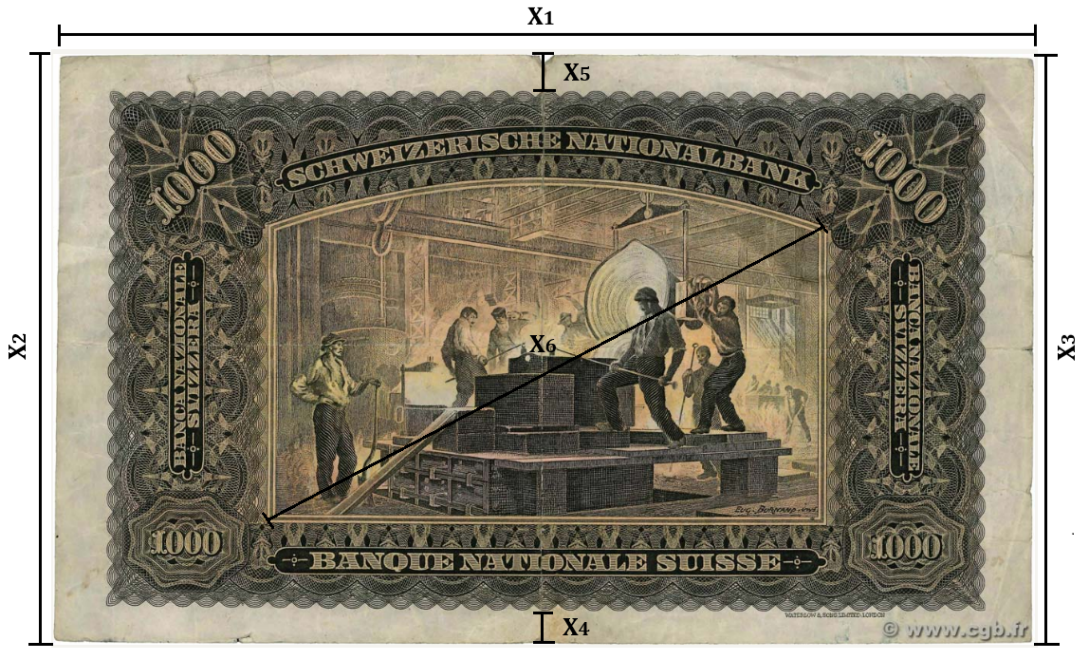


Figura 2 – Cédula Antiga de 1000 Francos Suiços

A Tabela 1 apresenta a descrição das variáveis que compõe o conjunto de dados sob investigação.

Tabela 1 – Variáveis Componentes do Banco de Dados.

Variável	Descrição
X_1	Comprimento da nota.
X_2	Largura da borda esquerda.
X_3	Largura da borda direita.
X_4	Distância do quadro interno para a borda inferior.
X_5	Distância do quadro interno para a borda superior.
X_6	Comprimento da diagonal da imagem central.

4.2 Análise Descritiva

A análise descritiva é preponderante para compreensão e interpretação de conjuntos de dados. É um procedimento que oferece uma visão detalhada das características fundamentais das variáveis em estudo. A Tabela 2 apresenta uma análise das principais estatísticas descritivas (mínimo, mediana, média, máximo, desvio padrão) associadas às variáveis nas notas genuínas e falsas desse estudo.

Tabela 2 – Estatísticas Descritivas das Variaveis Presentes nas Notas Genuínas e Falsas

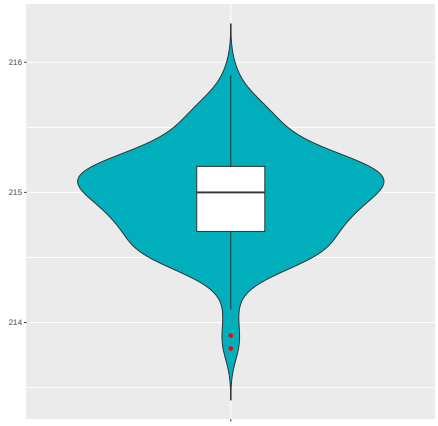
Variável	Mínimo		Mediana		Média		Máximo		Desvio Padrão	
	Genuína	Falsa	Genuína	Falsa	Genuína	Falsa	Genuína	Falsa	Genuína	Falsa
X_1	213,8	214,5	215,0	215,0	215,0	215,0	215,9	215,5	0,387	0,384
X_2	129,0	130,2	129,9	130,2	129,9	130,4	131,0	130,7	0,364	0,248
X_3	129,0	129,9	129,7	130,3	129,7	130,2	131,1	130,4	0,355	0,194
X_4	7,2	8,2	8,3	10,2	8,3	10,5	10,4	12,3	0,642	1,624
X_5	7,7	10,2	10,2	11,8	10,2	11,32	11,7	11,9	0,648	0,746
X_6	139,6	137,8	141,5	139,6	141,5	139,3	142,4	140,0	0,447	0,866

Ao observar as variáveis X_1, X_2 e X_3 é possível perceber que as notas genuínas e falsas exibem similaridade nos valores mínimos, medianos e máximos. No entanto, ao analisar as variáveis X_4, X_5 e X_6, diferenças mais expressivas são identificadas.

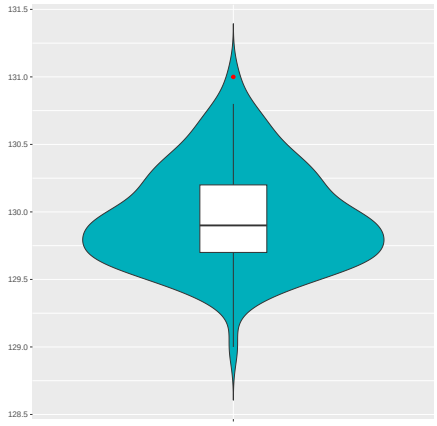
As variáveis X_4 e X_5 revelam algumas discrepâncias significativas. As notas falsas tendem a apresentar valores elevados para essas variáveis. A variável X_6 também destaca uma diferença na média, o que sugere uma característica distintiva potencialmente relevante nas notas falsas.

A análise das médias revela que, para X_4 e X_5, as notas falsas exibem valores ligeiramente superiores, o que indica uma tendência consistente para essas variáveis.

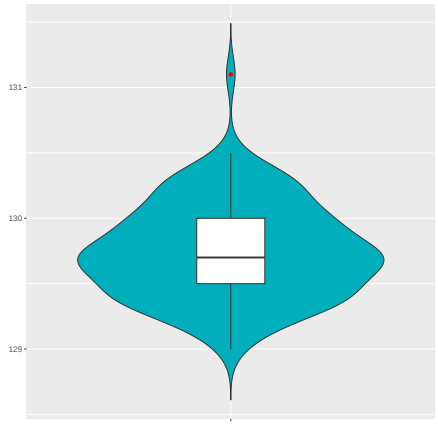
Em relação ao desvio padrão, X_4 se destaca com uma diferença substancial ao apresentar maior variabilidade nas notas falsas para esta variável específica. Uma representação que oferece uma visão abrangente da distribuição e estatísticas resumidas das variáveis pode ser vista na Figura 3.



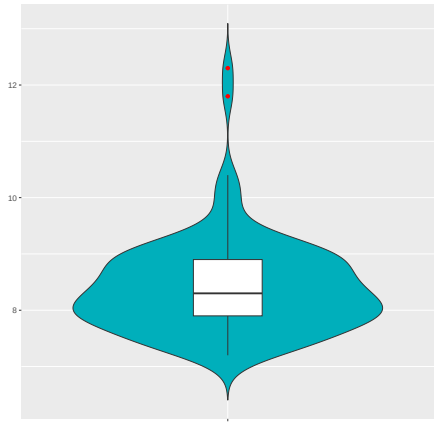
(a) variável X_1



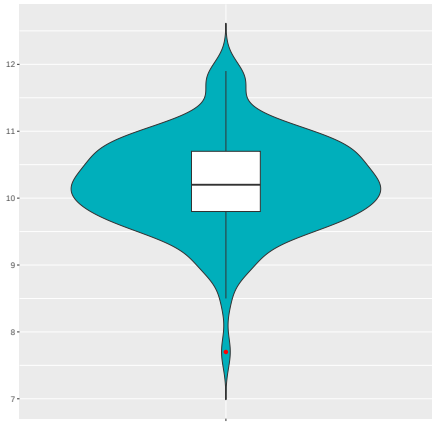
(b) variável X_2



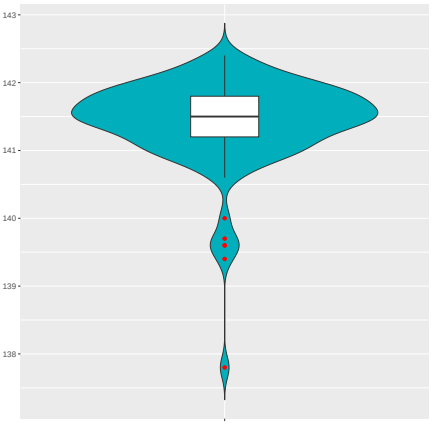
(c) variável X_3



(d) variável X_4



(e) variável X_5



(f) variável X_6

Figura 3 – Apresentação Gráfica para as Variáveis em Estudo

O gráfico usualmente nominado violino (do inglês *violin plot*) é uma alternativa adequada para a representação desse tipo de análise. O gráfico violino é um gráfico que apresenta uma distribuição ajustada para os dados através de uma curva de densidade estimada. Além disso, esse gráfico inclui em seu interior um *boxplot*.

Ao analisar as densidades ajustadas, efeitos de assimetria são visualizados em todas as variáveis, em geral estes efeitos decorrem da presença de *outliers* univariados identificados nos gráficos boxplot. Apesar do efeito ser visualizado em todas as variáveis, trata-se de um efeito mais evidenciado nas variáveis X_4, X_5, e X_6. Particularmente o efeito é bastante significativo na variável X_6.

Dentro de cada densidade ajustada, os boxplots fornecem informações sobre as medidas estatísticas usuais. As medianas estão posicionadas centralmente quanto ao primeiro e terceiro quartil nas variáveis X_4, X_5, e X_6, entretanto este efeito não se repete nas variáveis X_1, X_2, e X_3. Não existe um padrão claro acerca de identificação de *outliers* univariados de máximo ou de mínimo. As variáveis X_1, X_5, e X_6 apresentam valores extremos univariados mínimos ao passo que as demais variáveis apresentam valores extremos univariados máximos.

4.3 Aplicação de Métodos para Identificação de *Outliers* Multivariados

Com interesse em avaliar a eficiência e aplicabilidade dos métodos DDCAM, FSRMCD e IRMCD, abordados neste estudo, os mesmos foram utilizados para análise na base de dados mencionada anteriormente. O interesse central é mensurar sua adaptabilidade e precisão na detecção de notas falsas, que serão consideradas neste estudo como *outliers*. A base de dados processada é constituída pelas 100 primeiras notas autênticas do banco de dados, e ampliada com inclusão de cinco notas falsas selecionadas aleatoriamente. As notas fraudulentas foram incorporadas à base de dados original. Os índices associados à essas notas falsificadas foram determinados por meio de uma escolha completamente aleatorizada que resultou no índices 105, 109, 122, 128 e 180, dentre a sequência indexada de 101 até 200 com as notas falsas.

Inicialmente, na condução dos experimentos com os métodos FSRMCD e IRMCD para a detecção de possíveis notas falsas, verificou-se a detecção de 12 notas como fraudulentas dentre todas as analisadas. Notavelmente, ambos os métodos demonstraram a habilidade de identificar corretamente todas as notas que eram efetivamente falsificadas dentre as avaliadas. No entanto, foi observada uma classificação incorreta para 7 notas autênticas. A Tabela 3 apresenta detalhes sobre a identificação das notas classificadas como falsas.

Tabela 3 – Notas Identificadas como Falsas pelos Métodos FSRMCD e IRMCD

Índice	Legitimidade	X_1	X_2	X_3	X_4	X_5	X_6
1	Verdadeira	214,8	131,0	131,1	9,0	9,7	141,0
5	Verdadeira	215,0	129,6	129,7	10,4	7,7	141,8
13	Verdadeira	215,2	130,8	129,6	7,9	10,8	141,4
40	Verdadeira	213,9	130,3	129,0	8,1	9,7	141,3
57	Verdadeira	215,7	130,0	129,4	9,2	10,4	141,2
70	Verdadeira	214,9	130,2	130,2	8,0	11,2	139,6
71	Verdadeira	213,8	129,8	129,5	8,4	11,1	140,9
105	Falsa	214,7	130,2	130,3	11,8	10,9	139,7
109	Falsa	215,0	130,2	129,9	10,0	11,9	139,4
122	Falsa	215,1	130,6	130,3	12,3	10,2	139,6
128	Falsa	215,5	130,7	130,3	10,2	11,8	140,0
180	Falsa	214,5	130,2	130,4	8,2	11,8	137,8

Posteriormente o mesmo experimento foi conduzido com a utilização do método DDCAM na base de dados em análise e ocorreu a classificação de seis notas, das quais cinco foram corretamente identificadas como falsas. No entanto, uma nota genuína foi erroneamente categorizada como falsificada. A Tabela 4 apresenta detalhes da identificação das notas classificadas como não autênticas, incluindo informações sobre suas respectivas medidas.

Tabela 4 – Notas Identificadas como Falsas pelo Método DDCAM

Índice	Legitimidade	X_1	X_2	X_3	X_4	X_5	X_6
70	Verdadeira	214,9	130,2	130,2	8,0	11,2	139,6
105	Falsa	214,7	130,2	130,3	11,8	10,9	139,7
109	Falsa	215,0	130,2	129,9	10,0	11,9	139,4
122	Falsa	215,1	130,6	130,3	12,3	10,2	139,6
128	Falsa	215,5	130,7	130,3	10,2	11,8	140,0
180	Falsa	214,5	130,2	130,4	8,2	11,8	137,8

Embora os três métodos sob investigação tenham demonstrado precisão na identificação de *outliers*, evidenciada pela detecção correta das notas falsas, vale ressaltar que FSRMCD e IRMCD apresentaram uma diferença de desempenho em comparação ao método DDCAM. Ambos os métodos incorreram em classificações incorretas de sete notas autênticas, o que indica uma possível limitação na capacidade de distinguir entre notas autênticas e falsificadas. Por outro lado, o método DDCAM alcançou uma performance distinta ao classificar corretamente cinco das seis notas que classificou como falsas. Entretanto, é relevante notar que registrou um falso positivo, ao classificar erroneamente uma nota genuína como falsificada. Em análise comparativa, embora todos os métodos tenham demonstrado competência, a robustez do método DDCAM na detecção de *outliers* pode ser considerada superior neste estudo. Essa superioridade

é devida a capacidade de corretamente identificar todas as notas falsas, apesar da ocorrência de um falso positivo.

Para aferir a qualidade do procedimento de identificação de *outliers*, métricas específicas podem ser determinadas. Barbosa (2021) [1] define adequadamente estas medidas. Dado um conjunto de dados composto por n observações multivariadas, das quais algumas são de fato *outliers* multivariados, os seguintes conjuntos podem ser definidos:

- Ω: conjunto composto por todas as observações;
- $\mathbb{O}$: conjunto dos elementos *outliers*;
- $\mathbb{D}$: conjunto das observações identificadas pelo método utilizado como *outliers*.

A sensibilidade no processo de detecção é uma medida de probabilidade . Dado que uma observação pertence à $\mathbb{O}$, a sensibilidade é a probabilidade dela pertencer à $\mathbb{D}$, ou seja, $P(\mathbb{D}|\mathbb{O})$. Já para a especificidade, dado que uma observação não pertence à $\mathbb{O}$, a especificidade é a probabilidade dela não pertencer à $\mathbb{D}$, ou seja, $P(\overline{\mathbb{D}}|\overline{\mathbb{O}})$, em que para um dado conjunto A, tem-se que $\overline{A}$ é o complemento do conjunto A. O valor prediditivo positivo (VPP) é a probabilidade de uma observação ser de fato um *outlier* dado que foi identificada como *outlier* pela metodologia em uso, ou seja, $P(\mathbb{O}|\mathbb{D})$. Por fim, a acurácia é a proporção total de acertos dentre positivos e negativos, ou seja, $P[(\mathbb{D} \cap \mathbb{O}) \cup (\overline{\mathbb{D}} \cap \overline{\mathbb{O}})]$.

A Tabela 5 auxilia para um claro entendimento sobre as medidas de aferição da qualidade que serão utilizadas. A medida de sensibilidade é dada por $a/(a+c)$, a medida de especificidade por $d/(b+d)$, a medida de VPP por $a/(a+b)$ e a medida de acurácia por $(a+d)/n$.

Tabela 5 – Medidas de Aferição da Qualidade [1].

Método	*Outlier*		Total
	Sim	Não	
Positivo	a (verdadeiros positivos)	b (falsos positivos)	$a+b$ (positivos)
Negativo	c (falsos negativos)	d (verdadeiros negativos)	$c+d$ (negativos)
Total	$a+c$ (*outliers*)	$b+d$ (não *outliers*)	$a+b+c+d$ (n)

Com base nessas métricas, as medidas de desempenho foram calculadas após a utilização de cada uma das metodologias para o banco de dados em estudo. Na Tabela 6 é possível observar métricas que confirmam o desempenho de cada um dos três métodos aplicados ao estudo.

Tabela 6 – Desempenho dos Métodos na Detecção de Notas Falsas.

Método	Acurácia (%)	Especificidade (%)	VPP (%)	Sensibilidade (%)
DDCAM	99,05%	99,00%	83,33%	100,00%
FSRMCD	93,33%	93,00%	41,67%	100,00%
IRMCD	93,33%	93,00%	41,67%	100,00%

Os resultados confirmam a adaptabilidade dos três métodos. Todos eles apresentam efetiva qualidade em identificar *outliers*. Entretanto, a métrica de VPP reflete a clara tendência dos métodos FSRMCD e IRMCD em apresentar respostas que superestimam o verdadeiro conjunto solução de *outliers*. É notável a superioridade do método DDCAM em comparação com FSRMCD e IRMCD, especialmente quando se observa o Valor Preditivo Positivo (VPP). O DDCAM apresentou um desempenho aproximadamente duas vezes superior ao dos métodos concorrentes, indicando uma capacidade mais robusta na identificação de verdadeiros positivos. Essa diferença expressiva ressalta a eficácia do DDCAM, destacando seu desempenho significativamente superior na precisão de detecção de notas falsas em relação aos métodos comparativos.

5 Considerações Finais

No contexto dinâmico e interconectado dos sistemas financeiros modernos, a preservação da integridade monetária torna-se uma tarefa complexa e crucial. As ameaças crescentes de fraudes monetárias demandam abordagens inovadoras, capazes de enfrentar a sofisticação dos métodos empregados por agentes mal-intencionados. Nesse cenário, a fusão entre métodos estatísticos avançados e o poder computacional emergente desempenha papel fundamental na vanguarda da detecção de irregularidades financeiras.

A revolução computacional das últimas décadas ampliou não apenas a capacidade de processamento, mas também possibilitou a implementação eficiente de algoritmos complexos. A integração de métodos estatísticos inovadores em plataformas computacionais robustas capacita a exploração e análise de grandes conjuntos de dados de maneiras antes inimagináveis. Essa convergência proporciona uma visão abrangente e dinâmica sobre padrões, comportamentos e anomalias, elementos cruciais na identificação de transações fraudulentas.

No âmbito específico da detecção de notas falsas, a aplicação dos métodos de detecção de outliers multivariados apresentados neste estudo representa avanço significativo para o desenvolvimento e evolução de soluções de problemas neste âmbito. A estatística multivariada, ao considerar simultaneamente múltiplas variáveis interrelacionadas, proporciona compreensão mais profunda e precisa do comportamento dos dados. A evolução computacional, por sua vez, capacita a implementação eficaz desses métodos em grandes conjuntos de dados financeiros, proporcionando resposta mais rápida e precisa às potenciais ameaças.

Diante dos resultados obtidos, este estudo não apenas valida a eficácia dos métodos de detecção de outliers multivariados, mas também destaca a importância de escolhas criteriosas entre esses métodos, subsidiada pela consideração de suas peculiaridades e do desempenho específico em determinados contextos. Particularmente a metodologia DDCAM mostrou algum superioridade no presente estudo. Obviamente não é uma classificação definitiva e peremptória, mas um sinal efetivo da capacidade da metodologia DDCAM.

Para trabalhos futuros, é relevante a exploração de adaptações ou combinações entre métodos distintos, bem como a consideração do emprego de técnicas mais avançadas como técnicas de Inteligência Artificial, Aprendizado de Máquina, entre outras possibilidades. Essas abordagens podem aprimorar ainda mais a capacidade de detecção de fraudes monetárias em larga escala, e incorporar a evolução contínua tanto na

área Estatística quanto na Computacional. Essa perspectiva inovadora pode contribuir significativamente para o desenvolvimento de estratégias mais robustas e eficientes na prevenção e detecção de transações fraudulentas.

Referências

[1] Barbosa, Josino José: *Data-driven Cluster Analysis Method: Uma Nova Metodologia para Detecção de Outliers em Dados Multivariados*. Tese de Doutoramento, 2021. Citado 2 vezes nas páginas 15 e 20.

[2] Barbosa, Josino José, Anderson Ribeiro Duarte e Helgem Souza Ribeiro Martins: *A Performance Evaluation in Multivariate Outliers Identification Methods*. Ciência & Natura, 42:e16 1–14, 2020. Citado 2 vezes nas páginas 1 e 5.

[3] Aggarwal, Charu C.: *An Introduction to Outlier Analysis*, páginas 1–34. Springer International Publishing, 2017. Citado na página 2.

[4] Rousseeuw, Peter J. e Bert C. Van Zomeren: *Unmasking Multivariate Outliers and Leverage Points*. Journal of the American Statistical Association, 85(411):633–639, 1990. Citado na página 5.

[5] Filzmoser, Peter: *Identification of Multivariate Outliers: a Performance Study*. Austrian Journal of Statistics, 34(2):127–138, 2005. Citado na página 5.

[6] Filzmoser, Peter, Robert G. Garrett e Clemens Reimann: *Multivariate Outlier Detection in Exploration Geochemistry*. Computers & Geosciences, 31(5):579–587, 2005. Citado na página 5.

[7] Barbosa, Josino José, Tiago Martins Pereira e Fernando Luiz Pereira Oliveira: *Uma proposta para identificação de outliers multivariados*. Ciência & Natura, 40:e40 1–8, 2018. Citado 2 vezes nas páginas 5 e 7.

[8] Cerioli, Andrea: *Multivariate outlier detection with high-breakdown estimators*. Journal of the American Statistical Association, 105(489):147–156, 2010. Citado 3 vezes nas páginas 5, 11 e 12.

[9] Hardin, Johanna e David M. Rocke: *The distribution of robust distances*. Journal of Computational and Graphical Statistics, 14(4):928–946, 2005. Citado na página 5.

[10] Cerioli, Andrea, Marco Riani e Anthony C. Atkinson: *Controlling the size of multivariate outlier tests with the MCD estimator of scatter*. Statistics and Computing, 19:341–353, 2009. Citado na página 5.

[11] Jobe, John Marcus e Michael Pokojovy: *A cluster-based outlier detection scheme for multivariate data*. Journal of the American Statistical Association, 110(512):1543–1551, 2015. Citado 2 vezes nas páginas 5 e 6.

[12] Li, Jia, Surajit Ray e Bruce G. Lindsay: *A Nonparametric Statistical Approach to Clustering via Mode Identification*. Journal of Machine Learning Research, 8(8), 2007. Citado na página 6.

[13] Patel, Viresh, Aastha Kapoor, Ankush Sharma e Saikat Chakrabarti: *Taxonomy of outlier detection methods for power system measurements*. Energy Conversion and Economics, 2023. Citado na página 6.

[14] Duarte, Anderson Ribeiro, Josino José Barbosa, Helgem Souza Ribeiro Martins e Fernando Luiz Pereira Oliveira: *Data-driven cluster analysis method: a novel outliers detection method in multivariate data*. (*to appear*):1–25, 2024. Citado na página 7.

[15] Du, Qiang, Vance Faber e Max Gunzburger: *Centroidal Voronoi tessellations: Applications and algorithms*. SIAM review, 41(4):637–676, 1999. Citado na página 8.

[16] Rousseeuw, Peter J. e Katrien Van Driessen: *A Fast Algorithm for the Minimum Covariance Determinant Estimator*. Technometrics, 41(3):212–223, 1999. Citado na página 12.

[17] Scrucca, Luca, Michael Fop, T. Brendan Murphy e Adrian E. Raftery: *mclust 5: clustering, classification and density estimation using Gaussian finite mixture models*. The R Journal, 8(1):289–317, 2016. `https://doi.org/10.32614/RJ-2016-021`. Citado na página 15.