UNIVERSIDADE FEDERAL DO CEARÁ - UFC
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE FEAAC
PROGRAMA DE ECONOMIA PROFISSIONAL - PEP

LUCAS CAMINHA QUINTAS COLARES

DETECÇÃO DE ANOMALIAS NO TRANSPORTE DE MERCADORIAS PARA O CEARÁ: UMA ANÁLISE COM DADOS DO MDF-E E TÉCNICAS DE MACHINE LEARNING

FORTALEZA
2025

LUCAS CAMINHA QUINTAS COLARES

DETECÇÃO DE ANOMALIAS NO TRANSPORTE DE MERCADORIAS PARA O CEARÁ: UMA ANÁLISE COM DADOS DO MDF-E E TÉCNICAS DE MACHINE LEARNING

Dissertação apresentada à Coordenação do Curso de Mestrado Profissional em Economia do Setor Público da Universidade Federal do Ceará – UFC/CAEN, como requisito parcial à obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Orientador: Prof. Dr. Fabrício Carneiro Linhares.

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C649d Colares, Lucas Caminha Quintas.
Detecção de anomalias no transporte de mercadorias para o Ceará : uma análise com dados do MDF-e e técnicas de machine learning / Lucas Caminha Quintas Colares. – 2025.
47 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Mestrado Profissional em Economia do Setor Público, Fortaleza, 2025.
Orientação: Prof. Dr. Fabrício Carneiro Linhares .

1. Transporte rodoviário de mercadorias. 2. MDF-e. 3. Detecção de anomalias. 4. Fiscalização tributária. 5. Tempo de deslocamento. I. Título.

CDD 330

LUCAS CAMINHA QUINTAS COLARES

DETECÇÃO DE ANOMALIAS NO TRANSPORTE DE MERCADORIAS PARA O CEARÁ: UMA ANÁLISE COM DADOS DO MDF-E E TÉCNICAS DE MACHINE LEARNING

Dissertação apresentada à Coordenação do Curso de Mestrado Profissional em Economia do Setor Público da Universidade Federal do Ceará – UFC/CAEN, como requisito parcial à obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Aprovada em: 29/08/2025.

BANCA EXAMINADORA

_____
Prof. Dr. Fabrício Carneiro Linhares (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Prof. Dr. Leandro de Almeida Rocco
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Ricardo Brito Soares
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

A Deus, pela minha vida e por me sustentar até aqui, sem Ele nada seria possível.

À minha mãe e ao meu pai, por sempre estarem ao meu lado, incondicionalmente, e a quem devo toda a base de minha educação científica, moral e religiosa.

À minha esposa, pelo companheirismo, apoio, dedicação e amor à nossa família, assumindo sozinha os cuidados de nossa filha recém-nascida nas minhas ausências para me dedicar ao mestrado e ir às aulas.

À minha filha, que é a minha maior motivação para me tornar um ser humano melhor.

Ao meu orientador, pela dedicação e pela orientação criteriosa, sem o qual este trabalho não seria possível.

Aos professores do programa, pelas contribuições valiosas, pelas reflexões instigantes e pelo comprometimento em compartilhar conhecimento.

Ao CAEN/UFC, por oferecer a estrutura e as oportunidades que tornaram este trabalho viável, bem como aos colegas de curso pelo convívio e pelos conhecimentos compartilhados ao longo dessa caminhada.

A todos que, de alguma forma, contribuíram para esta conquista, os meus mais sinceros agradecimentos.

RESUMO

Este trabalho investiga o uso de dados do Manifesto Eletrônico de Documentos Fiscais (MDF-e) no transporte rodoviário de cargas com destino ao estado do Ceará, com foco na detecção de anomalias que possam indicar fraudes fiscais. Foram analisadas mais de 580 mil viagens interestaduais realizadas em 2023, utilizando variáveis como distância geodésica, valor e peso da carga, bem como as Unidades da Federação percorridas. Aplicou-se uma arquitetura de detecção de anomalias composta por dois métodos complementares: Regressão Linear Robusta (RLM) e Isolation Forest (ISO). Os resultados revelaram padrões espaciais e sazonais de anomalias, com destaque para estados do Norte e Centro-Oeste. As anomalias identificadas não apresentaram relação significativa com valor ou peso das cargas, sugerindo influência de fatores logísticos e operacionais. Conclui-se que a análise de dados fiscais eletrônicos, aliada a técnicas de machine learning, pode auxiliar na priorização de rotas e contribuintes suspeitos, promovendo maior eficiência na fiscalização tributária e no combate à sonegação.

Palavras-chave: transporte rodoviário de mercadorias; MDF-e; detecção de anomalias; fiscalização tributária; machine learning; tempo de deslocamento.

ABSTRACT

This work investigates the use of data from the Electronic Manifest of Tax Documents (MDF-e) in the road transportation of cargo destined for the state of Ceará, focusing on detecting anomalies that may indicate tax fraud. More than 580,000 interstate trips made in 2023 were analyzed, using variables such as geodesic distance, value and weight of the cargo, as well as the Federation Units traveled. An anomaly detection architecture composed of three complementary methods was applied: Robust Linear Regression (RLM), Isolation Forest (ISO) and Local Outlier Factor (LOF). The results revealed spatial and seasonal patterns of anomalies, with emphasis on states in the North and Central-West. The identified anomalies did not present a significant relationship with the value or weight of the cargo, suggesting the influence of logistical and operational factors. It is concluded that the analysis of electronic tax data, combined with machine learning techniques, can help prioritize suspicious routes and taxpayers, promoting greater efficiency in tax inspection and combating tax evasion.

Keywords: road transport of goods; MDF-e; anomaly detection; tax inspection; machine learning; travel time.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CONFAZ	Conselho Nacional de Política Fazendária
CT-e	Conhecimento de Transporte Eletrônico
CTN	Código Tributário Nacional
DAMDFe	Documento Auxiliar do Manifesto Eletrônico de Documentos Fiscais
IBS	Imposto sobre Bens e Serviços
ICMS	Imposto de Circulação de Mercadorias e Prestação de Serviços
ISO	Isolation Forest
LOF	Local Outlier Factor
MDF-e	Manifesto Eletrônico de Documentos Fiscais
NF-e	Notas Fiscais Eletrônicas
OLS	Mínimos quadrados ordinários
OSRM	Open Source Routing Machine
RLM	Regressão Linear Robusta
SEFAZ/CE	Secretaria da Fazenda do Estado do Ceará
SINIEF	Sistema Nacional Integrado de Informações Econômico-Fiscais
UF	Unidade Federativa
UFIRCE	Unidade Fiscal de Referência do Estado do Ceará

SUMÁRIO

1. INTRODUÇÃO

O Ceará é considerado um estado consumidor, quando se trata do Imposto de Circulação de Mercadorias e Prestação de Serviços (ICMS), exatamente, por isso, a Secretaria da Fazenda do Estado do Ceará (SEFAZ/CE) optou por cobrar, como regra, o ICMS na entrada do território estadual, conforme Art. 767, Decreto Nº 24.569/97. Desse modo, fiscalizar a entrada das mercadorias é uma prioridade para a SEFAZ/CE, isso tendo sido demonstrado ao longo do tempo pela política de forte atuação em Postos Fiscais nas divisas do estado.

Considerando também o crescimento populacional no Brasil, que leva ao aumento de circulação de mercadorias, a evolução tecnológica e a busca por maior eficiência nos processos, que tem como consequência a aceleração nos procedimentos do comércio de mercadorias, a fiscalização pessoal e presencial nos postos de todas as mercadorias que entram por caminhões, navios e aeronaves, atualmente, se torna inviável.

Aliado ao fato de que sonegadores fiscais, muitas vezes, não param nos postos fiscais, fazendo com que a mercadoria entre em território estadual sem a devida fiscalização, sem a emissão de documentos fiscais ou com a emissão contendo dados falsos ou inexatos, afetando, assim, a arrecadação do ICMS e distorcendo as estatísticas de logística.

Ademais, com a entrada em vigor da nova legislação sobre a reforma tributária, o tributo de competência estadual será o Imposto sobre Bens e Serviços (IBS), com arrecadação, exclusivamente, para o local onde o consumo ocorre, e não partilhada entre o local de produção e o local de consumo como atualmente acontece. Diante desse cenário, surge a necessidade de informatizar ainda mais os processos, utilizando-se das tecnologias disponíveis e da inteligência na análise de dados para se trabalhar de forma eficiente e direcionada. A emissão de documentos fiscais no formato eletrônico como, por exemplo, a Nota Fiscal Eletrônica (NF-e), foi um importante passo para que os Fiscos pudessem receber de forma mais organizada e rápida essa grande quantidade de dados.

Mas não só as notas fiscais se tornaram eletrônicas, o Conhecimento de Transporte Eletrônico (CT-e), documento emitido para acobertar as prestações de serviços de transporte interestadual e intermunicipal, e o Manifesto Eletrônico de Documentos Fiscais (MDF-e), também se tornaram digitais. Este último documento, talvez, pouco conhecido da população de uma forma geral, mas que tem um importante papel para o controle e registro do

trânsito de mercadorias, deve ser utilizado no transporte rodoviário, ferroviário, aquaviário ou aéreo.

Assim, segundo o Manual de Orientações ao Contribuinte do Portal do Manifesto Eletrônico de Documentos Fiscais, a finalidade do MDF-e é agilizar o registro em lote de documentos fiscais em trânsito e identificar a unidade de carga utilizada e demais características do transporte, porquanto, quando da consulta de um manifesto, obtém-se uma visão geral do que deve estar sendo transportado.

Atualmente, no âmbito da monitoramento de mercadorias em trânsito não se tem a informação de quanto tempo uma mercadoria leva para ser transportada de outros estados para o Ceará, e como não existem câmeras de fiscalização em todas as vias do país, um meio em que se poderia obter uma informação formal do tempo médio entre o carregamento e o descarregamento de um caminhão, seria através do MDF-e, posto que, segundo a Cláusula terceira do Ajuste SINIEF 21, o MDF-e deverá ser emitido no término do carregamento e antes do início do transporte, e segundo a Cláusula décima quarta do mesmo Ajuste SINIEF, deverá ser encerrado ao término do último descarregamento descrito no documento.

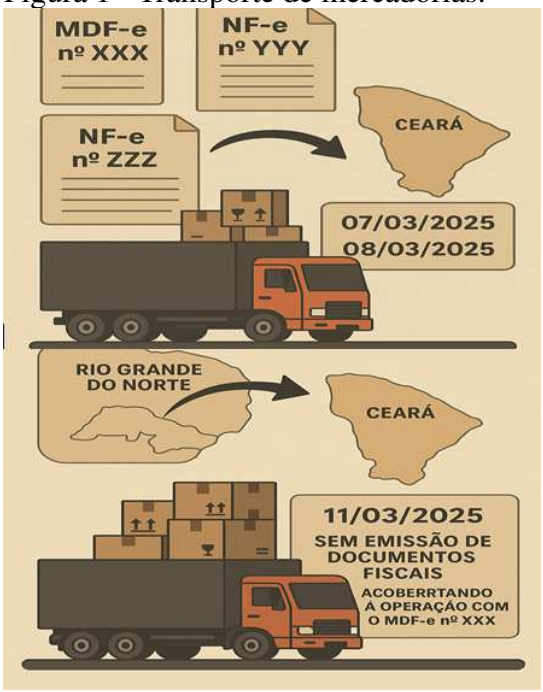
Essa informação é muito importante para identificação de possíveis veículos que estejam trafegando com manifestos já utilizados em outras viagens, o que não é permitido pela legislação. Importante também para interceptação de caminhões que estejam se deslocando para nosso estado e que muitas vezes pegam vias alternativas (fora de rota), desviando dos postos fiscais, pois, atualmente, não se tem uma estimativa de tempo desses deslocamentos, o que poderia subsidiar uma possível abordagem para verificação, já que o monitoramento via câmeras é uma realidade no estado do Ceará, com pontos espalhados pelo estado e nas divisas, identificando os caminhões que entram e trafegam pelo território.

Esse tipo de fraude, faz com que mercadorias sejam transportadas para o território cearense sem a devida emissão de notas fiscais, consequentemente, sem o devido recolhimento do ICMS, gerando graves prejuízos para o estado, perda de receitas, para os demais contribuintes, concorrência desleal, e para toda população do Ceará, falta de verba pública para prestação de serviço públicos.

Sendo mais claro, vamos supor que um MDF-e nº XXX foi utilizado para transportar as mercadorias relacionadas na NF-e nº YYY e na NF-e nº ZZZ provenientes do estado do Rio Grande do Norte; que o MDF nº XXX foi emitido no dia 07/03/2025 e as mercadorias entregues no Ceará no dia 08/03/2025. Vamos supor ainda que este mesmo

caminhão retorne para o Rio Grande do Norte no dia 09/03/2025, sem que o MDF-e nº XXX seja encerrado, e que faça uma nova viagem trazendo mercadorias novamente para o Ceará no dia 11/03/2025, mas dessa vez sem a emissão de documentos fiscais e tentando acobertar a operação com o MDF-e nº XXX. Assim, as mercadorias poderiam entrar no estado do Ceará sem o recolhimento de nenhum imposto.

Figura 1 - Transporte de mercadorias.



Fonte: Elaborada pelo autor.

A manutenção de um MDF-e aberto após o último descarregamento, seguida do uso desse mesmo documento para acobertar nova viagem sem emissão de NF-e, tende a produzir incongruências temporais entre a emissão e o encerramento. Trata-se de um artifício documental cuja materialidade se manifesta como anomalia no padrão temporal. Já o contorno de postos fiscais por rotas alternativas altera a geometria e a fricção do trajeto, gerando tempos de viagem atípicos, mais longos, ou em certos contextos mais curtos, dadas a distância e o corredor logístico.

Importa notar que os comportamentos acima não são observados de forma direta; infere-se sua presença a partir de padrões incompatíveis com o comportamento histórico de pares origem/destino.

Ademais, existem diferenças entre anomalias logísticas, que são as decorrentes de situações atípicas, como por exemplo, congestionamentos, condições climáticas, manutenção do veículo e incidentes de tráfego; e anomalias documentais, caracterizadas por inconsistências formais em campos do MDF-e, como por exemplo, erros nas datas de emissão

e encerramento.

Dessa forma, o objetivo deste trabalho é, utilizando os dados recebidos dos Manifestos Eletrônicos de Documentos Fiscais (MDF-e) pela Secretaria da Fazenda do Estado do Ceará, que possui em sua base de dados 581.138 viagens rodoviárias originadas em 2.546 municípios brasileiros com destino ao Ceará no ano de 2023, detectar anomalias no tempo de deslocamento rodoviário usando variáveis declaradas no MDF-e com métodos robustos de machine learning. Buscando, assim, identificar:

- estados de origem em que os registros anômalos estão concentrados;
- meses do ano em que os registros anômalos estão concentrados;
- perfis de carga (valor/peso) em que os registros anômalos estão concentrados.

Propondo uma arquitetura de detecção que combina Regressão Linear Robusta (RLM) e *Isolation Forest (ISO)*, construindo indicadores estaduais de suspeição para fins de fiscalização do trânsito de mercadorias e oferecendo evidências empíricas sobre a relação entre anomalias, valor/peso da carga e sazonalidade.

Os procedimentos analíticos empregados atuam como mecanismos de triagem: classificam e sinalizam suspeição, mas não estabelecem prova de infração nem determinam nexo causal conclusivo. A confirmação requer diligências fiscais e cruzamentos documentais adicionais, incluindo verificação in loco quando cabível, bem como a avaliação de eventuais fontes de vieses de mensuração e de erros de registro.

No Brasil, notadamente, o transporte rodoviário é muito relevante, segundo LEAL (2018), os modos de transporte de mercadorias têm evoluído desde meados do século XX: o transporte rodoviário tornou-se o modal dominante. Sendo assim, este trabalho será focado no modal rodoviário e nos manifestos que tenham como destino o estado do Ceará, alinhando-se à lógica da nova tributação baseada no estado de consumo.

2. LEGISLAÇÃO SOBRE INFORMAÇÃO OBRIGATÓRIA NO TRANSPORTE DE CARGAS

A tributação do transporte interestadual de cargas no Brasil está inserida no contexto mais amplo do Imposto sobre Circulação de Mercadorias e Serviços (ICMS), conforme disciplinado pela Constituição Federal de 1988, legislação complementar nacional e estadual, e normas infralegais específicas, como convênios e ajustes firmados no âmbito do Conselho Nacional de Política Fazendária (CONFAZ).

2.1 Marco normativo do ICMS no transporte interestadual

2.1.1. Fundamento constitucional e Lei Complementar 87/1996 (Lei Kandir)

O Artigo 155, inciso II da Constituição Federal estabelece que compete aos Estados e ao Distrito Federal instituir o ICMS, inclusive sobre prestações de serviços de transporte interestadual e intermunicipal. No §2º do mesmo artigo, estão previstas regras para a partilha do imposto entre os entes federativos de origem e destino, especialmente nos casos de operações interestaduais destinadas a consumidor final.

O inciso VII do Art. 155, incluído pela Emenda Constitucional nº 87/2015, alterou significativamente a sistemática de partilha, estabelecendo que o imposto relativo às operações e prestações que destinem bens e serviços a consumidor final não contribuinte do imposto deve ser repartido entre os estados de origem e de destino.

A Lei Complementar nº 87, de 13 de setembro de 1996, também conhecida como Lei Kandir, regulamenta o ICMS em nível nacional e define, no seu Art. 2º, inciso II, a incidência do imposto sobre prestações de serviço de transporte interestadual e intermunicipal, por qualquer via, de pessoas, bens, mercadorias ou valores.

A lei também estabelece que o local da operação ou da prestação, para os efeitos da cobrança do imposto e definição do estabelecimento responsável, é onde se encontre o transportador, quando em situação irregular pela falta de documentação fiscal ou quando acompanhada de documentação inidônea, segundo a alínea b, inciso II do Art. 11 da Lei Complementar nº 87/96, que se enquadraria nas situações de interceptações de mercadorias sem documentos fiscais.

2.1.2. Convênios, ajustes SINIEF e legislação do estado do Ceará

A operacionalização da cobrança do ICMS no transporte interestadual exige uma

série de normas infralegais complementares, celebradas entre os estados no âmbito do CONFAZ.

Destacam-se, no contexto desta pesquisa, o Convênio ICMS 142/2018, que dispõe sobre os regimes de substituição tributária e de antecipação de recolhimento do ICMS com encerramento de tributação, relativos ao imposto devido pelas operações subsequentes; o Ajuste SINIEF 09/2007, que institui o Conhecimento de Transporte Eletrônico – CT-e, e o Ajuste SINIEF 21/2010, que institui o Manifesto Eletrônico de Documentos Fiscais – MDF-e.

No estado do Ceará, a obrigatoriedade de emissão do MDF-e está ratificada na Subseção III do Decreto nº 35.061/2022, em que se destaca o § 1.º do Art. 107:

> § 1.º O MDF-e deverá ser emitido nas situações descritas no caput deste artigo, nas operações e prestações intermunicipais dentro deste Estado e interestaduais, e sempre que haja transbordo, redespacho, subcontratação ou substituição do veículo, do motorista, de contêiner ou inclusão de novas mercadorias ou documentos fiscais, bem como na hipótese de retenção imprevista de parte da carga transportada.

Essas normas determinam a obrigatoriedade de emissão, o conteúdo mínimo e a forma de armazenamento dos documentos fiscais eletrônicos no transporte de cargas, em especial, o MDF-e consolida informações dos CT-es e das NF-es, definindo os pontos de origem, destino, percurso estimado, horários de início e fim da viagem e a identificação dos veículos.

2.1.3. Fiscalização, cruzamento de informações e relevância para a pesquisa

As ações fiscais relacionadas ao trânsito de mercadorias podem ser realizadas de forma presencial ou remota, com o emprego de câmeras, balanças, scanners, entre outros, conforme as alíneas a e b do Art. 43 do Decreto nº 34.605/2022.

Além disso, a obrigatoriedade do MDF-e permite às Secretarias da Fazenda estaduais — como a SEFAZ/CE — monitorar o deslocamento de mercadorias em trânsito. Isso viabilizaria o uso de ferramentas automatizadas de detecção de inconsistências ou fraudes com base em tempo, distância, valor e rota, tema central desta dissertação.

O cruzamento entre o valor declarado, o tempo de trajeto e as Unidades Federativas (UFs) percorridas permitiria detectar desvios logísticos e práticas fraudulentas com potencial impacto na arrecadação de ICMS, fundamentais nas ações fiscais do trânsito de mercadorias de acordo com alínea c, III do Art. 43 do Decreto nº 34.605/2022:

> **Art. 43.** Relativamente às ações fiscais que se refiram às operações e prestações

> relacionadas ao trânsito de mercadorias, bens, valores ou pessoas observar-se-á o seguinte:
>
> III – considerar-se-á iniciada, inclusive para fins de cessação da espontaneidade do contribuinte, quando:
>
> c) os sistemas eletrônicos da SEFAZ detectarem remotamente, de forma eletrônica e automática, a emissão de documentos fiscais relativos a operações e prestações com mercadorias, bens, valores ou pessoas em trânsito, de modo a averiguar, inclusive via cruzamento de dados, inconsistências fiscais relacionadas ao cumprimento de obrigações tributárias;

O entendimento do marco normativo é essencial não apenas para fundamentar a legitimidade do uso de documentos fiscais como fonte de dados para modelagem estatística, mas também para identificar pontos críticos onde ocorrem divergências temporais suspeitas que podem estar associadas a práticas ilícitas, como, por exemplo, reutilização de documentos para acobertar viagens que já foram realizadas anteriormente.

Esses aspectos reforçam a importância da integração entre a legislação tributária e os métodos de detecção automatizada discutidos nesta dissertação, com vistas ao aprimoramento da fiscalização, à redução de fraudes e, consequentemente, ao aumento de arrecadação com a mitigação das perdas arrecadatórias.

2.2 Documentação fiscal eletrônica

Instituído pelo Ajuste SINIEF 21/10, o Manifesto Eletrônico de Documentos Fiscais (MDF-e) foi criado para consolidar em um único documento as informações relacionadas ao transporte de cargas, especialmente aquelas já constantes em NF-es e CT-es. Sua emissão é obrigatória para contribuintes que realizam transporte de mercadorias com veículos próprios, arrendados ou com contratação de transportador autônomo de cargas.

Conforme Leal (2018):

> As informações coletadas no Manifesto Eletrônico de Documentos Fiscais do DF - MDF-e existem, o que não significa que a coleta seja fácil. Pode-se argumentar que o uso dos dados de forma sistêmica provoca a sua evolução e oportuniza entender o comportamento da mobilidade da carga. Ainda, torna-se um vetor de controle das ações e reações no transporte de mercadoria, facilitando o comando de entrada e saída de divisas no território do Distrito Federal.

O MDF-e não só atende a uma finalidade fiscal, mas também possui relevância logística, possibilitando o monitoramento do percurso da carga, o controle sobre o tempo de viagem e a verificação dos pontos de carregamento e descarregamento. Ele representa uma

obrigação acessória cuja não observância pode acarretar penalidades previstas na legislação tributária.

A funcionalidade do MDF-e é reforçada pela emissão do Documento Auxiliar do Manifesto Eletrônico de Documentos Fiscais (DAMDF-e), que acompanha fisicamente a carga e possibilita a fiscalização nas estradas. Em que pese a utilização de documentos digitais serem a regra da atualidade, existem situações que a falta de internet, principalmente, em rodovias, onde, infelizmente, ainda não se tem uma cobertura de sinal completa, faz-se necessário o acompanhamento do DAMDF-e junto à mercadoria.

O Ajuste SINIEF 21/10, aprovado em 10 de dezembro de 2010, foi o marco regulatório para a criação e regulamentação do Manifesto Eletrônico de Documentos Fiscais (MDF-e). Esse ajuste estabelece as diretrizes para a emissão e utilização do MDF-e pelos contribuintes que realizam o transporte de mercadorias, seja com veículos próprios, arrendados ou com transportadores autônomos de cargas.

A obrigação de utilizar o MDF-e está vinculada à concessão da "Autorização de Uso" do documento, que ocorre mediante transmissão do arquivo digital à Secretaria da Fazenda, via protocolo de segurança e criptografia, conforme Cláusula 6ª do Ajuste SINIEF 21/10. Essa autorização é necessária para validar o manifesto e possibilitar o acompanhamento da carga pelo fisco.

A autorização de uso do MDF-e é um procedimento essencial, mas não garante a veracidade do transporte e que a carga esteja em conformidade com a legislação tributária e com o que foi declarado, que é responsabilidade legal do emitente. Essa autorização tem um papel fundamental na fiscalização e controle dos documentos fiscais eletrônicos, após a autorização de uso do MDF-e, o arquivo correspondente é disponibilizado para as unidades da federação envolvidas no carregamento e descarregamento das mercadorias, bem como para as unidades indicadas como percurso, caso estas sejam diferentes da unidade autorizadora, conforme a Cláusula 9ª do Ajuste SINIEF 21/2010.

O cancelamento do MDF-e, conforme previsto na Cláusula 13ª do Ajuste SINIEF 21/2010, caso seja necessário, deve ocorrer dentro de um prazo de 24 horas após a concessão de sua autorização de uso, desde que o transporte não tenha sido iniciado. Esse cancelamento é fundamental para evitar a circulação de documentos fiscais falsificados ou inconsistentes, garantindo a integridade do sistema de transporte e fiscalização.

Já o encerramento do MDF-e ocorre com o registro de evento de conclusão do transporte, como, por exemplo, o término do descarregamento, na hipótese de retenção imprevista e parcial da carga transportada, no caso de inclusão de novas mercadorias para a mesma UF de descarregamento, ou na substituição do veículo, de acordo com a Cláusula 14ª do Ajuste SINIEF 21/2010. O encerramento deve ser realizado de forma tempestiva para garantir que a carga tenha sido descarregada ou transportada corretamente e que não haja irregularidades no processo de transporte.

2.3 Penalidades e obrigações acessórias

A obrigação acessória é um conceito fundamental no direito tributário, que segundo o § 2º do Art. 113 do Código Tributário Nacional (CTN), decorre da legislação tributária e tem por objeto as prestações, positivas ou negativas, nela previstas no interesse da arrecadação ou da fiscalização dos tributos. Se refere à obrigação do contribuinte de fornecer informações ao fisco, sem que isso implique diretamente no pagamento de tributos. No caso do MDF-e, a obrigação acessória se traduz na necessidade de emissão e transmissão do documento para que a carga possa ser transportada dentro das exigências legais, permitindo o controle fiscal sobre o trânsito das mercadorias.

A ausência ou erro no preenchimento do MDF-e pode acarretar penalidades, como multas e restrições administrativas, conforme estabelecido no § 3º do Art. 113 do CTN, em que a obrigação acessória, pelo simples fato da sua inobservância, converte-se em obrigação principal relativamente à penalidade pecuniária. No Ceará, a Lei nº 18.665/2023 trouxe no inciso III do Art. 177 as seguintes penalidades:

> **Art. 177**. As infrações à legislação do ICMS sujeitam o infrator às seguintes penalidades, sem prejuízo do pagamento do imposto, quando for o caso:
>
> III - relativamente à documentação e à escrituração:
>
> m) deixar o contribuinte de emitir o Manifesto Eletrônico de Documentos Fiscais (MDF-e), quando obrigado nos termos da legislação pertinente: multa equivalente a 400 (quatrocentas) UFIRCEs por cada MDF-e não emitido;
>
> n) transportar mercadoria ou bem desacompanhado do Documento Auxiliar do Manifesto Eletrônico de Documentos Fiscais (DAMDFE), no formato impresso ou eletrônico: multa equivalente a 400 (quatrocentas) UFIRCEs por documento;
>
> o) transportar mercadoria ou bem cujo documento fiscal não esteja relacionado no Documento Auxiliar do Manifesto Eletrônico de Documentos Fiscais (DAMDFE) que

> acompanha a carga: multa equivalente a 100 (cem) UFIRCEs por cada documento omitido;

A exigência de que o MDF-e seja transmitido antes do início do transporte visa justamente permitir ao fisco a fiscalização prévia, de modo a identificar possíveis irregularidades, como a falta de documentos fiscais ou o transporte de mercadorias sem a devida autorização.

A emissão dos MDF-es também facilita o acompanhamento e monitoramento do transporte das mercadorias, promovendo um ambiente mais transparente e eficaz para a fiscalização tributária, além de fornecer dados valiosos para a administração pública e para os próprios operadores.

Caso o MDF-e esteja sendo "reutilizado" para uma outra viagem, a qual não foram emitidas notas fiscais, por exemplo, estaria acobertando uma situação em que as penalidades passam a ser relativas à obrigação principal.

Segundo o § 1º do Art. 113 do CTN, a obrigação principal surge com a ocorrência do fato gerador, tem por objeto o pagamento de tributo ou penalidade pecuniária e extingue-se juntamente com o crédito dela decorrente, assim os responsáveis podem ser penalizados com base nas alíneas a, b e f do inciso III do Art. 177 da Lei nº 18.665/2023, por exemplo:

> **Art. 177**. As infrações à legislação do ICMS sujeitam o infrator às seguintes penalidades, sem prejuízo do pagamento do imposto, quando for o caso:
>
> III - relativamente à documentação e à escrituração:
>
> a) entregar, remeter, transportar, receber, estocar ou depositar mercadorias, bem como prestar ou utilizar serviços:
>
> 1. sem documentação fiscal: multa equivalente a 30% (trinta por cento) do valor da operação ou da prestação;
>
> b) deixar de emitir documento fiscal:
>
> 1. em operações e prestações tributadas: multa equivalente a 30% (trinta por cento) do valor da operação ou da prestação;
>
> f) promover saída de mercadoria ou prestação de serviço acompanhada de documento fiscal já utilizado em operação ou prestação anterior, inclusive quando se tratar de documento fiscal eletrônico ou sua respectiva representação gráfica impressa: multa equivalente a 30% (trinta por cento) do valor da operação ou da prestação;

2.4 Estudos Correlacionados

Atualmente, técnicas de aprendizado de máquina para detecção de anomalias que possam apontar fraudes são amplamente utilizadas, Ludivia *et al* (2024) em estudo de 104 artigos, identificou que os modelos de regressão são amplamente utilizados em combate a fraudes financeiras, assim como as técnicas de *machine learning* supervisionadas e não supervisionadas. Os autores projetam perspectivas promissoras para pesquisas futuras com possibilidade de aprimoramento e aplicação dessas técnicas em outras áreas, o que possibilitaria avançar na detecção de anomalias em tempo real e minimizar os riscos das organizações.

Muhammad (2022) utilizou a capacidade dos algoritmos de *machine learning* na detecção de empresas fraudulentas na China, combinando classificadores independentes diversos e precisos, o que aprimorou o poder preditivo do modelo desenvolvido de detecção de fraudes, tendo como "metaclassificador" de melhor desempenho a combinação de Regressão Logística, *Random Forests* e *RusBoost*.

Souza *et al* (2021) desenvolveu uma abordagem de detecção de anomalias, baseada na técnica *Isolation Forest*, para monitoramento de sistemas. Apesar dos dados de interesse não apresentarem correlação entre si durante os experimentos, foi notada boa capacidade de generalização para a detecção de anomalias nos processos.

Tosta (2025) analisou o impacto da Inteligência Artificial no setor bancário e financeiro, especificamente na detecção de fraudes em transações financeiras, utilizando técnicas modernas e algoritmos avançados, como *Isolation Forest*, na identificação de padrões suspeitos, e concluindo pela imprescindibilidade da inteligência artificial para detecção e prevenção de fraudes nas instituições bancárias.

Ounacer *et al* (2018) comparou diferentes métodos não supervisionados para detecção de fraudes em cartões de crédito, com exemplo, *LOF, one class SVM, K-means and Isolation Forest*, para destacar a melhor abordagem de detecção de fraudes de cartão de crédito e que fosse capaz de detectar o maior número de novas transações em tempo real com alta precisão; concluindo que a floresta de isolamento é muito eficiente na detecção de anomalias no caso de cartões de crédito.

Vanini *et al* (2023) definiu três modelos, um para detecção de fraudes baseada em aprendizado de máquina, outro para otimização econômica dos resultados do aprendizado de máquina e outro para prever o risco de fraude, considerando contramedidas. Os modelos foram testados utilizando dados reais, e técnicas como *LOF* e *IF* para detecção de *outliers*.

Demonstrando que a detecção de anomalias não é apenas útil por identificar uma proporção significativa de fraudes e controlar alarmes falsos, mas porque a associação da detecção de anomalias a métodos estatísticos de gestão de risco pode reduzir significativamente o risco.

Na área tributária, Alsadhan (2023) empregou modelos supervisionados e não supervisionados para analisar declarações de imposto de renda, dividindo da seguinte forma: um módulo supervisionado, que utiliza um modelo em árvore para extrair conhecimento dos dados; um módulo não supervisionado, que calcula pontuações de anomalia; um módulo comportamental, que atribui uma pontuação de conformidade para cada contribuinte; e um módulo de previsão, que utiliza a saída dos módulos anteriores para gerar uma probabilidade de fraude para cada declaração de imposto de renda. Ele testou a estrutura com declarações de imposto de renda existentes fornecidas pela autoridade fiscal saudita, demonstrando sua eficácia.

Bittencourt Neto (2018) utilizou modelos estatísticos e métodos de mineração de dados para a análise de outliers sobre as informações da Notas Fiscais Eletrônicas e do Livro Fiscal Eletrônico, investigando novas modalidades de evasão fiscal no ICMS, e propondo um recurso computacional construído em linguagem R (plataforma R Studio) para identificação das circunstâncias anômalas, gerando, dessa forma, maior eficiência à atividade de programação fiscal de auditorias tributárias.

Vanhoeyveld (2019) aplicou técnicas de detecção de anomalias não supervisionadas para analisar um conjunto de dados exclusivo contendo as declarações do Imposto sobre Valor Agregado e as listagens de clientes belgas pertencentes a dez setores. Segundo ele, as altas taxas de acerto observadas na maioria dos setores demonstram o sucesso dessa abordagem e pode ser adotada por autoridades fiscais em todo o mundo. Existem diferenças setoriais devido às diferentes condições de mercado e requisitos legais entre os setores, e demonstrou-se que o método ideal depende do setor.

Savić (2022) examinou a possibilidade de um método híbrido não supervisionado para gerenciamento de risco de evasão fiscal que fosse capaz de validar e explicar internamente outliers detectados em um determinado conjunto de dados fiscais, declarações individuais de imposto de renda de pessoa física coletadas pela Administração Tributária da Sérvia. O método proposto, Hunod (*Hybrid Unsupervised Outlier Detection*), combina aprendizado de agrupamento e representação para detecção robusta de outliers. Os resultados obtidos mostram que o método indica entre 90% e 98% de outliers validados internamente.

Baghdasaryan (2022) desenvolveu um modelo de previsão de fraude baseado em ferramentas de aprendizado de máquina utilizando o universo de contribuintes empresariais armênios que operam sob um regime tributário padrão. O problema de detecção de fraude fiscal foi abordado usando técnicas de aprendizado supervisionado e não supervisionado, como regressão logística e algoritmos baseados em árvores — árvore de decisão, floresta aleatória e aumento de gradiente. Concluiu que o uso de dados sobre fraude na rede imediata de fornecedores e compradores do contribuinte em questão é quase tão informativo quanto seus registros históricos de auditoria e seus resultados.

González (2013) utilizou algoritmos de agrupamento como *Self-Organizing Map* (SOM) e gás neural para identificar grupos de comportamento semelhante no universo de contribuintes. Utilizando, em seguida, árvores de decisão, redes neurais e redes bayesianas para identificar as variáveis relacionadas à conduta de fraude e/ou não fraude, para detectar padrões de comportamento associados e para estabelecer até que ponto os casos de fraude e/ou não fraude podem ser detectados com as informações disponíveis. Concluindo que é possível caracterizar e detectar potenciais usuários de notas fiscais falsas em um determinado ano, a partir das informações de pagamento de impostos, do desempenho histórico e de suas características.

Percebe-se que a detecção de anomalia através de modelos estatísticos e *machine learning* está bem difundida no meio acadêmico. Logo, esse trabalho visa contribuir com a análise de detecção dessas anomalias, especificamente, no transporte de cargas para o Ceará através da utilização das informações contidas nos MDF-es.

3. BASE DE DADOS E ESTATÍSTICAS DESCRITIVAS

3.1 Fonte e período

Neste trabalho foram utilizadas 581.138 observações do banco de dados MDF-e da SEFAZ-CE referente ao ano-base de 2023 das operações interestaduais de entrada de mercadorias no Ceará.

Figura 2 - Base de dados MDF-e.

id	m	da	de	ei	ef	ep	cc	mc	cd	md	ee	t	qc	qn	v	p
2,42301E+43	1	01/01/2023 08:29	02/01/2023 05:36	RN	CE	NA	2401453	Barauna	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	3066338	34200
2,42301E+43	1	01/01/2023 08:46	03/01/2023 06:14	RN	CE	NA	2408003	MOSSORO	2307650	MARACANAU	RN	NA	NA	1	29232	10752
5,22301E+43	1	01/01/2023 08:55	07/01/2023 09:54	MT	CE	BA, TO, PI, GO	5101803	Barra do Garcas	2307601	Limoeiro do Norte	GO	NA	1	NA	7261891	279488
2,92301E+43	1	01/01/2023 08:57	02/01/2023 08:51	PE	CE	NA	2606804	IGARASSU	2312403	SAO GONCALO DO AMARANTE	BA	NA	1	NA	10	1198512
2,42301E+43	1	01/01/2023 09:00	02/01/2023 10:56	RN	CE	NA	2401453	Barauna	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	3249242	36240
2,92301E+43	1	01/01/2023 09:05	03/01/2023 13:05	PE	CE	PB	2606804	IGARASSU	2312403	SAO GONCALO DO AMARANTE	BA	NA	1	NA	10	1198512
2,22301E+43	1	01/01/2023 09:10	05/01/2023 08:08	PI	CE	NA	2207009	OEIRAS	2305506	IGUATU	PI	NA	1	NA	562302	48060
2,32301E+43	1	01/01/2023 09:15	02/01/2023 08:34	PE	CE	NA	2607752	Itapissuma	2304202	CRATO	CE	NA	NA	2	16582624	35238
2,42301E+43	1	01/01/2023 09:30	03/01/2023 12:13	RN	CE	NA	2401453	Barauna	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	3128203	34890
2,42301E+43	1	01/01/2023 09:35	02/01/2023 00:22	RN	CE	NA	2401453	Barauna	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	3158687	35230
2,42301E+43	1	01/01/2023 09:45	02/01/2023 10:15	RN	CE	NA	2401453	Barauna	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	3244759	36190
2,42301E+43	1	01/01/2023 09:45	02/01/2023 16:15	RN	CE	NA	2404408	Grossos	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	31860	56400000
2,62301E+43	1	01/01/2023 10:18	02/01/2023 13:08	PE	CE	NA	2607208	IPOJUCA	2303709	CAUCAIA	PE	NA	1	NA	12007804	28100
2,42301E+43	1	01/01/2023 10:23	03/01/2023 17:05	RN	CE	NA	2401453	Barauna	2312403	Sao Goncalo do Amarante	RN	NA	1	NA	3170342	35360
3,12301E+43	1	01/01/2023 10:32	05/01/2023 13:46	MG	CE	PB, BA, PE	3104205	ARCOS	2312403	SAO GONCALO DO AMARANTE	MG	NA	1	NA	530322	32880
2,62301E+43	1	01/01/2023 11:00	04/01/2023 06:17	PE	CE	NA	2611101	PETROLINA	2307650	MARACANAU	PE	NA	NA	1	7700	8300
2,62301E+43	1	01/01/2023 11:09	02/01/2023 13:08	PE	CE	NA	2607208	IPOJUCA	2303709	CAUCAIA	PE	NA	1	NA	11336906	26530
2,12301E+43	1	01/01/2023 11:58	02/01/2023 14:24	PI	CE	NA	2211001	TERESINA	2304400	FORTALEZA	MA	NA	1	NA	79999	28203
2,62301E+43	1	01/01/2023 12:05	08/01/2023 09:45	PE	CE	NA	2609600	OLINDA	2304400	FORTALEZA	PE	NA	1	NA	4669727	4800
2,62301E+43	1	01/01/2023 12:25	02/01/2023 13:08	PE	CE	NA	2607208	IPOJUCA	2303709	CAUCAIA	PE	NA	1	NA	11076238	25920
2,72301E+43	1	01/01/2023 12:26	03/01/2023 18:29	AL	CE	PB, RN, PE	2704302	MACEIO	2307650	MARACANAU	AL	NA	2	NA	6358772	2985472
2,32301E+43	1	01/01/2023 12:42	03/01/2023 10:38	PE	CE	NA	2610806	Pedra	2308708	Morada Nova	CE	NA	NA	1	66300	34000
2,62301E+43	1	01/01/2023 12:48	04/01/2023 17:34	PE	CE	PB, RN	2605202	ESCADA	2303709	CAUCAIA	PE	NA	1	NA	13840909	30000
3,52301E+43	1	01/01/2023 12:56	02/01/2023 18:02	SE	CE	PB, RN, AL, PE	2802106	ESTANCIA	2307650	MARACANAU	SP	NA	1	NA	60128	800
3,12301E+43	1	01/01/2023 13:33	10/01/2023 15:52	MG	CE	BA, PE, PB	3104205	ARCOS	2312403	SAO GONCALO DO AMARANTE	MG	NA	1	NA	772257	47880
3,52301E+43	1	01/01/2023 13:40	05/01/2023 08:40	MG	CE	PI, PE, PB, BA	3104205	Arcos	2312403	Sao Goncalo do Amarante	SP	NA	1	NA	620967	38500

Fonte: Elaborada pelo autor.

3.2 Variáveis

Após o processo de limpeza e filtragem, foram utilizadas as seguintes variáveis conforme a tabela abaixo.

Tabela 1 – Variáveis.

Símbolo	Descrição	Tipo
ei	Sigla da UF do Carregamento	Categórico
ef	Sigla da UF do Descarregamento	Categórico
ep	Sigla das Unidades da Federação do percurso do veículo	Categórico
v	Valor total da carga / mercadorias transportadas (R$)	Contínuo
p	Peso Bruto Total da Carga (Mercadorias transportadas) em Kg	Contínuo
tempo	Mínimo entre data e hora da abertura do MDFe (da) e data e hora que o MDFe foi encerrado (de)	Contínuo
dist_km_geo	Distância geodésica origem–destino	Contínuo
mc_ok, md_ok	Município origem/destino limpo	Categórico
vel_kmh, log_v, log_p	Derivadas	Contínuo

Fonte: Elaborada pelo autor.

3.3 Limpeza e filtragem

Em relação ao tratamento dos dados, aplicou-se a seguinte filtro: modal (m) é igual a “1” (rodoviário), (ei) é diferente de “CE” e (ef) é igual a “CE” para serem retornados apenas os MDF-es de operações interestaduais de entrada no Ceará, exclusivamente, por meio do transporte rodoviário.

Após isso, foram excluídos os registros em que havia mais de um município de carregamento ou descarregamento, por não ser possível identificar qual a ordem dos eventos, permanecendo apenas os registros em que continham a combinação de 1 município de carregamento e 1 munícipio de descarregamento. Dessa forma, 77,50% das observações foram mantidas e 22,50% foram excluídas por terem mais de um munícipio no carregamento ou no descarregamento.

Em seguida, foi realizada a padronização da unidade de medida da variável que representava o peso da carga (QCARGA). A coluna CUNID indicava a unidade de medida utilizada, sendo “1” correspondente a quilogramas (kg) e “2” a toneladas (t), com o intuito de uniformizar os dados, todos os valores expressos em toneladas foram convertidos para quilogramas, por meio da multiplicação por 1000. Após a conversão, foi criada uma variável nova contendo os pesos padronizados, e, posteriormente, as colunas originais QCARGA e CUNID foram removidas da base. A nova coluna foi renomeada como “p”, de forma a preservar a estrutura original da base, agora com valores exclusivamente em quilogramas.

Além disso, foram excluídos 52 registros com tempo ≤ 0 ou coordenadas faltantes, e considerado “rota direta”, quando a variável “ep”, estados de percurso, contivesse a expressão “NA”.

3.4 Estatísticas-chave

Conforme Tabela 2 abaixo, as estatísticas descritivas indicam uma expressiva variação nas distâncias e nos tempos de deslocamento. A variável referente à distância geodésica, menor distância entre dois pontos em uma superfície curva, como a Terra, entre o ponto de origem e o destino (dist_km_geo) apresenta média de aproximadamente 1.025,74 quilômetros, com desvio-padrão de 799,71 quilômetros.

O tempo médio de deslocamento (tempo) foi estimado em cerca de 7.279,75 minutos, com desvio-padrão de 7.596,05 minutos, denotando elevada dispersão. A velocidade média calculada (vel_kmh) é de 16,21 quilômetros por hora, valor significativamente inferior ao esperado para o modal rodoviário, o que pode indicar a ocorrência de atrasos, interrupções

logísticas ou inconsistências nos registros informacionais.

Quanto à localização geográfica, os pontos de origem (variáveis lat_cc e lon_cc) apresentam ampla distribuição ao longo do território nacional, enquanto os pontos de destino (lat_cd e lon_cd) concentram-se no território cearense, o que está em conformidade com o escopo regional do estudo. No que se refere às características das cargas transportadas, o peso médio declarado (p) é de aproximadamente 29.595 de quilogramas, enquanto o valor monetário médio (v) é de R$ 80.707,37. Ambas as variáveis apresentam ampla dispersão e valores máximos extremamente elevados, o que justifica o uso das transformações logarítmicas (log_p e log_v) com o intuito de reduzir a assimetria e viabilizar análises mais robustas.

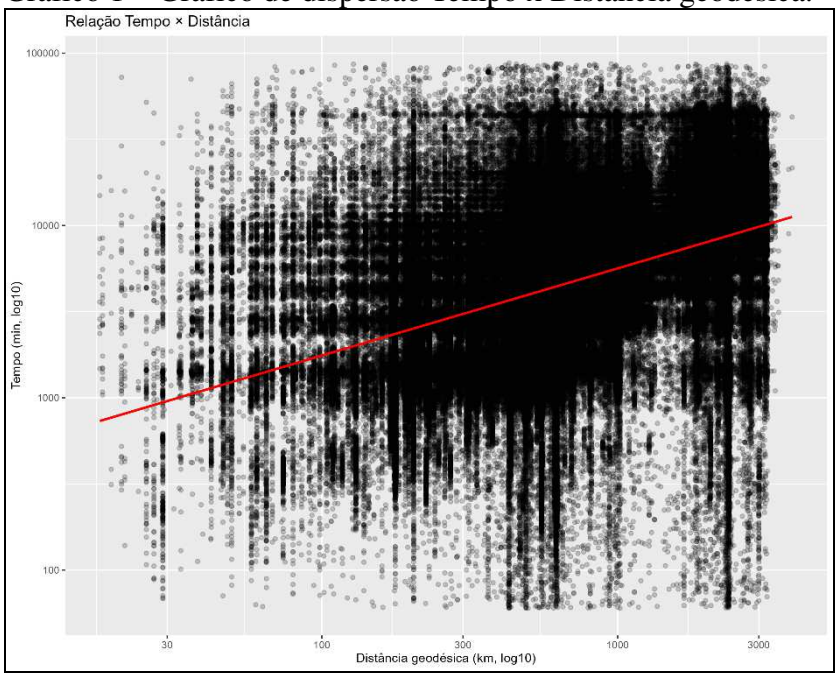
Tabela 2 - Estatística descritiva.

variável	mean	sd	min	q25	median	q75	max
dist_km_geo	1025.743	799.709	17.75666	470.1162	652.1247	1726.711	3888.237
lat_cc	-11.3019	7.126528	-32.2172	-16.0891	-8.23694	-6.11742	3.117848
lat_cd	-4.36295	1.143312	-7.78034	-4.10217	-3.81108	-3.78574	-2.8777
log_p	9.215092	2.571826	0	8.721113	9.818256	10.37727	30.02892
log_v	10.938	2.078292	0	10.06901	11.29021	12.23077	28.20763
lon_cc	-41.3351	5.491446	-73.4391	-46.4548	-40.2745	-36.1805	-34.8417
lon_cd	-38.8098	0.609791	-41.2327	-39.0605	-38.528	-38.528	-37.4105
p	29595844	1.5E+10	0	6130	18365	32120	1.1E+13
tempo	7279.753	7596.05	60.1	2747.7	5419.633	8838.45	87732.73
v	8070737	3.54E+09	0	23599.14	80033.51	205000	1.78E+12
vel_kmh	16.21059	54.36095	0.017426	5.448809	9.800304	15.98674	2865.962

Fonte: Elaborada pelo autor.

No Gráfico 1, percebe-se que, em geral, quanto maior a distância geodésica, maior é o tempo de deslocamento das cargas. No entanto, a dispersão dos pontos indica grande variabilidade nos tempos, mesmo para distâncias semelhantes, sugerindo possíveis anomalias ou fatores externos que afetam o trajeto. A linha vermelha evidencia essa tendência crescente entre tempo e distância.

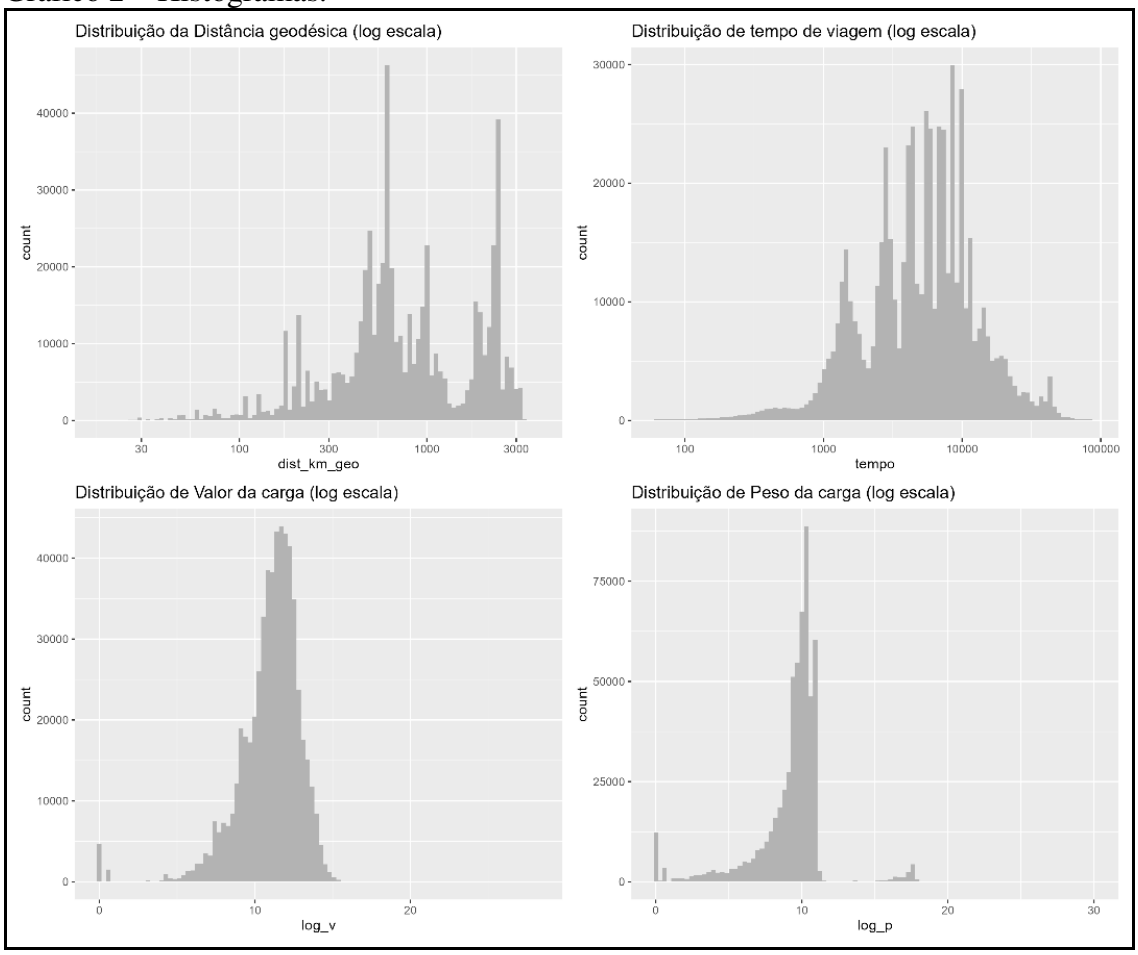
Gráfico 1 – Gráfico de dispersão Tempo x Distância geodésica.



Fonte: Elaborado pelo autor.

No Gráfico 2, tem-se as distribuições de frequências das variáveis distância geodésica, tempo de viagem, valor da carga e peso da carga, todas em escala logarítmica, o que facilita a visualização de dados com grande variação. Ressalta-se que na distribuição de peso da carga há um corte bem definido devido a capacidade máxima de carga dos caminhões. Percebe-se também alguns registros com pesos superiores a esse corte, que podem estar relacionados a cargas específicas transportadas por veículos especiais ou até mesmo por erros de registros.

Gráfico 2 – Histogramas.

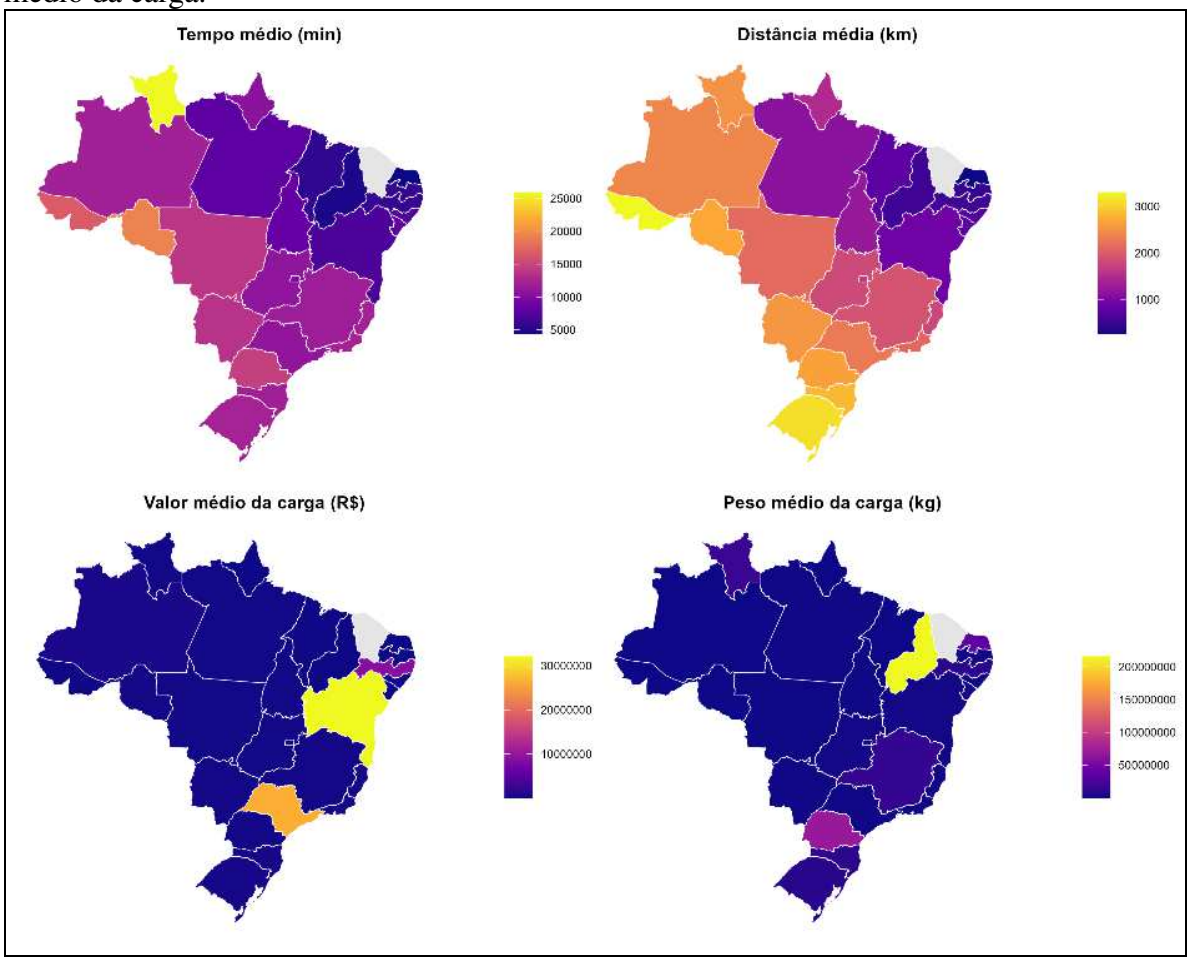


Fonte: Elaborado pelo autor.

No Gráfico 3 temos quatro mapas, o primeiro retrata o tempo médio (minutos), observa-se que estados mais distantes do Ceará registram maiores tempos médios de deslocamento, já os estados mais próximos apresentam os menores tempos, em razão da proximidade. No segundo mapa, a distribuição geográfica da distância média acompanha a do tempo, Norte, Sul e Centro-Oeste possuem os maiores trajetos até o Ceará, enquanto o entorno nordestino apresenta percursos mais curtos, o que faz bastante sentido.

Já no terceiro mapa, valor médio da carga (R$), tem-se informações interessantes com o estado da Bahia se destacando no envio de cargas de maior valor agregado, seguido por São Paulo (SP) e Pernambuco (PE). No quarto mapa, peso médio da carga (kg), Piauí (PI) e Paraná (PR) destacam-se com os maiores pesos médios das viagens.

Gráfico 3 - Mapa coroplético: tempo médio, distância média, valor médio da carga e peso médio da carga.



Fonte: Elaborado pelo autor.

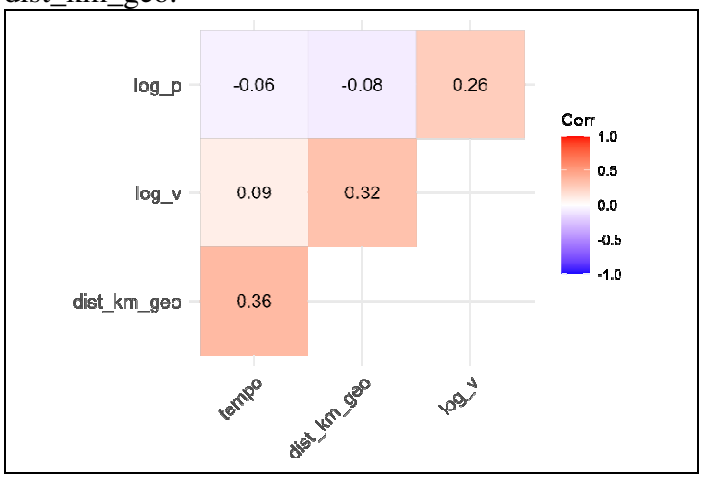
O Gráfico 4 apresenta a matriz de correlação entre variáveis numéricas relacionadas ao deslocamento de cargas: tempo de viagem (tempo), distância geodésica (dist_km_geo), logaritmo do peso (log_p) e do valor da carga (log_v).

Verifica-se que dist_km_geo apresenta a correlação mais significativa com tempo (r = 0,36), indicando que viagens mais longas tendem a durar mais. No entanto, o valor moderado desse coeficiente sugere a influência de outros fatores logísticos ou operacionais sobre a duração das viagens.

As correlações entre tempo e log_p (-0,06) e entre tempo e log_v (0,09) são muito fracas, indicando que peso e valor da carga não têm relação linear relevante com o tempo de deslocamento.

Tais resultados reforçam a importância do uso de técnicas de detecção de anomalias, capazes de identificar padrões atípicos no tempo de viagem não explicados por variáveis tradicionalmente consideradas.

Gráfico 4 – Matriz de correlação log_p, log_v e dist_km_geo.



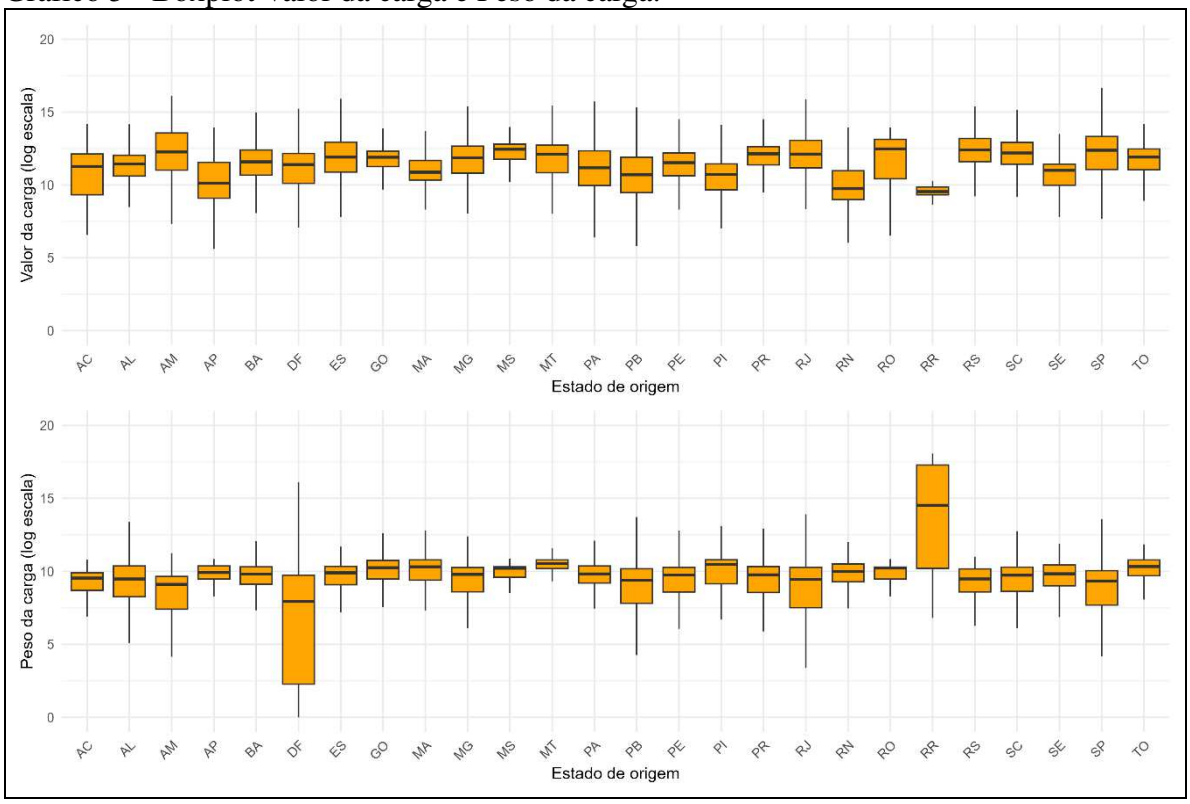
Fonte: Elaborado pelo autor.

O gráfico 5 apresenta dois diagramas de caixa (*boxplots*) com a distribuição do valor e do peso das cargas, conforme os estados de origem das mercadorias com destino ao Ceará.

No primeiro gráfico, referente ao valor da carga, observa-se que a mediana se mantém relativamente estável entre os estados, sugerindo certa homogeneidade no perfil econômico das cargas transportadas. Contudo, alguns estados, como Rondônia (RO) e Acre (AC), apresentam maior dispersão, indicando a presença de cargas com valores significativamente distintos — de itens de baixo a alto valor agregado. Em contraste, estados como Roraima (RR), Goiás (GO) e Mato Grosso do Sul (MS) revelam distribuições mais concentradas e com menor variabilidade.

No segundo gráfico, que representa o peso da carga, verifica-se uma maior heterogeneidade entre os estados. O estado de Roraima (RR) se destaca por apresentar cargas com pesos consideravelmente superiores aos demais, tanto pela mediana elevada quanto pela amplitude da distribuição. Já o Distrito Federal (DF) apresenta cargas significativamente mais leves e uma grande variação. Estados como Amapá (AP), Mato Grosso (MT), Mato Grosso do Sul (MS) e Rondônia (RO) exibem distribuições mais regulares, com variações moderadas e sem presença expressiva de outliers.

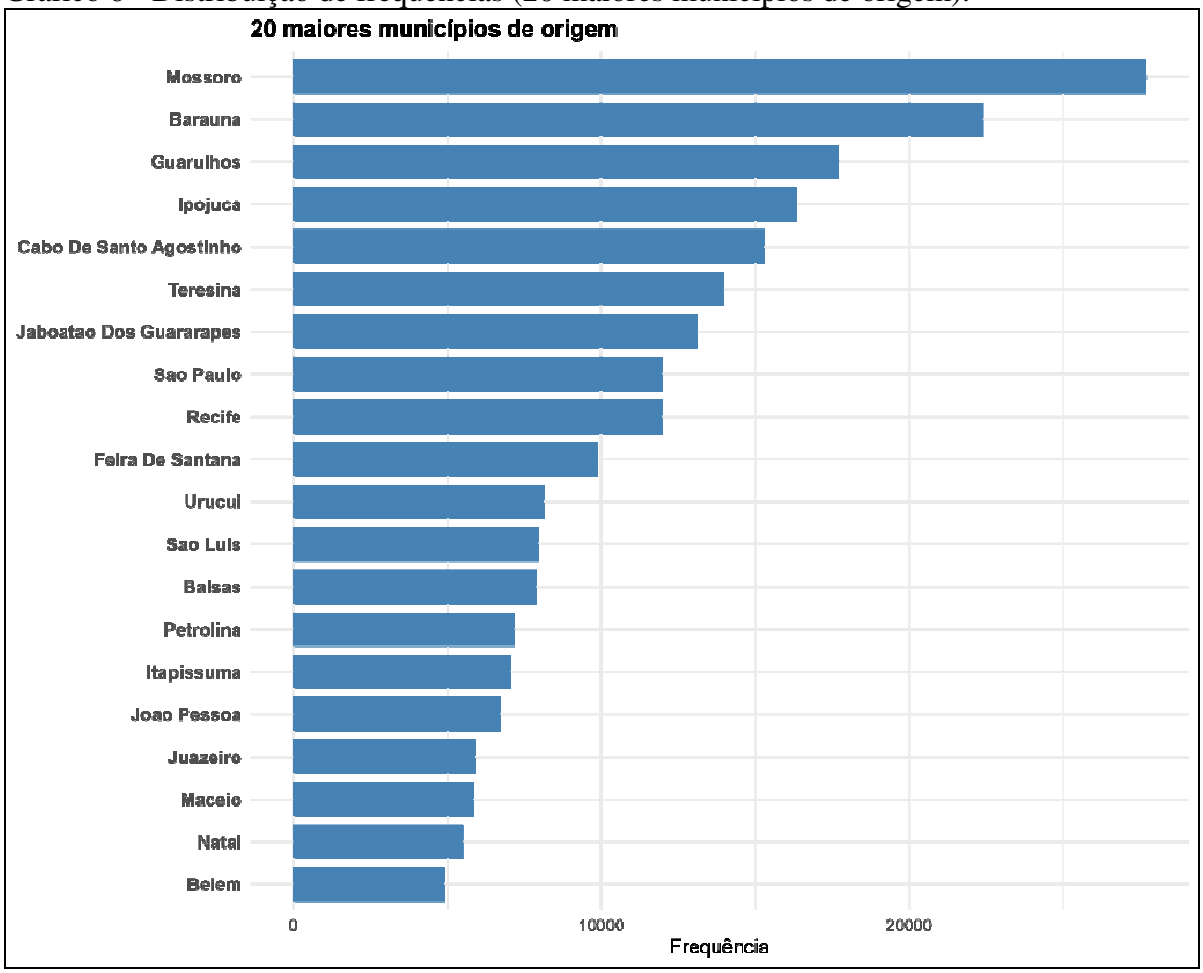
Gráfico 5 - Boxplot Valor da carga e Peso da carga.



Fonte: Elaborada pelo autor.

O Gráfico 6 traz os 20 municípios de origem com maior frequência de expedição de MDF-es, ou seja, os locais de onde mais partiram cargas com destino ao Ceará. Destaca-se os municípios do nordeste, em especial dos estados de Pernambuco, Rio Grande do Norte e Piauí, além dos municípios de Guarulhos, São Paulo e Belém.

Gráfico 6 - Distribuição de frequências (20 maiores municípios de origem).



Fonte: Elaborado pelo autor.

4. METODOLGIA

Nesta subseção, descreve-se a aplicação de dois métodos complementares de detecção de anomalias em tempos de transporte rodoviário registrados no Manifesto Eletrônico de Documentos Fiscais (MDF-e): a Regressão Linear Robusta (RLM) e o *Isolation Forest* (IF). O objetivo principal é identificar padrões incomuns no tempo de deslocamento, calculado como a diferença entre os horários de abertura e de fechamento do MDF-e, entendendo-se por anomalias as observações raras ou não compatíveis com o comportamento geral da amostra (Pang et al., 2021).

Embora o tempo de viagem aparente seja a métrica de interesse, é fundamental reconhecer que esse indicador pode ser influenciado por fatores estruturais, como distância percorrida, rota escolhida e tipo de carga. Em decorrência disso, classificações puramente univariadas podem resultar em falsos positivos; por exemplo, viagens longas naturalmente demandam mais tempo. Para mitigar esse viés, os modelos consideram simultaneamente as variáveis relevantes (distância, peso e valor da carga etc.) ou ajustam-se à distribuição geral dos tempos observados, permitindo diferenciar desvios pontuais de variações esperadas.

A ausência de um "rótulo" que identifique previamente quais registros são anômalos impõe o uso de métodos não supervisionados. Nesse contexto, ambos os algoritmos detectam automaticamente os casos que destoam da maioria das observações, sem necessidade de amostras rotuladas. A RLM atenua a influência de valores extremos sobre a estimação dos parâmetros de regressão, realçando desvios significativos na relação entre tempo e variáveis explicativas. Já o *Isolation Forest* identifica anomalias isolando iterativamente as observações em subárvores, de modo que pontos raros são separados com menos divisões (Liu et al., 2008).

Por fim, a aplicação conjunta de RLM e IF visa reforçar a confiabilidade dos achados: enquanto o *Isolation Forest* destaca casos isolados em um espaço multidimensional de atributos, a Regressão Linear Robusta sinaliza desvios em um modelo de dependência linear. A comparação contrastiva entre os resultados de ambos os métodos fornece uma avaliação mais robusta e ajuda a compreender a origem e a natureza das anomalias detectadas no transporte de mercadorias para o Ceará.

4.1 Engenharia de variáveis

A etapa inicial consistiu na construção de variáveis explicativas com potencial para influenciar o tempo de deslocamento das cargas. Dentre as variáveis derivadas, destaca-

se a *dummificação* das Unidades da Federação (UFs) percorridas, representadas pela variável binária passou_UF. Foram criadas 27 variáveis indicadoras, correspondentes a cada uma das unidades federativas brasileiras, marcando a presença ou ausência de passagem por determinado estado.

Para lidar com a assimetria e a heterocedasticidade de algumas variáveis contínuas, foram aplicadas transformações logarítmicas às variáveis valor da carga (v) e peso da carga (p). Além disso, foi calculada a velocidade implícita média do percurso, obtida pela razão entre a distância total e o tempo registrado no MDF-e, como uma proxy para padrões anômalos de desempenho logístico.

4.2 Modelos de detecção

Dois métodos distintos de detecção de anomalias foram aplicados, com a utilização do software R, de forma complementar, com o objetivo de aumentar a robustez dos achados: Regressão Linear Robusta (RLM) e *Isolation Forest* (IF). A seguir, descreve-se cada abordagem.

4.2.1. Regressão Linear Robusta (RLM)

O método de Regressão Linear Robusta baseado na função de perda de Huber distingue-se pela detecção de anomalias por meio do exame de resíduos padronizados. Trata-se de uma abordagem clássica de diagnóstico de regressão, que resiste à presença de outliers no processo de estimação e adapta-se bem a dados heterocedásticos, mas exige a suposição de linearidade na relação entre a variável dependente e os preditores. Para uma apresentação detalhada das propriedades teóricas deste estimador, veja Rousseeuw e Leroy (1987) e Hodge e Austin (2004).

Considere um conjunto de n observações $\{(y_i, X_i)\}_{i=1}^{n}$, em que y_i representa o tempo de viagem registrado pelo MDF-e e $X_i = (distância_i, \log(peso_i), \log(valor_i), \dots)$ é um vetor de características do transporte, incluindo variáveis contínuas e *dummies* para categoriais (mês, município de origem, município de destino, UFs percorridas). Em vez de minimizar a soma dos quadrados dos resíduos (MQO), a RLM obtém o vetor de parâmetros β que minimiza

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - X_i'\beta}{s}\right), \qquad (1)$$

onde s é uma estimativa de escala robusta e $\rho(\cdot)$ é a função de perda de Huber:

$$\rho(\omega) = \begin{cases} \frac{1}{2}\omega^2 & |\omega| \leq c \\ c|\omega| - \frac{1}{2}c^2 & |\omega| > c \end{cases}, \quad (2)$$

sendo o parâmetro de corte c geralmente fixado em 1,345 (aproximadamente 95% de eficiência sob normalidade).

A estimação utiliza o algoritmo Iteratively Reweighted Least Squares (IRLS), com ponto de partida $\beta^{(0)}$ igual à solução MQO. Em cada iteração j, calculam-se os resíduos padronizados

$$r_i^{(j-1)} = \frac{y_i - X_i'\beta^{(j-1)}}{s^{(j-1)}}, \quad (3)$$

e os pesos

$$w_i^{(j)} = \frac{\varphi\left(r_i^{(j-1)}\right)}{r_i^{(j-1)}}, \quad (4)$$

onde $\varphi(\omega) = \rho'(\omega)$ é a função de influência de Huber. Em seguida, resolve-se o problema

$$min_\beta \sum_{i=1}^{n} w_i^{(j)}(y_i - X_i'\beta)^2, \quad (5)$$

atualiza-se a estimativa de escala s e repete-se o procedimento até que $\left\|\beta^{(j)} - \beta^{(j-1)}\right\| < \delta$ ou até atingir o número máximo de iterações.

Após a convergência, calculam-se os resíduos padronizados finais,

$$\hat{r}_i = \left(y_i - X_i'\hat{\beta}\right)/\hat{s}, \quad (6)$$

e considera-se como outlier qualquer observação com $\hat{r}_i > 3$, critério clássico em diagnóstico robusto de regressão. Essa estratégia permite identificar desvios significativos na relação linear entre tempo de viagem e as variáveis explicativas, minimizando a influência de valores extremos na estimação dos parâmetros.

4.2.2. Isolation Forest (IF)

O *Isolation Forest* (IF) é um método não supervisionado de detecção de anomalias que se baseia no princípio de que observações "poucas e diferentes" são mais fáceis de isolar do que pontos típicos da amostra. Em vez de estimar densidades ou distâncias, o IF constrói uma floresta de árvores de partição aleatória (*iTrees*), em que cada árvore tenta isolar cada instância por meio de divisões recursivas no espaço de atributos (Liu et al., 2012).

Em cada *iTree*, seleciona-se aleatoriamente uma variável $k \in \{1,\ldots,K+1\}$ e um ponto

de corte $\boldsymbol{p}$ uniformemente entre o valor mínimo e máximo observado para essa *feature* na amostra de treinamento (tamanho ψ). Os dados com valor menor ou igual a $\boldsymbol{p}$ seguem para o nó esquerdo e os demais para o direito. O processo repete-se até que cada observação $Z_i = \{(y_i, X_i)\}_{i=1}^{n} \in \mathbb{R}^{K+1}$ seja isolada (formando um nó folha) ou até atingir a profundidade máxima $D_{max} = log_2(\psi)$ (Liu et al., 2012).

Para cada observação Z_i, calcula-se sua profundidade de isolamento $h(Z_i)$ em cada árvore e depois obtém-se a média $E[h(Z_i)]$. A pontuação de anomalia é então definida como

$$\pi(Z_i) = 2^{-\frac{E[h(Z_i)]}{c(\psi)}}, \tag{7}$$

onde

$$c(\psi) = 2\left(H_{\psi-1} - \frac{(\psi-1)}{\psi}\right), \tag{8}$$

é o fator de normalização que aproxima o comprimento médio de busca em uma árvore binária de ψ nós, e $H_{\psi-1}$ é o $(\psi - 1)$-ésimo número harmônico (Liu et al., 2012). Valores de $\pi(Z_i)$ próximos de 1 indicam isolamento rápido (forte suspeita de anomalia), enquanto valores próximos a 0,5 refletem comportamento típico.

Nesta aplicação, utilizamos ψ =256 e número de árvores construídas igual a 500, conforme sugerido em Liu et al. (2008) para equilibrar sensibilidade e custo computacional. Antes do treino, mantivemos apenas variáveis numéricas relevantes (distância geodésica, log(peso), log(valor), velocidade média) e as dummies para município e UF. Eliminamos variáveis de variância quase nula e imputamos valores faltantes pela mediana. Em seguida, cada atributo contínuo foi centralizado em zero e padronizado para desvio-padrão um, evitando que variáveis de maior escala influenciassem excessivamente as divisões aleatórias.

Como critério de classificação, consideramos como anomalias as observações cuja pontuação $\pi(Z_i)$ supera o percentil 95 da distribuição de scores na amostra, correspondendo aos 5 % mais “isolados”. Essa escolha equilibra a detecção de casos verdadeiramente atípicos e a contenção de falsos positivos.

Em comparação com a Regressão Linear Robusta, o IF captura interações e padrões não lineares sem assumir forma funcional específica, mas sacrifica parcialmente a interpretabilidade, pois não fornece coeficientes nem relações explícitas entre preditores e resposta. Por isso, adotamos ambas as técnicas de forma complementar: a RLM destaca

desvios na dependência linear clássica, enquanto o IF identifica estruturas de anomalia multidimensionais mais complexas (Aggarwal e Sathe, 2015; Hariri et al., 2019).

4.3 Consolidação

Após a execução dos dois métodos, os resultados foram consolidados por meio de uma regra de votação. Cada observação foi marcada com os métodos que a classificaram como anômala (RLM, IF, ou combinações entre eles), resultando na variável categórica *"metodo_outlier"*.

Além disso, foi criada uma métrica-alvo binária, denominada *"outlier_any"*, que assume o valor 1 caso a observação tenha sido considerada anômala por pelo menos um dos métodos aplicados, e 0 caso contrário. Esta variável foi utilizada como principal indicador na análise estatística e espacial das anomalias detectadas ao longo do território nacional.

Cabe destacar que, até o presente momento, não foram conduzidos procedimentos formais de validação externa ou testes de consistência para os modelos aplicados. A ausência de uma etapa de validação cruzada, comparação com dados de telemetria (como GPS), ou verificação manual de casos anômalos limita a generalização e a robustez das conclusões. A incorporação desses procedimentos em estudos futuros poderá fortalecer a credibilidade dos achados e fornecer subsídios mais confiáveis para aplicações práticas na fiscalização do trânsito de mercadorias.

5. ANÁLISE DOS RESULTADOS

5.1 Frequência geral de anomalias

A aplicação dos métodos de detecção de anomalias aos dados de tempo de deslocamento de cargas com destino ao estado do Ceará revelou percentuais distintos de identificação de observações atípicas. O método de Regressão Linear Múltipla (RLM) identificou 4,2% de anomalias na base analisada, já o método *Isolation Forest* (IF) apresentou uma frequência de anomalias de 5,0%.

Quando considerada a interseção dos resultados, ou seja, observações identificadas como anômalas pelos dois métodos simultaneamente, o percentual foi reduzido para 1,8%. Essa interseção representa os casos com maior robustez na identificação de padrões anômalos, indicando uma convergência entre os diferentes algoritmos quanto à atipicidade de determinados registros.

5.2 Padrões espaciais

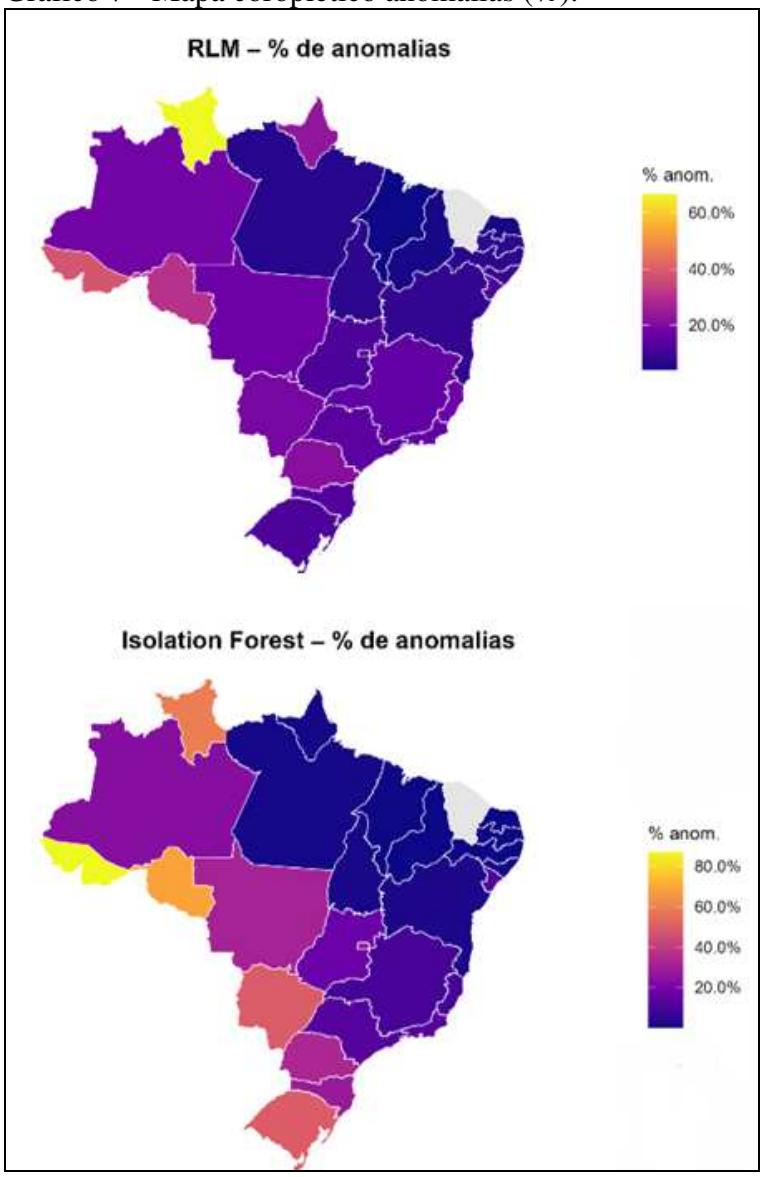
O Gráfico 7 apresenta mapas com a distribuição espacial das anomalias por unidade federativa de origem, de acordo com os dois métodos utilizados. Nota-se que a incidência de anomalias não é uniforme entre os estados. Em particular, observou-se que o estado de Roraima apresentou os maiores percentuais de anomalias segundo o método RLM, ultrapassando 60% das observações, seguido por Acre e Rondônia. Já o método IF indicou elevados índices de anomalias nos mesmos estados, mas na seguinte ordem: Acre, Rondônia e Roraima, sendo de aproximadamente 80%, 60% e 60%.

A recorrência de altos percentuais de anomalias nesses estados, especialmente quando detectados por mais de um método, sugere a existência de padrões atípicos persistentes. Tais padrões podem estar relacionados a diferentes fatores, como limitações na infraestrutura rodoviária, distâncias mais elevadas, uso de rotas alternativas (anomalias logísticas), ou mesmo inconsistências nos registros documentais (anomalias documentais). A presença simultânea desses fatores tende a influenciar significativamente o tempo de deslocamento das cargas, justificando a maior incidência de registros classificados como anômalos.

Diante desse cenário, destaca-se a importância de uma investigação mais aprofundada sobre os fluxos logísticos provenientes dessas unidades federativas. A análise detalhada poderá contribuir para ajustes metodológicos nos modelos preditivos, bem como

para a melhoria da qualidade dos dados utilizados na gestão do transporte rodoviário de cargas, especialmente em contextos de fiscalização e planejamento tributário interestadual.

Gráfico 7 - Mapa coroplético anomalias (%).



Fonte: Elaborado pelo autor.

5.3 Padrões sazonais

O Gráfico 8 apresenta a frequência mensal de anomalias identificadas pelos dois métodos (RLM e ISO) ao longo dos doze meses do ano. O eixo x representa o mês de início do transporte, enquanto o eixo y indica o percentual de anomalias detectadas em relação ao total de observações do mês.

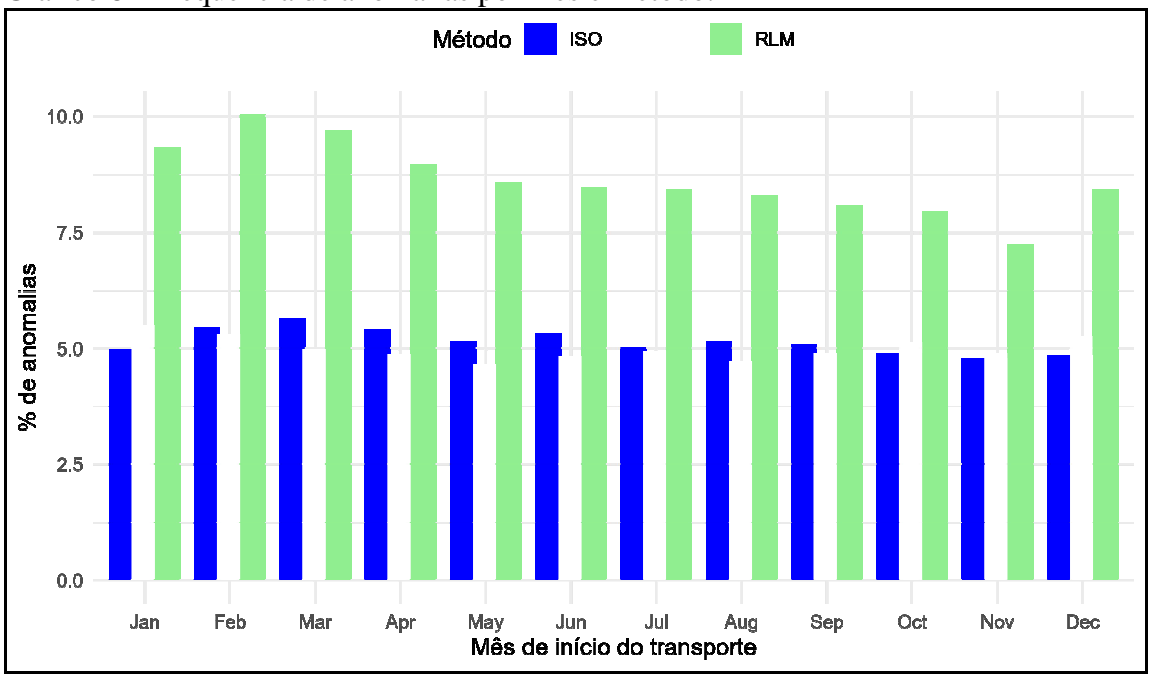
Observa-se que o método RLM (barras verdes) apresenta, de forma sistemática, os maiores percentuais de anomalias em todos os meses, variando aproximadamente entre 7% e 10%. Essa maior sensibilidade pode indicar que o modelo estatístico capta desvios em relação

à média esperada com maior intensidade, embora nem sempre esses desvios sejam considerados anômalos por métodos mais sofisticados de aprendizado de máquina. Já o método IF (azul) apresentou resultados muito próximos ao longo dos meses, oscilando entre 5% e 6% de anomalias.

Não se verifica uma variação significativa ao longo dos meses, o que indica estabilidade na ocorrência de anomalias ao longo do tempo. Isso pode apontar para uma distribuição relativamente homogênea dos fatores que geram desvios, como atrasos, erros de registro ou variações nas rotas utilizadas.

O gráfico revela que a detecção de anomalias nos dados de transporte apresenta padrões estáveis ao longo do ano, com RLM identificando mais anomalias do que o método IF, e que os meses de fevereiro e março foram os picos das observações. A diferença entre os métodos reforça a importância de utilizar abordagens complementares, especialmente quando se busca uma triangulação robusta para identificar registros efetivamente anômalos.

Gráfico 8 - Frequência de anomalias por mês e método.



Fonte: Elaborado pelo autor.

5.4 Relação com valor/peso

O Gráfico 9 apresenta a comparação entre registros classificados como normais e anômalos quanto ao valor e ao peso médio das cargas transportadas.

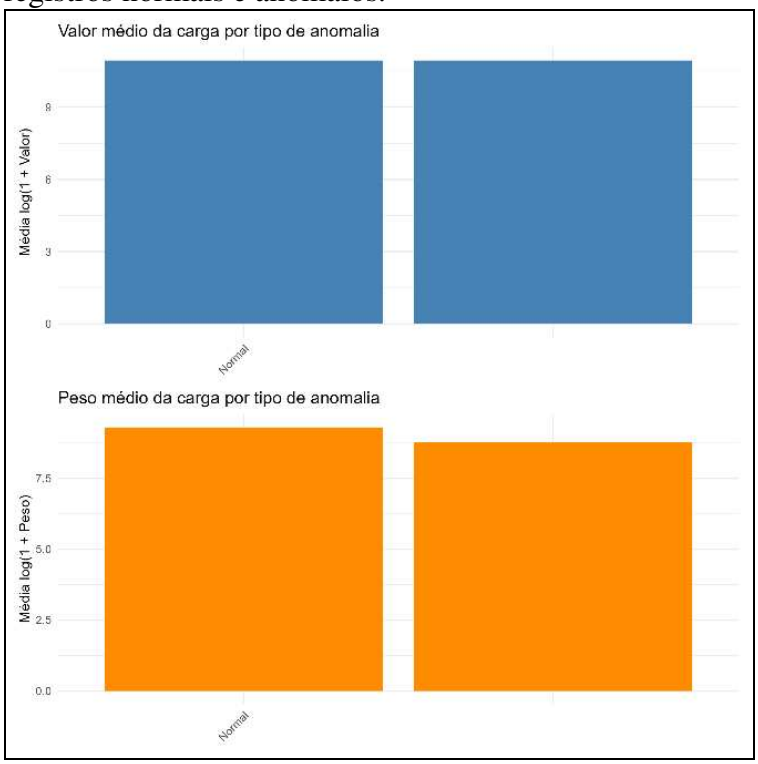
No primeiro painel, observa-se a distribuição do valor médio das cargas por tipo de anomalia, os resultados indicam que não há diferenças substanciais entre os registros normais e os registros identificados como anômalos. Essa constatação sugere que o valor

monetário da carga transportada não constitui um fator discriminante relevante na ocorrência de anomalias nos tempos de deslocamento analisados.

De forma semelhante, o segundo painel apresenta a comparação do peso médio das cargas entre os dois grupos. Assim como observado no caso do valor, os resultados não evidenciam variações significativas no peso das cargas entre os registros normais e os anômalos. Essa estabilidade entre os grupos aponta para a ausência de correlação relevante entre a massa transportada e a probabilidade de ocorrência de padrões atípicos.

Diante desses achados, infere-se que as anomalias detectadas nos tempos de deslocamento das cargas não estão associadas diretamente às características como peso ou valor declarado. Tais resultados reforçam a hipótese de que outros fatores, como origem geográfica, distância percorrida, condições operacionais da malha rodoviária ou possíveis inconsistências nos registros fiscais, exercem maior influência na identificação de desvios temporais no transporte.

Gráfico 9 - Valor médio e peso médio das cargas entre registros normais e anômalos.



Fonte: Elaborado pelo autor.

6. CONCLUSÃO

O presente trabalho investigou o uso de técnicas de aprendizado de máquina aplicadas aos dados do Manifesto Eletrônico de Documentos Fiscais (MDF-e), com o objetivo de identificar padrões e anomalias temporais nos deslocamentos rodoviários de cargas com destino ao estado do Ceará, de modo a contribuir com a fiscalização tributária da Secretaria da Fazenda do Estado do Ceará (SEFAZ-CE).

Foi possível identificar os estados de origem e os períodos do ano com maior concentração de anomalias, permitindo traçar um panorama geográfico e temporal do comportamento atípico dos deslocamentos, apesar da pouca variação no tempo. Por outro lado, observou-se que o valor e o peso da carga, não apresentaram correlação expressiva com o tempo de deslocamento. Essas informações são particularmente relevantes para o planejamento das ações de fiscalização, permitindo que a SEFAZ-CE atue de forma mais estratégica.

Dessa forma, os padrões extraídos por meio dos modelos de detecção de anomalias podem orientar o direcionamento das ações de fiscalização no trânsito das mercadorias pela SEFAZ-CE, inclusive com antecipação, já que no momento da abertura do MDF-e, antes da chegada das mercadorias no estado, a secretaria já recebe os dados utilizados nos modelos; podendo, assim, priorizar rotas e transportadoras com maior incidência de anomalias. Isso viabilizaria uma abordagem mais eficiente na alocação de recursos públicos e no combate à evasão fiscal, fortalecendo a arrecadação estadual.

Apesar dos avanços obtidos, este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados e na formulação de políticas públicas. Em primeiro lugar, destaca-se a ausência de dados de geolocalização (GPS) ou telemetria embarcada, o que limita a precisão na verificação do trajeto real percorrido e do tempo efetivo de deslocamento. A análise baseou-se exclusivamente nas informações declaradas no MDF-e, que podem não refletir com exatidão a dinâmica logística das operações.

Também não foi possível obter informações de câmeras espalhadas pelo país que se conectam ao Operador Nacional dos Estados (ONE), sistema responsável por integrar os MDF-es, que registram a data e hora das passagens de caminhões que possuem manifesto de carga aberto, isso talvez reduzisse os falsos positivos na detecção de anomalias.

Além disso, erros ou inconsistências nos registros do MDF-e — como horários incorretos de início e fim da viagem — podem comprometer a qualidade das estimativas. Tais problemas decorrem de falhas humanas ou dos sistemas de emissão e não puderam ser

corrigidos integralmente durante o tratamento dos dados.

Outro ponto a ser considerado é a generalização dos resultados. O modelo foi calibrado com dados de um período e região específicos (transporte rodoviário com destino ao Ceará em determinado ano), o que pode limitar a aplicabilidade dos achados a outros contextos temporais e geográficos. Também não foi possível aplicar procedimentos formais de validação externa ou testes de consistência para os modelos utilizados.

Essas limitações não invalidam as contribuições do estudo, mas indicam caminhos para o seu aprimoramento em pesquisas futuras. A incorporação de dados externos, como rotas estimadas por APIs como a do *Open Source Routing Machine* (OSRM) ou dados reais de GPS; a incorporação de dados internos, data e hora das passagens de caminhões, registradas nas câmeras associadas ao ONE; bem como a aplicação de modelos mais sofisticados de aprendizado não supervisionado, como *AutoEncoders* e a ampliação da amostra para diferentes anos, podem aumentar a robustez dos modelos e a confiabilidade das inferências, separando as anomalias meramente logísticas das anomalias potencialmente fraudulentas.

Do ponto de vista de política fiscal, este estudo reforça a importância de medidas que ampliem a rastreabilidade das operações de transporte, e da discussão do modo que o MDF-e está sendo preenchido atualmente, se ele realmente permite a fiscalização em tempo real das mercadorias transportadas, visto que essa é a sua função. Indaga-se se o tempo de 30 dias concedidos para o encerramento de um MDF-e é necessário e razoável, e se esse período não poderia ser reduzido.

Assim, conclui-se que a utilização analítica dos dados fiscais eletrônicos, aliada a ferramentas de ciência de dados, pode desempenhar papel relevante na modernização e eficácia da gestão tributária estadual. A continuidade e o aprofundamento dessas iniciativas representam uma oportunidade estratégica para a SEFAZ-CE no enfrentamento a fraudes no transporte de cargas e das perdas de receita com ICMS, fortalecendo, assim, a arrecadação estadual e promovendo maior equidade tributária.

REFERÊNCIAS

AGGARWAL, Charu; SATHE, Saket. (2015). ***Theoretical Foundations and Algorithms for Outlier Ensembles?***. *ACM SIGKDD Explorations Newsletter*. 17. 24-47. 10.1145/2830544.2830549.

ALSADHAN, Nasser. (2023). **A Multi-Module Machine Learning Approach to Detect Tax Fraud**. Computer Systems Science and Engineering. 46. 241-253. 10.32604/csse.2023.033375. disponível em: https://www.researchgate.net/publication/367405432_A_Multi-Module_Machine_Learning_Approach_to_Detect_Tax_Fraud. Acesso em: 07 julho 2025.

AYUB, Mohammed; ANNARUMMA, Mauro; LIBERMAN, Gaston; et al., **Efficient and Explainable Anomaly Detection in Financial System Logs Using Statistical Modeling**. 2025 IEEE/ACM 2nd Workshop on Software Engineering Challenges in Financial Firms (FinanSE), Ottawa, ON, Canadá, 2025, pp. 7-14, doi: 10.1109/FinanSE66659.2025.00006.

BAGHDASARYAN, Vardan; DAVTYAN, Hrant; SARIKYAN, Arsine; NAVASARDYAN, Zaruhi. (2022). **Improving Tax Audit Efficiency Using Machine Learning: The Role of Taxpayer's Network Data in Fraud Detection.** *Applied Artificial Intelligence*, *36*(1). Disponível em: https://doi.org/10.1080/08839514.2021.2012002. Acesso em: 13 julho 2025.

BASTOS, Cleverson Leite; KELLER, Vicente. **Aprendendo a aprender**: introdução à metodologia científica. 19. ed. Petrópolis: Vozes, 2006.

BITTENCOURT NETO, Sérgio Augusto Pará. **Análise de "outliers" para controle do risco de evasão tributária do ICMS**. Brasília, 2018. Disponível em: http://repositorio2.unb.br/bitstream/10482/33031/1/2018_S%C3%A9rgioAugustoPar%C3%A1BittencourtNeto.pdf.

BOTTER, Rui Carlos; TACLA, Douglas; HINO, Celso Mitsuo. **Estudo e aplicação de transporte colaborativo para cargas de grande volume.** São Paulo, 2006. Disponível em: https://www.scielo.br/j/pope/a/vxqQFxHQDLxCmNjDVpjgW8n/. Acesso em: 25 jun 2025.

BRASIL. Lei nº 5.172, de 25 de outubro de 1966. **Dispõe sobre o Sistema Tributário Nacional e institui normas gerais de direito tributário aplicáveis à União, Estados e Municípios**. Diário Oficial da União: seção 1, Brasília, DF, 27 out. 1966. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l5172compilado.htm. Acesso em: 16 maio 2025.

BRASIL. Lei nº 11.442, de 5 de janeiro de 2007. **Dispõe sobre o transporte rodoviário de cargas por conta de terceiros e mediante remuneração**. Diário Oficial da União, Brasília, DF, 08 jan 2007.

CASTRO, Newton Rabello de. **Formação de preços no transporte de carga.** Rio de Janeiro, 2003. Disponível em: https://repositorio.ipea.gov.br/handle/11058/4297. Acesso em: 25 jun 2025.

CEARÁ. Lei nº 18.665, de 28 de dezembro de 2023. **Dispõe acerca do ICMS**. Diário Oficial do Estado, Fortaleza, CE, 28 dez. 2023.

CEARÁ. Decreto nº 35.061, de 21 de dezembro de 2022. Diário Oficial do Estado, Fortaleza, CE, 21 dez. 2022.

CEARÁ. Decreto nº 34.605, de 24 de março de 2022. Diário Oficial do Estado, Fortaleza, CE, 24 mar. 2022.

CEARÁ. Decreto nº 24.569, de 31 de julho de 1997. Diário Oficial do Estado, Fortaleza, CE, 08 ago. 1997.

CONFAZ - CONSELHO NACIONAL DE POLÍTICA FAZENDÁRIA. Ajuste SINIEF 21, de 10 de dezembro de 2010. **Institui o Manifesto Eletrônico de Documentos Fiscais (MDF-e)**. Disponível em: https://www.confaz.fazenda.gov.br/legislacao/ajustes/2010/aj021_10. Acesso em: 11 maio 2025.

GONZÁLEZ, Pamela Castellón; VELÁSQUEZ, Juan D. ***Characterization and detection of taxpayers with false invoices using data mining techniques***. *Expert Systems with Applications*, Volume 40, Issue 5, 2013, Pages 1427-1436, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2012.08.051.

HARIRI, Sahand; KIND, Matias; BRUNNER, Robert. (2019). ***Extended Isolation Forest with Randomly Oriented Hyperplanes***. *IEEE Transactions on Knowledge and Data Engineering*. PP. 1-1. 10.1109/TKDE.2019.2947676.

HODGE, Victoria; AUSTIN, Jim. (2004). ***A Survey of Outlier Detection Methodologies***. *Artificial Intelligence Review.* 22. 85-126. 10.1023/B:AIRE.0000045502.10941.a9.

HUERTAS, Daniel Monteiro. **Gênese e expansão dos agentes do transporte rodoviário de carga no Brasil.** Osasco, 2025. Disponível em: https://www.scielo.br/j/ecos/a/HQNZkj3qdfj4FdJrzRSfFZC/?lang=pt. Acesso em: 24 jun 2025.

LEAL, Carlos Chagastelis Martins. **Transporte de carga no Distrito Federal Questões e desafios**. Brasília, 2018. Disponível em: https://www.periodicos.capes.gov.br/index.php/acervo/buscador.html?task=detalhes&source=all&id=W4312917670. Acesso em: 14 maio 2025.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. **Isolation Forest.** In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008. p. 413–422. DOI: 10.1109/ICDM.2008.17.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. ***Isolation-based anomaly detection***. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 1 – 39. 2012.

LUDIVIA, Hernandez Aros; LUISA XIMENA, Bustamante Molano; FERNANDO, Gutierrez-Portela; et al. **Financial fraud detection through the application of machine learning techniques: a literature review**. Humanit Soc Sci Commun 11, 1130, 2024. Disponível em: https://doi.org/10.1057/s41599-024-03606-0.

MUHAMMAD, Atif Khan Achakzai; PENG Juan. **Using machine learning Meta-Classifiers to detect financial frauds**. Finance Research Letters, Volume 48, 2022, 102915,

ISSN 1544-6123, Disponível em: https://www.sciencedirect.com/science/article/pii/S1544612322001866. Acesso em: 06 julho 2025.

OPERADOR NACIONAL DOS ESTADOS. Sobre. Portal do Operador Nacional dos Estados. Disponível em: https://dfe-portal.svrs.rs.gov.br/One. Acesso em: 01 julho 2025.

OUNACER, Soumaya et al. (2018). ***Using Isolation Forest in anomaly detection: the case of credit card transactions.*** *Periodicals of Engineering and Natural Sciences.* Vol.6, No.2, December 2018, pp.394-400. International University of Sarajevo. doi: 10.21533/PEN.V6I2.533.

PANG, Guansong; SHEN, Chunhua; CAO, Longbing; VAN DEN HENGEL, Anton. ***Deep learning for anomaly detection: A review***. *ACM Computing Surveys*, 54(2):1 – 38, 2021.

PORTAL NACIONAL DA MDF-e. **Manual de Orientação do Contribuinte – MDF-e** – Versão 3.00a. Brasília, DF: ENCAT/SEFAZ, 2023. Disponível em: https://dfe-portal.svrs.rs.gov.br/Mdfe. Acesso em: 11 maio 2025.

ROUSSEEUW, Peter; LEROY, Annick. (1987). **Robust Regression and Outlier Detection**. John Wiley & Sons. Journal of Educational Statistics. 13. 358-364.

SAVIĆ, Miloš; ATANASIJEVIĆ, Jasna; JAKOVETIĆ, Dušan; KREJIĆ, Nataša. **Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method, Expert Systems with Applications**. Volume 193, 2022, 116409, ISSN 0957-4174. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417421016973.

SOUZA, A. A.; JÚNIOR, A. A. F.; SILVA, D. A.; ANDRADE, J. V. R. **Detecção de Anomalias em Aplicações de Monitoramento de Sistemas utilizando Isolation Forest**. Revista Processando o Saber, [s. l.], v. 17, n. 01, 21-37, 6 jun. 2025. DOI 10.5281/zenodo.15477217. Disponível em: https://www.fatecpg.edu.br/revista/index.php/ps/article/view/341. Acesso em: 6 julho 2025.

TOSTA, Pedro Lucas Maranini; DIAS, Jonatas Cerqueira. **Detecção de fraudes em transações bancárias utilizando inteligência artificial.** Revista Processando o Saber, São Paulo, v.17, p. 21-37, junho, 2025. Disponível em: https://www.fatecpg.edu.br/revista/index.php/ps/article/download/341/256/1583.

VANHOEYVELD, Jellis; MARTENS, David; PEETERS, Bruno. (2019). **Value-added tax fraud detection with scalable anomaly detection techniques.** Applied Soft Computing. 86. 105895. 10.1016/j.asoc.2019.105895. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1568494619306763#preview-section-references. Acesso em: 12 julho 2025.

VANINI, Paullo *et al.* **Online payment fraud: from anomaly detection to risk management.** *Financ Innov* 9, 66 (2023). https://doi.org/10.1186/s40854-023-00470-w.

APÊNDICE

Tabela 3 - Estatística descritiva completa.

variável	obs	Na	mean	sd	min	q25	median	q75	max
cc	NA	NA	2802406	734563.7	1100023	2403103	2607307	3125101	5300108
cc_n	574805	0	NA	NA	NA	NA	NA	NA	NA
cd	NA	NA	2306023	3133.881	2300101	2304400	2304400	2307650	2314102
cd_n	574805	0	NA	NA	NA	NA	NA	NA	NA
dist_km_geo	NA	NA	1025.743	799.709	17.75666	470.1162	652.1247	1726.711	3888.237
dist_km_geo_n	574805	0	NA	NA	NA	NA	NA	NA	NA
lat_cc	NA	NA	-11.3019	7.126528	-32.2172	-16.0891	-8.23694	-6.11742	3.117848
lat_cc_n	574805	0	NA	NA	NA	NA	NA	NA	NA
lat_cd	NA	NA	-4.36295	1.143312	-7.78034	-4.10217	-3.81108	-3.78574	-2.8777
lat_cd_n	574805	0	NA	NA	NA	NA	NA	NA	NA
log_p	NA	NA	9.215092	2.571826	0	8.721113	9.818256	10.37727	30.02892
log_p_n	574805	0	NA	NA	NA	NA	NA	NA	NA
log_v	NA	NA	10.938	2.078292	0	10.06901	11.29021	12.23077	28.20763
log_v_n	574805	0	NA	NA	NA	NA	NA	NA	NA
lon_cc	NA	NA	-41.3351	5.491446	-73.4391	-46.4548	-40.2745	-36.1805	-34.8417
lon_cc_n	574805	0	NA	NA	NA	NA	NA	NA	NA
lon_cd	NA	NA	-38.8098	0.609791	-41.2327	-39.0605	-38.528	-38.528	-37.4105
lon_cd_n	574805	0	NA	NA	NA	NA	NA	NA	NA
p	NA	NA	29595844	1.5E+10	0	6130	18365	32120	1.1E+13
p_n	574805	0	NA	NA	NA	NA	NA	NA	NA
tempo	NA	NA	7279.753	7596.05	60.1	2747.7	5419.633	8838.45	87732.73
tempo_n	574805	0	NA	NA	NA	NA	NA	NA	NA
v	NA	NA	8070737	3.54E+09	0	23599.14	80033.51	205000	1.78E+12
v_n	574805	0	NA	NA	NA	NA	NA	NA	NA
vel_kmh	NA	NA	16.21059	54.36095	0.017426	5.448809	9.800304	15.98674	2865.962
vel_kmh_n	574805	0	NA	NA	NA	NA	NA	NA	NA

Fonte: Elaborada pelo autor.