Desempenho Logístico no Brasil: Análise Geoespacial sobre Prazos de Entrega, Custos de Frete e Satisfação do Cliente no E-commerce

Lucas Bruno F. Praxedes[1] Isaac de L. Oliveira Filho[1]

[1]Programa de Pós Graduação em Ciência da Computação – Universidade Federal Rural do Semiárido e Universidade do Estado do Rio Grande do Norte (UFERSA-UERN)
Mossoró – RN – Brazil

lucas.praxedes@alunos.ufersa.edu.br, isaacoliveira@uern.br

Abstract. *This study performs a quantitative analysis of logistics performance in Brazilian e-commerce, using a large transactional dataset. Through geospatial analysis, statistical tests, and machine learning, critical delivery routes were identified, the negative and significant impact of delays on customer satisfaction was quantified, and a predictive model for freight cost was developed. The results demonstrate that product weight is a more decisive factor than distance in determining the cost, providing important information for the optimization of logistics strategies and for the improvement of the consumer experience.*

Resumo. *Este estudo realiza uma análise quantitativa do desempenho logístico no e-commerce brasileiro, utilizando um grande conjunto de dados transacionais. Por meio de análise geoespacial, testes estatísticos e machine learning, foram identificadas rotas de entrega críticas, quantificado o impacto negativo e significativo dos atrasos na satisfação do cliente, e desenvolvido um modelo preditivo para o custo do frete. Os resultados demonstram que o peso do produto é um fator mais decisivo que a distância na determinação do custo, fornecendo informações importantes para a otimização de estratégias logísticas e para a melhoria da experiência do consumidor.*

1. Introdução

O crescimento do *e-commerce* brasileiro impôs um desafio logístico de grande magnitude, um componente central do "Custo Brasil" (ABComm, 2024; ILOS, 2023). Em um país de dimensões continentais e infraestrutura heterogênea, a performance da entrega, custo e prazo, tornou-se um pilar para o sucesso e uma grande vantagem competitiva, impactando diretamente a conversão de vendas e a satisfação do cliente (World Bank, 2023; NielsenIQ Ebit, 2023; Parasuraman et al., 1988).

Apesar da relevância do tema, estudos quantitativos que mapeiam os gargalos logísticos em escala nacional com base em dados transacionais reais são escassos na literatura. Esta pesquisa busca preencher essa lacuna, analisando a fundo o desempenho logístico no Brasil para identificar os fatores que governam os custos, os prazos e, em última instância, a satisfação do cliente. Os objetivos são: (1) mapear o fluxo de pedidos; (2) desenvolver um modelo preditivo para o custo do frete; (3) identificar as rotas mais problemáticas em termos de atraso; e (4) quantificar o impacto da pontualidade na satisfação do cliente. A investigação é guiada pelas seguintes hipóteses:

H1: As características físicas do produto são preditores mais fortes do custo de frete do que a distância geográfica.

H2: Atrasos na entrega têm um impacto negativo e estatisticamente significativo na avaliação do cliente.

H3: Rotas inter-regionais apresentam desempenho logístico inferior a rotas intrarregionais.

2. Conceitos Fundamentais

2.1 Desafios Logísticos e a Experiência do Cliente no *E-commerce*

A gestão da cadeia de suprimentos no Brasil é uma tarefa de alta complexidade. A literatura aponta a logística como um dos principais entraves operacionais, destacando-se a massiva dependência do modal rodoviário, que responde por mais de 60% da matriz de transportes de cargas do país (CNT, 2022). Essa dependência é agravada pela condição frequentemente precária da malha viária e pelos elevados custos operacionais, que impactam diretamente os prazos e a previsibilidade das entregas (ILOS, 2023). Soma-se a isso o desafio da "última milha" (*last-mile delivery*), que, especialmente em grandes centros urbanos e áreas remotas, adiciona uma camada de complexidade com custos elevados e riscos de segurança (Duarte et al., 2019).

Nesse contexto, onde a entrega é o principal ponto de contato físico entre a marca e o consumidor, a performance logística transcende a operação e compõe a estrutura da experiência do cliente. A teoria de qualidade de serviço (SERVQUAL) postula que a confiabilidade, que é a capacidade de executar o serviço prometido de forma precisa, é uma dimensão central na percepção do cliente (Parasuraman et al., 1988). No *e-commerce*, isso se traduz diretamente no cumprimento do prazo de entrega. Atrasos, portanto, não são meros inconvenientes operacionais; eles minam a confiança e a satisfação, afetando negativamente a lealdade e a probabilidade de recompra (Agatz et al., 2008; Esper et al., 2003).

2.2 Abordagens Quantitativas para Análise Logística

Para decifrar a complexa dinâmica logística, a pesquisa se apoia em um conjunto de ferramentas quantitativas. No campo da modelagem preditiva, o uso de *Machine Learning* é fundamental, especialmente para prever custos e estimar tempos de chegada (ETA) (Rokoss et al., 2024). Algoritmos de *Gradient Boosting*, como o LightGBM, são bastante eficazes para dados tabulares, superando modelos de regressão tradicionais devido à sua capacidade de capturar interações não lineares complexas entre as variáveis (Kim et al., 2023). O princípio do LightGBM é construir um modelo preditivo de forma aditiva, onde cada novo modelo fraco (uma árvore de decisão) é treinado para corrigir os erros do anterior. Este processo é representado pela equação 1:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$$

Onde o modelo final $F_m(x)$ é o resultado da soma do modelo anterior $F_{m-1}(x)$ com uma nova árvore $h_m(x)$ ponderada por uma taxa de aprendizado v. Paralelamente, a análise geoespacial emprega métricas como a fórmula de Haversine para calcular a

distância ortodrômica, que é o caminho mais curto sobre a superfície de uma esfera, entre as coordenadas de origem e destino. Por fim, para validar hipóteses e quantificar relações, são empregadas ferramentas estatísticas consolidadas: o Teste *t* de *Student* é usado para verificar se a diferença na satisfação média entre grupos é estatisticamente significativa, enquanto a Correlação de *Pearson* mede a força e a direção da relação linear entre variáveis, como os dias de atraso e as notas de avaliação.

3. Metodologia

Este estudo caracteriza-se como uma pesquisa aplicada, de abordagem quantitativa e com objetivos descritivos e explicativos. Foi conduzido como um estudo documental, analisando dados secundários provenientes do "*Brazilian E-Commerce Public Dataset by Olist*". Os dados utilizados foram coletados entre 2016 e 2018, e correspondem a aproximadamente 100 mil pedidos reais de e-commerce brasileiro, cujas informações foram anonimizadas. A análise foi realizada em *Python*, com o suporte de seu ecossistema de bibliotecas científicas (VanderPlas, 2016), como *Pandas, GeoPandas, Scikit-learn* e LightGBM.

3.1. Fluxo de Trabalho da Pesquisa

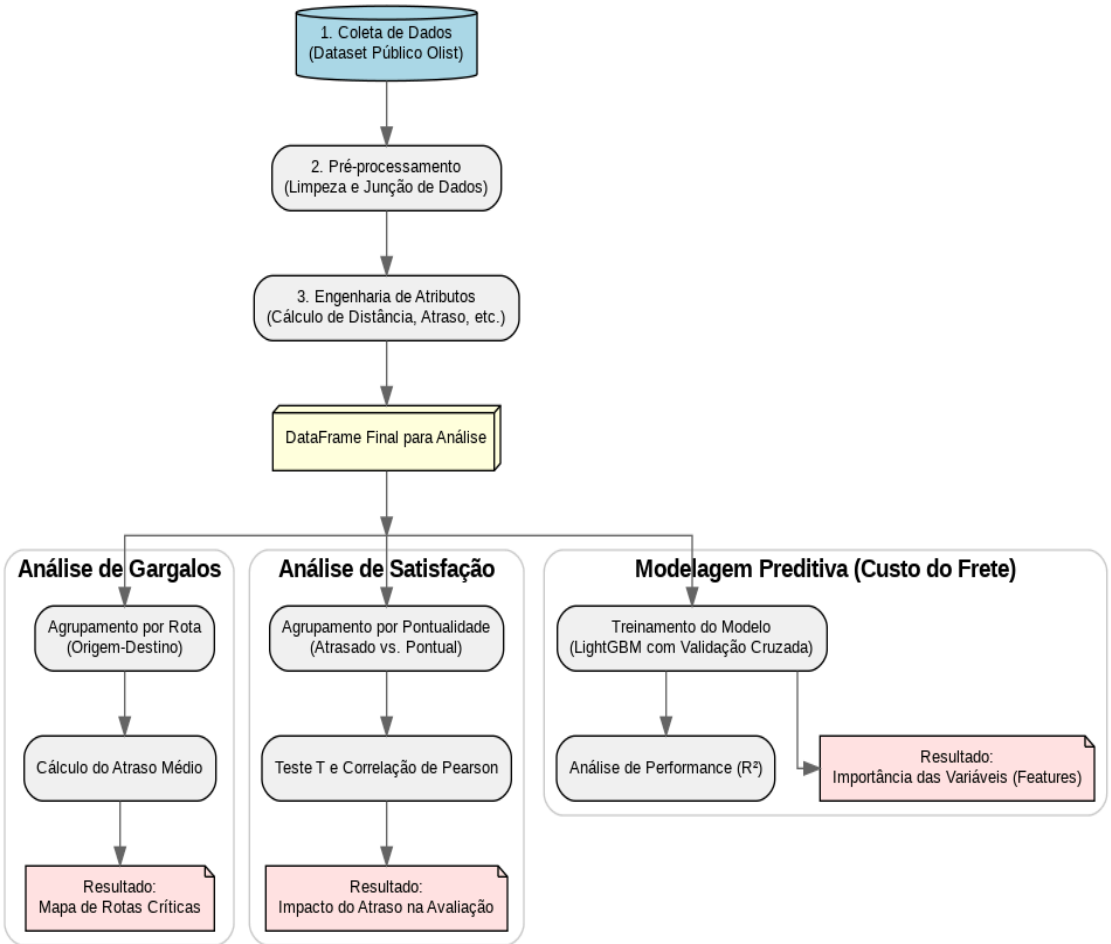

Figura 1: Fluxograma da metodologia de análise de dados.

A metodologia foi executada em três etapas sequenciais. Primeiro, na preparação de dados, os arquivos-fonte foram unidos, limpos (removendo-se registros com dados nulos essenciais) e submetidos à engenharia de atributos, onde variáveis como *distancia_km* (via fórmula de Haversine), *dias_atraso* e *is_interregional* foram criadas. Em seguida, na fase de análise, os dados foram agrupados por rota (origem-destino) para identificar gargalos logísticos, e a relação entre atrasos e a nota de avaliação (*review_score*) foi investigada com o Teste *t* de *Student* e a correlação de *Pearson*. Finalmente, na modelagem preditiva, um modelo LightGBM foi treinado para prever o *custo_frete*, utilizando validação cruzada (*k=10*) para aferir o R^2 e a importância das variáveis.

3.2. Detalhamento das Etapas

3.2.1. Pré-processamento de Dados

Os arquivos CSV, contendo os dados brutos, foram consolidados em um único *DataFrame* com o auxílio da biblioteca *Pandas*. Foi realizada uma etapa de limpeza, na qual registros com valores nulos em colunas essenciais, como datas de entrega, coordenadas geográficas e peso do produto, foram sistematicamente removidos. Esta ação foi importante para garantir a integridade e a precisão dos cálculos subsequentes. Após a limpeza, o *dataset* final consistiu em 107.859 registros únicos.

3.2.2. Engenharia de Atributos

Nesta fase, foram criadas variáveis analíticas chave. A distância em quilômetros (*distancia_km*) entre a origem e o destino de cada pedido foi calculada utilizando a fórmula de Haversine. Adicionalmente, os seguintes atributos foram desenvolvidos:

1. *dias_atraso*: Diferença, em dias, entre a data de entrega efetiva e a data estimada. Valores negativos indicam entrega adiantada.
2. *is_interregional*: Variável booleana que identifica se um pedido cruza fronteiras estaduais.

3.2.3. Análise de Gargalos

Para identificar gargalos logísticos, os pedidos foram agrupados por rota (definida pelo par de estado de origem e destino). Para cada rota, foram calculadas métricas de desempenho médio, como tempo de entrega, custo do frete e dias de atraso.

3.2.4. Análise de Satisfação do Cliente

A relação entre a pontualidade e a satisfação do cliente foi investigada com análises estatísticas. A diferença na nota média de avaliação (*review_score*) entre os grupos "pontuais" e "atrasados" foi avaliada com o Teste *t* de *Student* para amostras independentes, a fim de verificar sua significância estatística. Para quantificar a força da relação, foi calculado o coeficiente de correlação de *Pearson*.

3.2.5. Modelagem Preditiva

Para prever o valor do frete (*custo_frete*), foi treinado um modelo de regressão baseado no algoritmo LightGBM (*Light Gradient Boosting Machine*). A performance do modelo foi aferida utilizando a técnica de validação cruzada (*K-Fold* com *10 folds*) para garantir uma estimativa eficaz de sua capacidade de generalização.

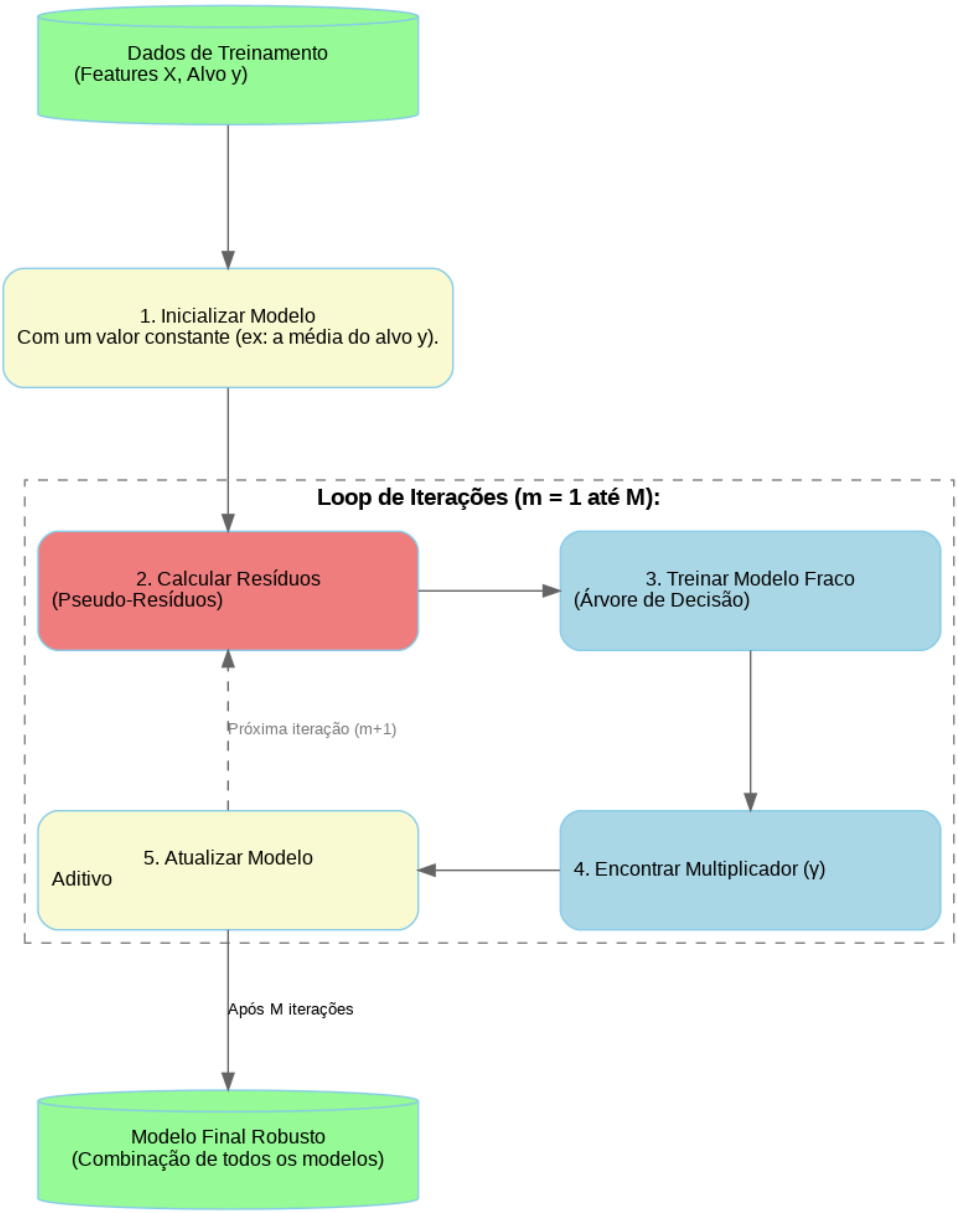

Figura 2: Fluxograma demonstrativo do LightGBM

4. Resultados

A análise dos dados revela a magnitude do desafio logístico brasileiro, com uma distância média de entrega de aproximadamente 600 km e alta variabilidade no tempo de espera (desvio padrão de 9,5 dias), conforme detalhado na Tabela 1. Para investigar as fontes dessa variabilidade, foram mapeados os gargalos em rotas com tráfego minimamente consistente (mais de 5 pedidos). A análise, visualizada na Figura 3,

confirmou a Hipótese 3, apontando claros pontos de estrangulamento em rotas inter-regionais. Destacam-se as rotas que envolvem a região Nordeste, como BA -> MA (atraso médio de 7,94 dias), MA -> SE (6,80 dias) e RJ -> CE (5,34 dias), que apresentam os maiores tempos médios de atraso.

Tabela 1: Estatísticas descritivas das variáveis de análise

	Média	Desvio Padrão	Mín.	Mediana	Máx.
custo_frete (R$)	19.96	15.55	0.0	16.28	409.68
tempo_entrega_dias	12.11	9.54	1.0	9.0	209.0
distancia_km	598.36	588.64	0.0	430.82	8725.21
peso_produto_kg	2.09	3.68	0.05	0.95	40.43

Além do tempo, o estudo investigou os determinantes do custo de frete. Para isso, foi desenvolvido um modelo preditivo LightGBM, cujo desempenho se mostrou muito bom, conforme detalhado na Tabela 2. O modelo obteve um coeficiente de determinação (R^2) de 0.6755, indicando que é capaz de explicar aproximadamente 67,5% da variância no custo do frete, com um erro médio absoluto de R$ 4,36 por entrega.

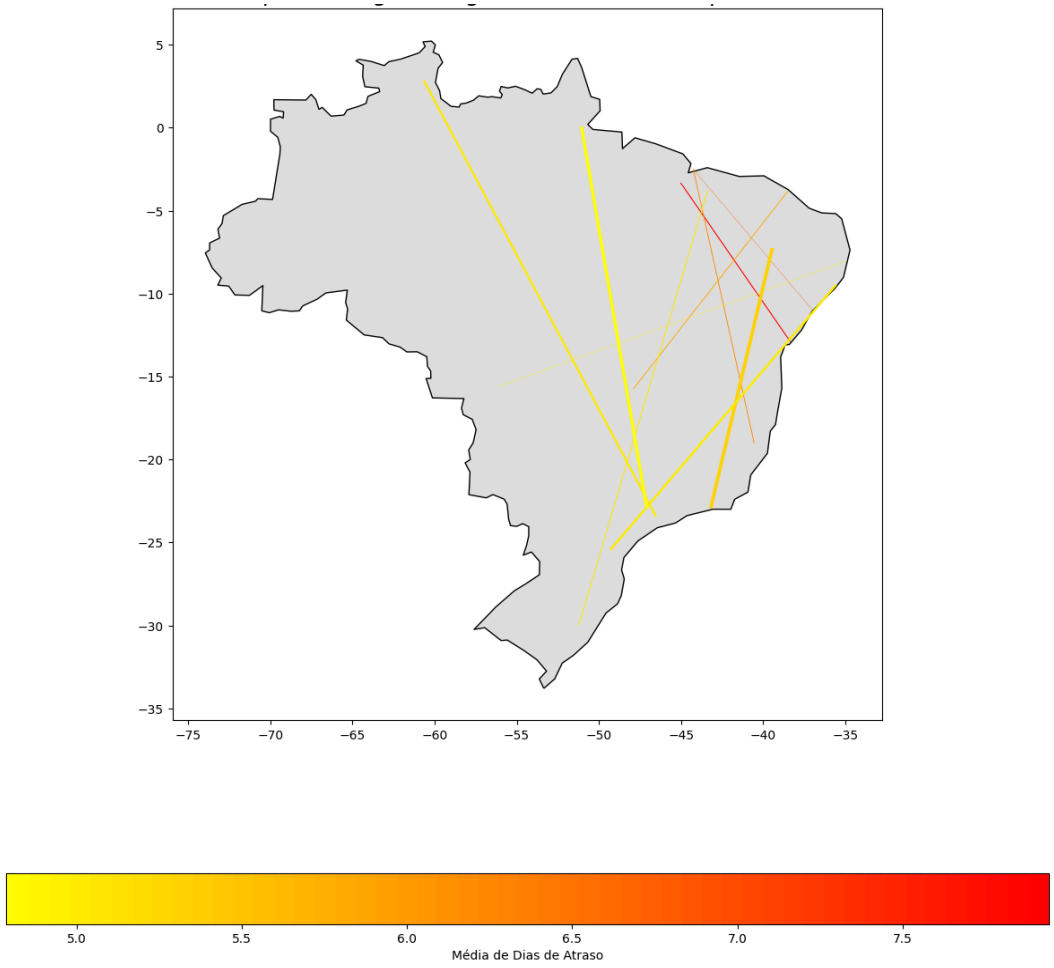

Figura 3: Mapa de Gargalos Logísticos: Piores Rotas por Atraso

A análise dos fatores do modelo, ilustrada na Figura 4, validou parcialmente a Hipótese 1. O peso do produto (*peso_produto_kg*) revelou-se a variável mais decisiva para determinar o custo, superando a distância (*distancia_km*), que figura como o segundo fator mais importante. Contrariando a intuição comum, o modelo também indicou que o

fato de uma localidade ser capital não influencia significativamente o preço, e o fator inter-regional teve apenas um impacto modesto no custo final.

Tabela 2: Métricas de desempenho do modelo de LightGBM

	Valor
R^2 (R-squared)	0.6755
Desvio Padrão do R^2	± 0.0186
MAE (Erro Absoluto Médio)	R$ 4.36
RMSE (Raiz do Erro Quadrático Médio)	R$ 8.95

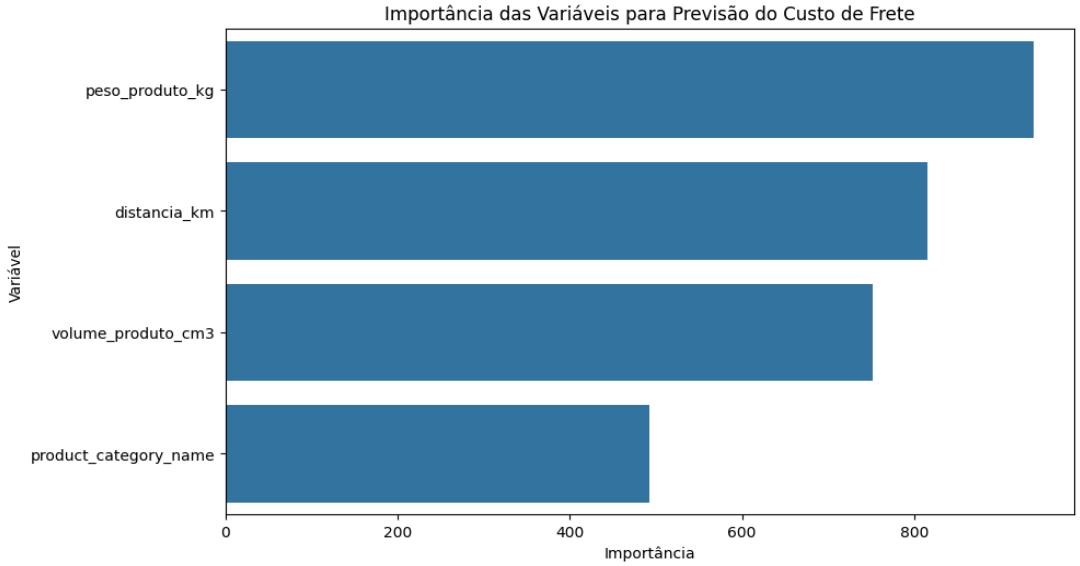

Figura 4: Importância das variáveis para previsão do custo de frete

5. Discussão e Conclusão

Esta pesquisa demonstrou, por meio de uma análise de dados em larga escala, os principais determinantes do desempenho logístico no *e-commerce* brasileiro. O mapeamento de gargalos críticos, como a rota entre Bahia e Maranhão, e a quantificação do impacto dos atrasos forneceram evidências empíricas de que falhas na pontualidade da entrega afetam negativamente a avaliação do consumidor. Um dos achados mais relevantes foi a constatação de que o peso do produto supera a distância como principal fator na composição do custo de frete, desafiando a percepção comum sobre a precificação logística.

Esses resultados geram implicações diretas para os agentes do setor. Para os vendedores, a dominância do peso no custo do frete sinaliza que a otimização de embalagens é uma estratégia de economia fundamental. Para as plataformas de *e-commerce*, a identificação de rotas-gargalo e a clara correlação entre atraso e insatisfação são insumos estratégicos para a renegociação de contratos com transportadoras e o planejamento da malha logística. No âmbito das políticas públicas, os dados oferecem um diagnóstico preciso que pode subsidiar decisões de investimento. A evidência de ineficiência crônica em rotas específicas, por exemplo, justifica a alocação de recursos para a melhoria da infraestrutura rodoviária, o estudo de incentivos para modais alternativos (como

cabotagem e ferrovias) ou o fomento a centros de distribuição em regiões estrategicamente carentes, visando reduzir as disparidades logísticas regionais.

O estudo possui limitações, como a ausência de variáveis sobre a transportadora específica, o valor do pedido ou o tipo de produto. Pesquisas futuras poderiam enriquecer a análise ao incorporar dados externos sobre a malha rodoviária e segurança pública. Adicionalmente, o desenvolvimento de modelos preditivos focados na probabilidade de atraso por rota ofereceria uma ferramenta de gestão de risco ainda mais poderosa para o setor.

Referências

Agatz, N. A., Fleischmann, M. and van Nunen, J. A. (2008) "E-fulfillment and Multi-channel Distribution: A Review", *European Journal of Operational Research*, 187(2), p. 339-356.

Associação Brasileira de Comércio Eletrônico (ABComm) (2024) "Relatório E-commerce no Brasil".

Confederação Nacional do Transporte (CNT) (2022) "Transporte & Desenvolvimento: Panorama do Transporte de Cargas", Brasília, DF: CNT.

Duarte, A. L. C. M., Macau, F., Silva, C. F. E. and Sanches, L. M. (2019) "Last Mile Delivery to the Bottom of the Pyramid in Brazilian Slums", *International Journal of Physical Distribution & Logistics Management*, 49(5), p. 473-491.

Esper, T. L., Jensen, T. D., Turnipseed, F. L. and Burton, S. (2003) "The Last Mile: An Examination of Effects of Online Retail Delivery Strategies on Consumers", *Journal of Business Logistics*, 24(2), p. 177-203.

Instituto de Logística e Supply Chain (ILOS) (2023) "Custos Logísticos no Brasil: Pesquisa Anual".

Kim, J., Park, J. and Jeong, D. (2023) "Prediction of Shipping Cost on Freight Brokerage Platform Using Machine Learning", *Sustainability*, 15(2), 1122.

NielsenIQ Ebit (2023) "Webshoppers, 47ª Edição".

Olist (2018) "Brazilian E-Commerce Public Dataset by Olist", Kaggle.

Parasuraman, A., Zeithaml, V. A. and Berry, L. L. (1988) "SERVQUAL: A Multiple-item Scale for Measuring Consumer Perceptions of Service Quality", *Journal of Retailing*, 64(1), p. 12-40.

Rokoss, A., Valmorbida, V. G., Pescador, G. V. and Pellerin, R. (2024) "Case Study on Delivery Time Determination Using a Machine Learning Approach in Small Batch Production Companies", *Journal of Intelligent Manufacturing*, 35, p. 3937–3958.

VanderPlas, J. (2016) "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly Media.

World Bank (2023) "Logistics Performance Index (LPI) 2023: Connecting to Compete", Washington, D.C.: The World Bank.