UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
CURSO DE ENGENHARIA DE TRANSPORTES E LOGÍSTICA

ALAN ARTHUR PRIEBE

MODELAGEM MATEMÁTICA DA PRECIFICAÇÃO DE FRETES POR MOTORISTAS AUTÔNOMOS ATRAVÉS DO USO DO ALTERYX DESIGNER

Joinville
2024

ALAN ARTHUR PRIEBE

MODELAGEM MATEMÁTICA DA PRECIFICAÇÃO DE FRETES POR MOTORISTAS AUTÔNOMOS ATRAVÉS DO USO DO ALTERYX DESIGNER

Trabalho apresentado como requisito para obtenção do título de bacharel em Engenharia de Transportes e Logística, no Centro Tecnológico de Joinville, da Universidade Federal de Santa Catarina.

Orientadora: Dra Simone Becker Lopes.

Joinville

2024

ALAN ARTHUR PRIEBE

MODELAGEM MATEMÁTICA DA PRECIFICAÇÃO DE FRETES POR MOTORISTAS AUTÔNOMOS ATRAVÉS DO USO DO ALTERYX DESIGNER

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de bacharel em Engenharia de Transportes e Logística, na Universidade Federal de Santa Catarina, Centro Tecnológico de Joinville.

Joinville (SC), 04 de julho de 2024.

Banca Examinadora:

_____

Profa. Dra. Simone Becker Lopes
Orientadora/Presidente
Universidade Federal de Santa Catarina

_____

Profa. Dra. Elisete Santos da Silva Zagheni
Membro(a)
Universidade Federal de Santa Catarina

_____

Eng. Murilo Collin
Membro
Urban Analytics and Complex Systems

_____

Eng. Vangunther Bohn Junior
Membro
Bialog Transportes e Logística

Dedico este trabalho à Sheila Priebe, *in memorian*.

AGRADECIMENTOS

Primeiramente a Deus pela vida.

Aos meus pais, Adelbert e Elisete por me proporcionarem a oportunidade de estudar e me formar engenheiro, pela educação e pelo caráter que tenho.

As minhas irmãs Sabrina e Simone, pela ajuda incondicional em momentos difíceis e felizes, tanto na vida acadêmica quanto fora dela, nunca duvidando da minha capacidade.

A minha irmã Sheila, mesmo não estando mais presente fisicamente, por estar ao meu lado em meus pensamentos em todas as minhas decisões, me guiando.

Aos meus amigos Everton e Mayra, pelos tantos momentos compartilhados durante a graduação.

RESUMO

A matriz de transportes no Brasil é predominantemente rodoviária, sendo o transporte rodoviário de cargas um meio barato e eficiente em comparação aos demais, com diversas empresas atuando no ramo, seja utilizando frota própria ou contratando transportadores autônomos de cargas, como os caminhoneiros autônomos. No entanto, atuar e gerar lucro estando entre duas pontas tem suas dificuldades, pois ao mesmo tempo que é preciso oferecer valores satisfatórios para o contratante do frete, é preciso também oferecer preços atrativos para a contratação dos caminhoneiros. De tal forma, precificar o serviço é uma dificuldade, dadas diversas características e fatores, como a variação de diesel e a extensão territorial em proporções continentais. Este estudo é baseado nos dados disponibilizados pela empresa Bialog, sendo uma startup logtech que atua na contratação de motoristas autônomos para fretes de embarcadores. Foram levantados dados de dois anos de fretes praticados pela transportadora. Os dados foram coletados e tratados para definir um conjunto de variáveis candidatas aos modelos preditivos testados. Foram calibrados e validados, através do Alteryx Designer, modelos de Regressão Linear, Spline e Floresta Randômica. Os modelos mostraram eficácia na previsão, com vantagem para o modelo de Floresta Randômica. Escolhido esse modelo, foi comparado às formas de precificação já praticadas pela empresa, sendo o método de tabelas de frete mínimos da ANTT e próprio método Bialog. O modelo de Floresta Randômica ainda se sobressaiu em nova comparação aos demais em 83,65% das observações, obtendo R^2 de 0,9602, α de 134,42, β igual a 0,9458 e RMSE de 393,0512, o que destaca que o objetivo desse trabalho foi alcançado.

Palavras-chave: modelos preditivos; machine learning; precificação de frete.

ABSTRACT

The transportation matrix in Brazil is predominantly road-based, with road freight transport being a cheap and efficient means compared to others, and various companies operating in the sector, either using their own fleets or hiring autonomous freight carriers, such as independent truck drivers. However, operating and generating profit between two ends comes with its challenges. While it's necessary to offer satisfactory rates to the freight contractors, it's also important to provide attractive prices for hiring truck drivers. Pricing the service is difficult due to various factors, such as diesel price fluctuations and the country's continental-sized territory. This study is based on data provided by Bialog, a logtech startup that hires autonomous drivers for shippers' freights. Two years of freight data from the transporter were collected and processed to define a set of candidate variables for the predictive models tested. Linear Regression, Spline, and Random Forest models were calibrated and validated using Alteryx Designer. The models demonstrated effectiveness in prediction, with the Random Forest model showing a clear advantage. This model was compared to the pricing methods already used by the company, namely the ANTT minimum freight rate tables and Bialog's own method. The model outperformed the others in 83.65% of observations, achieving an R^2 of 0.9602, α of 134.42, β of 0.9458, and RMSE of 393.0512, which demonstrates that the objective of this work has been achieved.

Keywords: predictive methods; machine learning; freight price.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANP – Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
ANTT – Agência Nacional de Transportes Terrestres
API – *Application Programming Interface*
CNT – Confederação Nacional dos Transportes
CTC – Cooperativas do Transporte Rodoviário de Cargas
CTe – Conhecimento de Transporte Eletrônico
DACTE – Documento Auxiliar de Conhecimento de Transporte Eletrônico
DAMDFE – Documento Auxiliar do Manifesto Eletrônico de Documentos Fiscais
ETC – Empresas de Transporte Rodoviário de Cargas
IBGE – Instituto Brasileiro de Geografia e Estatística
INPC – Índice Nacional de Preços ao Consumidor Amplo
LPC – Levantamento de Preços de Combustíveis
MDFe – Manifesto Eletrônico de Documentos Fiscais
NFe – Nota Fiscal Eletrônica
PNL – Plano Nacional de Logística
PNPM-TRC – Política Nacional de Pisos Mínimos do Transporte Rodoviário de Cargas
RNTRC – Registro Nacional de Transportadores Remunerados de Carga
SEFAZ – Secretaria de Estado da Fazenda
SIG – Sistemas de Informações Geográficas
SQL – *Structured Query Language*
TAC – Transportadores Autônomos de Cargas
TMS – *Transportation Management System*
UF – Unidade Federativa

SUMÁRIO

1. INTRODUÇÃO

O Brasil tem elevados índices de participação do modo rodoviário na sua matriz de transportes, atingindo a marca de 65% de sua composição através desse em 2015, segundo dados do Plano Nacional de Logística (PNL, 2015). Além disso, operam no Brasil 300,5 mil empresas de transporte de cargas e mais de 938,7 mil transportadores autônomos de carga, segundo a Agência Nacional de Transportes Terrestres (ANTT, 2024).

Empresas do ramo de logística inseridas no transporte de cargas buscam atender o mercado ofertando o menor preço possível às empresas que fazem a contratação de um frete, ao mesmo tempo que buscam atingir os valores solicitados por transportadores autônomos e empresas donas de veículos. Desta forma, encontrar um equilíbrio entre as demandas, garante o lucro e bom funcionamento do seu negócio. A principal dificuldade em lidar com a contratação de caminhoneiros autônomos, é conseguir ofertar valores que sejam satisfatórios tanto ao transportador quanto ao embarcador.

A ANTT (2024) dispõe por meio de decreto, tabelas mínimas de frete, no entanto, as empresas possuem diferentes formas de precificar o serviço, seja por experiência de mercado, subjetividade ou percentuais em valores de carga. Segundo a Confederação Nacional dos Transportes (CNT, 2019) em pesquisa realizada com caminhoneiros, os principais motivos de insatisfação com as tabelas de frete mínimo disponibilizados pela ANTT são o não cumprimento da mesma pelas empresas contratantes, valores mínimos de frete baixos e custos elevados de operação.

Para a resolução de problemas de previsão de custos, a Machine Learning pode oferecer uma alternativa eficiente quando o custo e o tempo de desenvolvimento são as principais preocupações, ou quando o problema parece ser muito complexo para ser estudado em sua totalidade (SIMEONE, 2018).

O presente estudo busca entender os fatores determinantes para a formulação do preço de frete, estimando um modelo matemático capaz de prever os valores, considerando o peso de cada aspecto. Para isso, será realizado um estudo de caso com a Bialog, empresa que atua no transporte de cargas.

A Bialog Transporte e Logística é uma startup *logtech* fundada em 2019 e sediada em São Paulo, além de possuir outras duas filiais situadas nas cidades de

Aracaju e Joinville. A empresa tem a missão de conectar empresas embarcadoras aos caminhoneiros de forma rápida, segura e econômica. Para isso, atua na contratação de fretes de carga, sendo através do meio digital um elo entre empresas que necessitam do serviço de transporte e motoristas que o ofertam. Para atender satisfatoriamente os clientes contratantes e caminhoneiros, a empresa precisa conseguir precificar de forma acurada os valores de frete, porém os métodos que já possui não atingem a eficácia desejada.

Por meio deste estudo, espera-se desenvolver um modelo de previsão de preços de fretes que seja mais preciso, proporcionando satisfação tanto para motoristas quanto para empresas embarcadoras, e garantindo a lucratividade da Bialog.

1.1. OBJETIVOS

Para o entendimento das variáveis explicativas na formulação do preço de fretes por caminhoneiros autônomos, propõe-se os seguintes objetivos.

1.1.1. Objetivo Geral

Estimar um modelo acurado de previsão de valores de fretes através do uso de Alteryx Designer, garantindo maior assertividade nas contratações.

1.1.2. Objetivos Específicos

- Tratar os dados de entrada para os modelos de previsão de preços de frete selecionando as variáveis candidatas;
- Calibrar diferentes modelos de previsão de preços de fretes;
- Validar os modelos de previsão de preços de fretes e selecionar o melhor;
- Comparar os resultados obtidos através do melhor modelo de previsão validado com os modelos já praticados pela empresa.

2. FUNDAMENTAÇÃO

Para que se faça completo o entendimento sobre o estudo, faz-se necessário a revisão sobre a fundamentação teórica, com a finalidade de apresentar conceitos sobre métodos preditivos, documentação de frete e os agentes atuantes no transporte rodoviário de cargas e suas ferramentas.

2.1. O TRANSPORTE RODOVIÁRIO DE CARGAS NO BRASIL

O transporte de cargas é a principal forma de transporte de bens e mercadorias no país. A predominância do modo, se deve a diversos fatores históricos e geográficos. Uma das principais causas do uso elevado, se iniciou na década de 50 com a construção de rodovias a partir da política de incentivo à indústria automobilística.

O baixo custo de criação de infraestrutura para esse modo é o diferencial para sua popularidade. O Brasil possui segundo a ANTT (2024), cerca 1,7 milhão de quilômetros de estradas, que interligam regiões e estados. No entanto, boa parte das estradas é considerada ruim ou sem manutenção, acarretando custos financeiros, além da falta de segurança. Assim como os altos índices de acidentes, os roubos de cargas são um problema constante e que refletem negativamente no mercado.

Ainda assim, o transporte de cargas rodoviário é o que movimenta o mercado logístico brasileiro, permitindo a evolução e prosperidade da sociedade.

2.1.1 Agência Nacional de Transportes Terrestres

A ANTT é uma autarquia federal que faz parte do Ministério da Infraestrutura, ou seja, é uma entidade administrativa criada pela Lei 10.233 de 5 de junho de 2001 que é fundamental para a implantação de políticas públicas e prestação de serviços relacionados ao transporte terrestre à sociedade.

O órgão tem papel fundamental na sociedade, pois tem como missão a regulação e fiscalização dos transportes terrestres, sendo assim contribui para o desenvolvimento nacional garantindo serviços e infraestruturas de qualidade (ANTT, 2024).

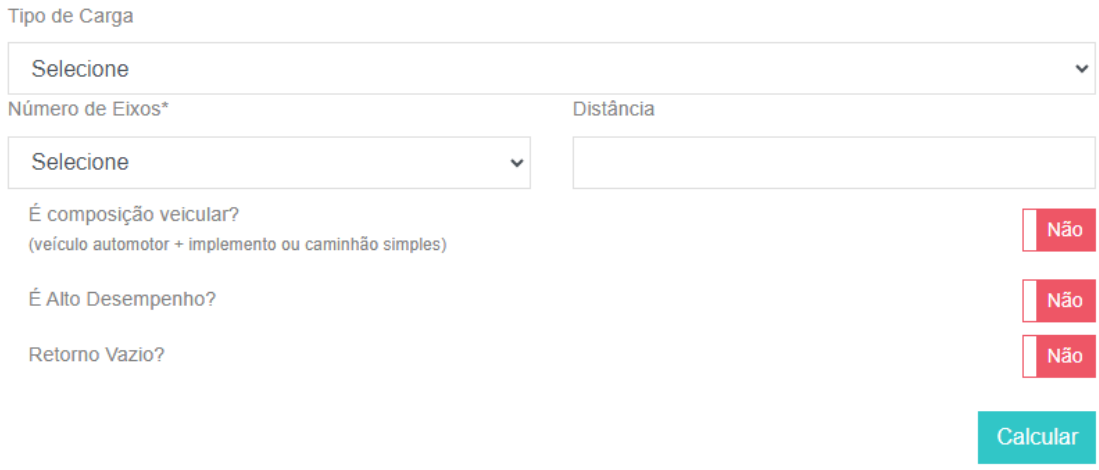
Possui caráter de regulação e fiscalização, pois define normas que ditam a prestação dos serviços de transporte rodoviário e ferroviário, além de fiscalizar a operação de vias concedidas. Faz ainda parte do seu escopo, permitir concessões nos serviços, regulando contratos e investimentos na área.

Um instrumento utilizado pela ANTT é o Registro Nacional de Transportadores Remunerados de Carga (RNTRC), que serve para cadastrar Transportadores Autônomos de Cargas (TAC), Empresas de Transporte Rodoviário de Cargas (ETC) e Cooperativas do Transporte Rodoviário de Cargas (CTC). Todo atuante do processo do transporte precisa ter seu vínculo ativo e regularizado. Em levantamento realizado em 2024, no Brasil existe 300,5 mil ETCs, 938,7 mil TACs e mais de 2,7 milhões de veículos de cargas.

Para garantir a sustentabilidade financeira do setor rodoviário, foi instituída a Política Nacional de Pisos Mínimos do Transporte Rodoviário de Cargas (PNPM-TRC) através da Lei nº 13.703 de 8 de agosto de 2018. A política é o principal instrumento de regulação de valores de frete. Segundo a ANTT (2023), “foi criada pelo Governo Federal em resposta à manifestação dos caminhoneiros, ocorrida em maio de 2018.”

A política busca manter condições mínimas para a realização de fretes, estipulando valores mínimos por quilometragem rodada e tipo de carroceria e carga carregada. Como ferramenta, o órgão disponibiliza em seu site uma calculadora de piso mínimo de fretes. A ferramenta se baseia na tabela de fretes atual do órgão, sendo assim, utiliza-se do preenchimento de características do frete para a previsão de um valor mínimo (Figura 1).

Figura 1 - Calculadora de fretes ANTT



Fonte: ANTT (2024).

O preenchimento consiste na identificação do tipo de carga, sendo possível a especificação conforme trezes tipos de cargas, sendo o mais comum a carga lotação. O tipo de veículo determina a quantidade de eixos dele, através dessa quantidade há um parâmetro que determina o valor por quilômetro rodado por tipo de veículo e um valor fixo de descarga. Por fim, a quantidade de quilômetros estimada da rota deve ser preenchida. Os parâmetros que determinam o cálculo do valor final são dispostos através de resoluções e adaptados através da Figura 2. Cada tabela traz uma especificidade quanto ao perfil contratado.

Figura 2 - Tabela de coeficientes disposta na resolução

#	Tipo de carga	Coeficiente de custo	Unidade	Número de eixos carregados do veículo combinado						
				2	3	4	5	6	7	9
1	Granel sólido	Deslocamento (CCD)	R$/km	3,433	4,382	5,06	5,52	6,148	6,9381	7,8196
		Carga e descarga (CC)	R$	397,3	481,9	521,5	513,4	540,1	690,73	735,06
2	Granel líquido	Deslocamento (CCD)	R$/km	3,506	4,481	5,041	5,666	6,378	7,0775	8,1032
		Carga e descarga (CC)	R$	411,9	503,5	505,8	543,1	592,8	718,5	802,46
3	Frigorificada ou Aquecida	Deslocamento (CCD)	R$/km	4,134	5,25	6,047	6,718	7,413	8,7727	9,6599
		Carga e descarga (CC)	R$	476,9	573,2	622,5	647,9	665	972,25	983,92
4	Conteinerizada	Deslocamento (CCD)	R$/km		4,356	4,937	5,462	6,075	6,9249	7,7762
		Carga e descarga (CC)	R$		474,7	487,6	497,7	520,2	687,11	723,12
5	Carga Geral	Deslocamento (CCD)	R$/km	3,41	4,368	5,006	5,491	6,068	6,945	7,8685
		Carga e descarga (CC)	R$	391	478	506,5	505,6	518,4	692,62	748,51
6	Neogranel	Deslocamento (CCD)	R$/km	3,078	4,368	5,018	5,491	6,068	6,945	7,8685
		Carga e descarga (CC)	R$	391	478	509,8	505,6	518,4	692,62	748,51
7	Perigosa (granel sólido)	Deslocamento (CCD)	R$/km	4,091	5,04	5,757	6,217	6,845	7,6497	8,5379
		Carga e descarga (CC)	R$	523,2	607,9	655,5	647,5	674,2	828,67	874,83
8	Perigosa (granel líquido)	Deslocamento (CCD)	R$/km	4,184	5,158	5,74	6,365	7,077	7,7907	8,823
		Carga e descarga (CC)	R$	548,9	640,4	650,8	688,2	737,8	867,42	953,22
9	Perigosa (frigorificada ou aquecida)	Deslocamento (CCD)	R$/km	4,649	5,766	6,596	7,267	7,962	9,3406	10,2364
		Carga e descarga (CC)	R$	568,2	664,5	724,3	749,7	766,8	1.079,10	1.093,15
10	Perigosa (conteinerizada)	Deslocamento (CCD)	R$/km		4,663	5,284	5,809	6,422	7,2863	8,1442
		Carga e descarga (CC)	R$		556,2	577,2	587,2	609,8	780,58	818,41
11	Perigosa (carga geral)	Deslocamento (CCD)	R$/km	3,718	4,675	5,353	5,838	6,416	7,3063	8,2366
		Carga e descarga (CC)	R$	472,5	559,5	596,1	595,1	607,9	786,09	843,8
12	Carga Granel Pressurizada	Deslocamento (CCD)	R$/km				5,874	6,574		8,4113
		Carga e descarga (CC)	R$				610,9	657,4		897,78

Fonte: Adaptado de ANTT (2024).

A Equação 1 representa o cálculo do valor mínimo de frete. Os parâmetros são encontrados na tabela da Figura 2 e substituídos na equação.

$$Valor\ mínimo = (KM * CCD) + CC \quad (1)$$

Onde KM corresponde à quantidade de quilômetros totais da rota, CCD representa o fator de deslocamento encontrado na tabela da Figura 2 e CC o valor de descarga, ambos a partir da quantidade de eixos.

2.1.2 Documentação de frete

A documentação para o transporte rodoviário de cargas é fundamental para assegurar a legalidade, segurança e eficiência das operações logísticas. Os principais documentos tratados neste estudo são o de conhecimento de transporte e o manifesto de documentos fiscais.

2.1.2.1 Conhecimento de transporte

O documento de conhecimento de transporte é o documento principal na contratação de um serviço de transporte de carga através do modo rodoviário. O documento é definido por Ballou (2006, p. 182) como “um contrato legal entre o embarcador e o transportador para a movimentação de carga com razoável rapidez até um destino especificado, e com entrega sem danos ou perdas”.

Nele, estarão especificados os envolvidos na negociação, definindo quem é o tomador do serviço, quem são o remetente e o destinatário, além de questões fiscais de recolhimento de impostos nacionais. As notas fiscais dos produtos estarão vinculadas a este documento.

O conhecimento de transporte é um documento, portanto, necessita ser autenticado e de tal maneira, é validado digitalmente através de chaves junto a Secretaria do Estado da Fazenda (SEFAZ), órgão responsável pelo controle de receitas e despesas dos estados. Atualmente é chamado de Conhecimento de Transporte Eletrônico (CTe) e normalmente utilizado através do Documento Auxiliar de Conhecimento de Transporte Eletrônico (DACTe) que nada mais é que uma representação visual simplificada.

2.1.2.1 Manifesto de documentos fiscais

O Manifesto Eletrônico de Documentos Fiscais (MDFe) é um documento obrigatório no transporte de cargas intermunicipal ou interestadual, pois é através dele que são identificadas as cargas que estão em trânsito. De maneira simplificada, o MDFe agrupa os CTEs referentes a determinada carga.

No documento fiscal, estão contidas as informações das chaves dos CTEs que foram autorizados junto a SEFAZ, além das informações de quem é o agente que transporta a carga, identificando o motorista, veículo e registro junto a ANTT.

O MDFe também pode ser conhecido através da nomenclatura Documento Auxiliar do Manifesto Eletrônico de Documentos Fiscais (DAMDFe), que é a representação simplificada das informações transmitidas à SEFAZ.

2.2. COLETA, ORGANIZAÇÃO E ANÁLISE DE DADOS

A primeira etapa para a criação de um modelo preditivo é a compreensão sobre os dados obtidos.

A hipótese inicial da criação de um modelo preditivo é de que os dados obtidos contenham indícios das causas ou possuam padrões que ajudam o modelo a ser criado. Sendo assim, os grupos de dados têm de ser capazes de explicar o fenômeno que se busca prever.

As informações coletadas nunca serão perfeitas, existem sempre erros contidos e que acontecem de forma inesperada e até mesmo criativa (CHAI, 2020). A coleta dos dados tem suma importância, porém Witten, Frank e Hall (2005) definem a etapa de tratamento de dados como a parte de maior esforço investido dentro do processo de modelagem dos dados.

Witten, Frank e Hall (2005) apresentam os desafios na coleta de dados, exemplificando que diferentes departamentos ou fontes terão suas próprias formas de registrar dados, em períodos diferentes, com convenções diferentes, com diferentes chaves ou identificadores e que isto resultará em inúmeros erros, sendo assim os dados precisam ser tratados e limpos.

Para Polyzotis et al. 2018, as ferramentas de análise e visualização podem ajudar na visualização e compreensão dos dados, desvendando propriedades surpreendentes.

2.2.1 A ferramenta Alteryx Designer

No mercado existem diversas ferramentas que auxiliam na criação de análise estatística e visualização ou tratamento de dados. Nesta seção serão apresentadas algumas ferramentas e pacotes, além dos principais pontos sobre o Alteryx Designer.

Quando se trata da manipulação de uma base de dados, o Microsoft Excel é a ferramenta líder. O software é um editor de planilhas robusto, capaz de desenvolver cálculos e tratamentos com dados, apresentando-os de forma clara, direta e fácil. Os arquivos gerados em Excel são amplamente utilizados como forma de entrada ou saída para outras ferramentas, porém a análise de dados em grande volume pode ser uma de suas vulnerabilidades.

Semelhantemente, para análises estatísticas existe a linguagem R, que é um conjunto integrado de recursos de software para manipulação de dados, cálculo e exibição gráfica, com maior potencial de exploração. A linguagem é bem desenvolvida, simples e eficaz que inclui condicionais, loops, funções recursivas definidas pelo usuário e recursos de entrada e saída, fornecendo diversas técnicas estatísticas, como modelagem linear e não linear, testes estatísticos clássicos, entre outros (Fundação R, 2024).

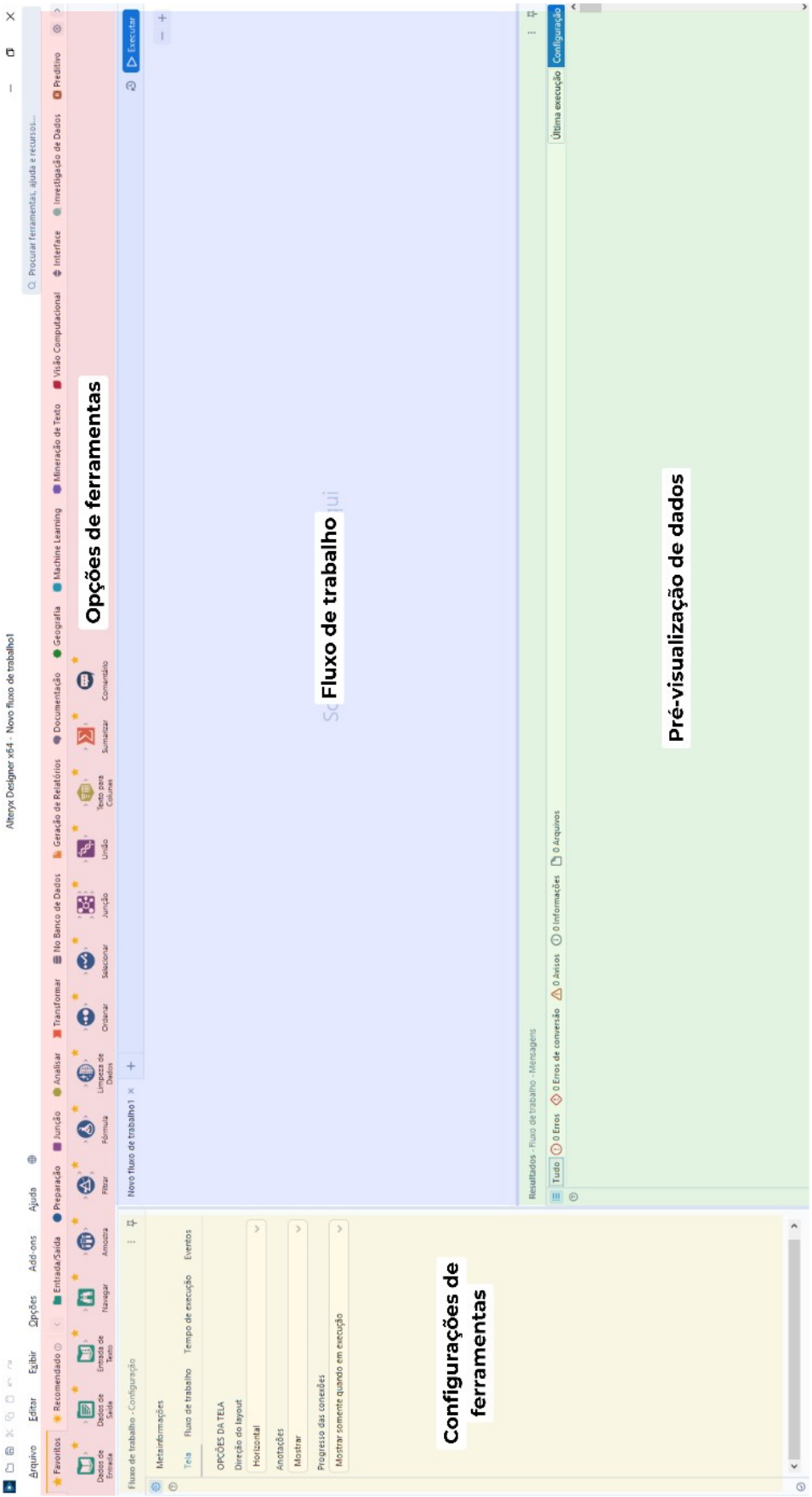
O ambiente R está presente em muitas ferramentas, especialmente no Alteryx Designer. O software é voltado aos analistas de dados, emponderando-os com um programa de preparação, união e análise de dados – preditiva, estatística e espacial -, em uma interface simples (Alteryx Inc., 2024).

O Alteryx é baseado no estilo “arrastar e soltar”, onde dentro de um fluxo de trabalho, uma ferramenta é arrastada e inserida com algumas predefinições básicas. Cada fluxo de trabalho possui uma sequência lógica, onde o usuário insere uma entrada de dados, seja por integração externa ou inserindo diretamente, conecta às ferramentas de preparação, tratamento e modelagem e no fim visualiza e exporta suas análises. Essa abordagem facilita a visualização e a compreensão das transformações e análises aplicadas aos dados, tornando a exploração mais acessível para os usuários (Alteryx Inc., 2024).

A interface pode ser subdividida entre quatro áreas com finalidades diferentes. Na Figura 3, tem-se na parte superior em vermelho as opções de ferramentas, que são categorizadas em grupos ou salvas em favoritos. Essas ferramentas têm finalidades específicas, que podem ser configuradas através da barra lateral que auxilia na configuração de parâmetros, com destaque amarelo.

As ferramentas são utilizadas uma vez que arrastadas para a área em destaque azul, o fluxo de trabalho. É no fluxo de trabalho onde serão esquematizados os processos envolvidos desde a entrada até a saída de dados. Pós-tratamento, os dados estarão disponíveis para pré-visualização na área em destaque amarelo.

Figura 3 – Interface Alteryx Designer



Fonte: Autor (2024).

O Quadro 1 a seguir relaciona as ferramentas presentes na interface do software com sua respectiva descrição.

Quadro 1 - Ferramentas do Alteryx Designer

Nome da ferramenta	Descrição
Visão Computacional	Use o Designer para inserir, processar e analisar imagens e, depois, gerar a saída delas.
Conectar	Ferramentas para o Alteryx Connect.
Mineração de Texto	Analise os dados do texto.
Ferramentas de entrada/saída	Forneça entradas e saídas para os fluxos de trabalho.
Machine Learning	Crie modelos de aprendizado de máquina.
Preparação	Prepare os dados para análise posterior.
Junção	Combine dois ou mais fluxos de dados, agregando os dados a um esquema largo/longo.
Analisar	Separe os valores de dados em um esquema de tabela padrão.
Transformar	Sumarize ou reorganize os dados.
No Banco de Dados	Conecte uma base de dados para combinar e visualizar os dados.
Geração de Relatórios	Ajuda para a apresentação e organização dos dados.
Documentação	Melhore a apresentação de um fluxo de trabalho adicionando anotações e organizando ferramentas.
Geografia	Manipulação e processamento espacial de dados e edição de objetos geográficos.
Ferramentas de interface	Projete elementos de interface de usuário para aplicativos e macros.
Investigação de Dados	Entenda os dados a serem usados em um projeto de análise preditiva.
Preditivo	Modelagem preditiva, comparação de modelos e ferramentas para teste de hipóteses.
Teste AB	Realize experimentos de teste A/B.
Série Temporal	Plotagem de gráficos de séries temporais regulares e univariadas e ferramentas de previsão.
Agrupamento Preditivo	Agrupe registros ou campos em um número menor de grupos.
Prescritivo	Determine a melhor estratégia ou resultado para diferentes situações.
Conectores	Recupere ou envie dados para a nuvem ou Internet.
Endereço	Padronize listas de endereços e realize a geocodificação para o nível de código postal com 9 dígitos.
Análise Demográfica	Extraia dados utilizando o mecanismo do Allocate dentro do Alteryx.
Análise comportamental (obsoleto)	Extraia dados utilizando o mecanismo do Solocast.

Calgary	Recupere dados de contagem listados e realize análises em bases de dados de grande escala.
Desenvolvedor	Crie macros e aplicativos analíticos e execute programas externos.
Laboratório	Ferramentas que não são para uso em produção.
Exemplos de SDK	Exemplos de ferramentas de SDK.

Fonte: Adaptado de SILVEIRA (2023).

O presente trabalho faz uso de parte das ferramentas disponibilizadas. Foram utilizadas “Ferramentas de entrada/saída” para a integração dos dados com as planilhas do Microsoft Excel, “Preparação”, “Junção”, “Analisar” e “Transformar” para o tratamento de dados e para a organização melhor do fluxo, foi utilizada a ferramenta “Documentação”. Os modelos foram criados utilizando-se da aba “Preditivo”, onde estão disponibilizados os algoritmos e testes utilizados.

2.3. MÉTODOS DE PREDIÇÃO

Os métodos de predição são algoritmos e ferramentas capazes de oferecer de forma satisfatória uma previsão de um fenômeno.

Eles são calibrados e validados a partir de dados de série histórica, com variáveis que contém informações relevantes sobre o fenômeno a ser explicado. Através da identificação de padrões e correlações presentes nas variáveis, os métodos vão convergindo em previsões aproximadas.

2.3.1 Regressão linear

Segundo Devore (2006, p. 433), “A análise de regressão é a parte da estatística que investiga a relação entre duas ou mais variáveis relacionadas de maneira não-determinística”.

Os métodos de regressão linear simples (RLS) e regressão linear múltipla (RLM) são procedimentos estatísticos que procuram entender e quantificar os fatores que influenciam um fenômeno. Estes modelos podem relacionar duas ou mais variáveis quantitativas ou qualitativas.

As regressões lineares se baseiam na construção de funções lineares como representadas na Equação 3.

$$Y = X\beta + \alpha + \varepsilon \quad (2)$$

Onde:

- Y – Variável dependente
- X – Variáveis independentes
- β – Constantes e coeficientes estimados
- α – Intercepto
- ε – Erro aleatório

Desta forma, Y é a variável ou evento que busca ser explicado e que pode ser chamada também de variável resposta. As demais variáveis que influenciam na explicada, podem ser chamadas de variáveis independentes ou explicativas, representadas pelos termos X (Devore, 2006).

Para estudar um fenômeno e explicá-lo através de uma RLM, torna-se necessário entender como as variáveis independentes se relacionam à variável dependente e como elas também se relacionam entre si. Para estimar tal relação, o cálculo do coeficiente de correlação de Pearson é aplicado através da fórmula:

$$p_{x_i y} = \frac{Covariância_{x_i y}}{\sqrt{Variância_{x_i}} \cdot \sqrt{Variância_y}} = \frac{\sum_k^{\square}(x_{ik} - \bar{x}) \cdot (y_k - \bar{y})}{\sqrt{\sum_k^{\square}(x_{ik} - \bar{x})^2} \cdot \sqrt{\sum_k^{\square}(y_k - \bar{y})^2}} \quad (3)$$

Onde:

$$-1 \leq p_{xy} \leq 1 \quad (4)$$

Duas variáveis podem estar positivamente ou negativamente relacionadas. Quando há o crescimento de uma ao mesmo tempo em que há crescimento da segunda, diz-se que estas estão positivamente correlacionadas. De forma análoga, duas variáveis são ditas negativamente correlacionadas quando há o crescimento de uma delas em detrimento do decrescimento da segunda.

Variáveis independentes altamente correlacionadas, não devem ser adicionadas simultaneamente a um modelo de regressão, pois não têm significância estatística.

2.3.2 Floresta Aleatória

O Método de Floresta Aleatória ou Random Forest é um algoritmo de *machine learning* que tem amplo uso e aplicações. Através do algoritmo são combinadas as saídas de múltiplas árvores de decisão, alcançando um único objetivo. O método é capaz de encontrar soluções eficientes para problemas ditos como classificação ou regressão. Criador do método, Breiman (2001), define que as florestas aleatórias são um conjunto com várias árvores de decisão, onde cada árvore é baseada em um vetor aleatório com amostra independente. A combinação de cada uma dessas árvores é o que possibilita a previsão em de um resultado.

Biau (2016) cita que o Método de Floresta Aleatória é um dos modelos mais utilizados atualmente na área de pesquisa e aplicação de machine learning, pois se adapta a uma grande quantidade de problemas, com a capacidade de resolvê-los com poucos hiperparâmetros. Três parâmetros são necessários no método: quantidade de variáveis aleatórias em cada árvore, o tamanho do nó de cada árvore e o parâmetro *bootstrapping,* que faz com que amostras únicas sejam escolhidas para cada árvore.

Como todo método, possui vantagens e desvantagens. O método tem risco reduzido de *overfitting,* que se manifesta quando um modelo se adapta muito bem aos dados de treinamento, mas não consegue fazer boas previsões quando exposto a um novo conjunto de dados. O modelo também tem a facilidade na definição das variáveis que contribuem para a previsão.

Os pontos negativos do método estão concentrados no tempo e recursos para execução. Por se tratar de um método com grande entrada de dados, a limpeza e armazenamento desses dados pode apresentar um problema. O alto custo computacional pode também ser um ponto negativo, visto que o modelo precisa de muito processamento.

2.3.3 Spline

Cunha (2003) define as splines como métodos com boas propriedades de aproximação, convergência e estabilidade, que foram desenvolvidas a partir de necessidades práticas de aproximação.

As splines são definidas por Gomes et al. (2017, p. 223) como “modelos de regressão com um conjunto de variáveis fictícias e que se ajusta linhas de regressão

separadas dentro das regiões entre os nós, e os nós ligam os ajustes de regressão segmentada.”

Assim, a previsão através de métodos spline é uma técnica que visa ajustar os valores previstos conforme uma curva suave que se modela aos dados.

Por se tratar de um modelo de interpolação, o método spline necessita de uma quantidade de dados considerável para que os resultados previstos sejam satisfatórios, pois quantos mais pontos inseridos, mais suave será a curva.

2.4. MÉTRICAS DE DESEMPENHO E VALIDAÇÃO DE MODELOS

Para avaliar-se os resultados dos métodos obtidos e suas predições, são necessárias algumas métricas de desempenho. Essas, utilizam-se de fórmulas para quantificar a eficácia dos modelos. A seguir serão apresentadas as principais métricas utilizadas para a avaliação de resultados e validação de modelos

2.4.1 Erro relativo

Para o comparativo entre métodos, necessita-se de um método de avaliação dos erros observados em relação aos valores verdadeiros. Para tal, utiliza-se a medida de erro relativo, que é uma forma de expressar a magnitude do erro em relação ao tamanho do valor verdadeiro. Ele pode ser calculado através da fórmula disposta através da Equação 4.

$$Er = \frac{|\hat{y} - y|}{|y|} \times 100\% \quad (5)$$

Onde:

- $\hat{y}$ – Valor estimado
- y – Valor observado

2.4.2 Raiz Quadrada do Erro Quadrático Médio - RMSE

A RMSE ou em português, raiz quadrada do erro quadrático médio, é uma métrica que mede a precisão geral do modelo. Bruce e Bruce (2019, p. 185) a definem

como a métrica de desempenho mais importante da perspectiva da ciência de dados. Através da fórmula representada na Equação 5, são calculados os erros com base no valor previsto e o valor real.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

Onde:

- $\hat{y}$ – Valor estimado
- y – Valor observado

De forma direta, o valor encontrado na métrica pode ser explicado como quanto o modelo pode estar errando, tanto para menos quanto para mais.

2.4.3 Coeficiente de Determinação - R²

O Coeficiente de Determinação R² é uma medida muito utilizada juntamente aos modelos de regressão, frequentemente aplicada para julgar a adequação de um modelo de regressão (MONTGOMERY, 2009). O coeficiente varia entre os valores de 0 e 1, medindo a proporção da variabilidade total da variável dependente. Quanto mais próximo de 1, as variáveis estão mais ajustadas ao modelo, isto é, percentualmente o valor indica a proporção de Y explicada pela presença da variável X.

O coeficiente pode ser determinado através do uso da Equação 6:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

Onde:

- $\hat{y}$ – Valor estimado
- y – Valor observado
- n – Número total de observações

2.4.4 Validação dos modelos

Para a comparação de desempenho, se faz necessária a utilização de indicadores iguais para todos os métodos analisados, assim são definidos α, β, R^2 e o *Root Mean Squared Error*. Lopes (2010) define como modelo bem ajustado, os que apresentam, na Regressão Simples entre valor observado e valor estimado, o intercepto α mais próximo de 0, com a inclinação β o mais próximo de 1 e R^2 mais próximo de 1. Para a avaliação do RMSE, temos que os menores valores possíveis indicam a menor presença de erros.

3. METODOLOGIA

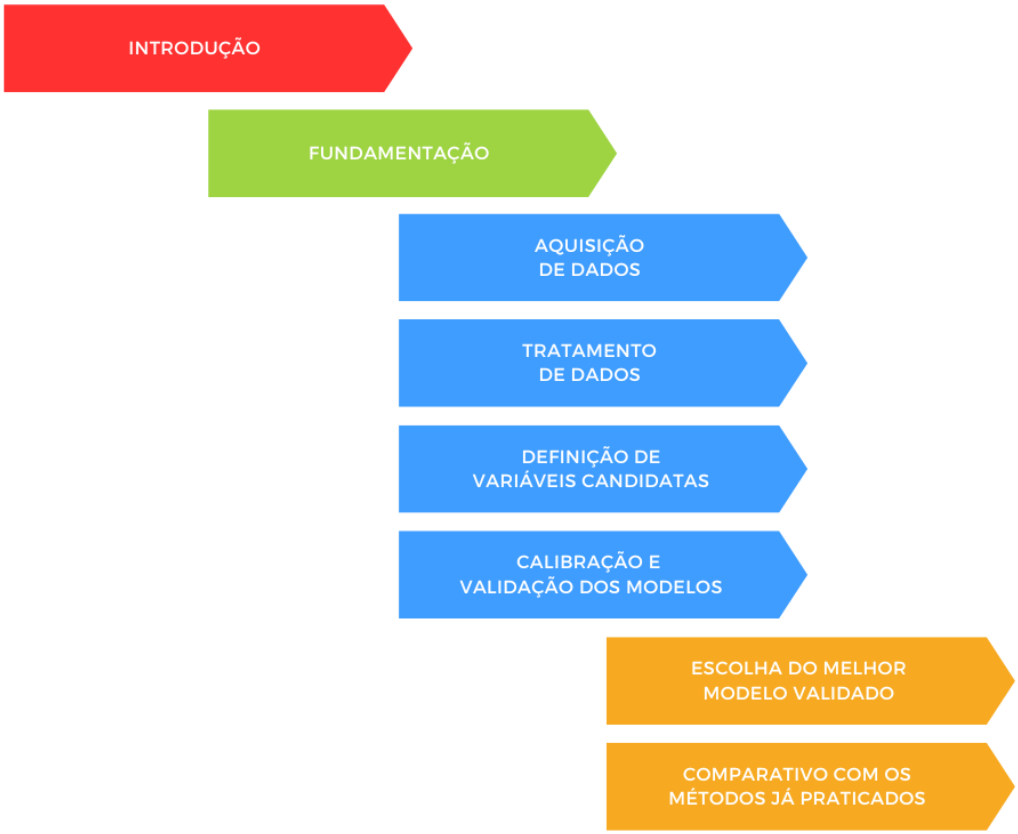
O presente estudo tem como objetivo desenvolver um modelo acurado através do uso de Alteryx Designer, capaz de equilibrar os fatores da precificação de fretes por caminhoneiros autônomos, garantindo a maior lucratividade para a Bialog.

Para atingir tal propósito, é realizado um estudo de caso com a empresa, que atua na contração de fretes, unindo motoristas aos clientes embarcadores. A fim de calibrar os modelos, faz-se necessária a obtenção de dados referentes a fretes anteriormente praticados, e para isso, as informações obtidas são provenientes dos sistemas de informação utilizados pela empresa.

A disponibilização de dados facilitada é uma característica determinante para a escolha do estudo de caso: a empresa conta com registros desde sua implantação e os cede a fim de estudos. A falta de padronização de alguns registros pode acarretar dificuldades de tratamento, porém não são impeditivos ao uso dos dados. Os dados sobre valores de diesel praticados em datas retroativas foram coletados através dos relatórios de levantamentos de preços da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

A elaboração da metodologia se fez através das etapas de aquisição e tratamento de dados, definição da amostra de trabalho, calibração e validação, e comparação de resultados entre o modelo criado e os modelos já praticados previamente. A Figura 4 ilustra a sequência lógica do estudo.

Figura 4 - Metodologia



Fonte: Autor (2024).

A etapa de fundamentação, apresentada no Capítulo 2, teve como finalidade elucidar os conceitos presentes no estudo, a teoria e os métodos utilizados, além de contextualizar o tema.

A aquisição dos dados fez-se necessária para definir quais tipos de dados seriam utilizados. A empresa registra dentro do sistema diversas informações sobre os fretes já realizados, mas é de suma importância entender quais estão minimamente relacionados à formação do preço de frete. Assim, existem centenas de fontes de dados, as quais devem ser escolhidas para utilização no estudo como variáveis candidatas.

Para evitar a perda de informações, a etapa de tratamento de dados visa organizá-las de forma concisa, unindo diferentes fontes de dados em uma única base, removendo informações repetidas, incompletas ou imprecisas.

Dadas as características de distribuição geográfica das rotas atendidas pela empresa e a localização dos clientes, é definida uma amostra dos fretes atendidos. Para a definição de quais clientes serão analisados, faz-se uso de ferramentas de

Sistemas de Informações Geográficas (SIG), especificamente o QGIS, para o mapeamento e representação gráfica.

Após coleta, tratamento e definição de amostra de trabalho inicia-se a construção dos modelos através do uso dos algoritmos preditivos no Alteryx Designer, comparando os diferentes algoritmos e selecionando o com melhores características.

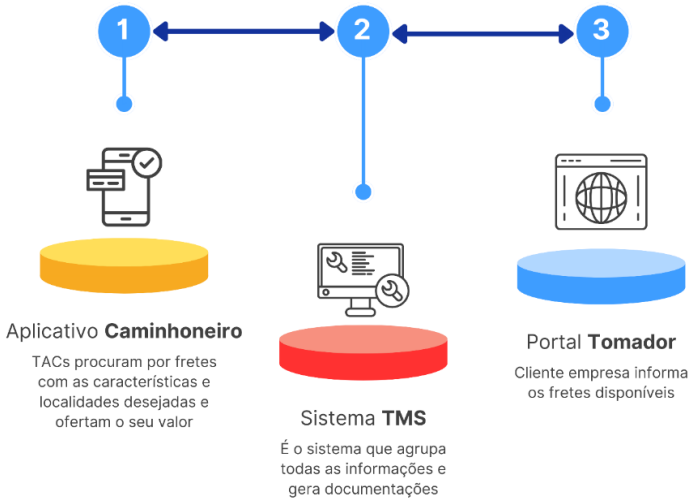
Por fim, compara-se o resultado do melhor modelo encontrado com os métodos já utilizados pela empresa e é definido o de melhor acurácia.

3.1 AQUISIÇÃO DE DADOS

Para o entendimento sobre os dados coletados, é necessário primeiro entender como eles são registrados em sistema. O Transportation Management System (TMS) utilizado pela empresa faz o registro das informações em tabelas, relacionando-as através de chaves. Desta forma, todos os dados foram disponibilizados através de planilhas, onde as informações possuem códigos identificadores, que se unem entre diferentes fontes de dados.

A empresa não possui frota própria, portanto, todos os fretes são realizados por TACs ou parceiros, contratados à distância, seja via aplicativo ou via atendimento digital. Assim, a empresa se denomina como uma transportadora digital, pois é através desse meio que se viabiliza sua operação. A Figura 5 mostra o relacionamento entre softwares da empresa.

Figura 5 - Fluxo de informações no sistema



Fonte: Autor (2024).

A Bialog denomina oferta de frete, toda possibilidade de frete disponibilizada pelo cliente tomador através do portal. O valor do motorista é chamado de lance. Uma oferta se concretizará em frete realmente, quando há acordo de valores entre motorista e cliente tomador, assim se iniciando o procedimento de carregamento.

Após o carregamento, serão emitidos os documentos de frete e realizado o pagamento, automaticamente. Quando então há geração da documentação com o frete acordado, é criada a viagem dentro do sistema. Dessa forma, oferta de frete no portal é uma possibilidade ainda não concretizada, enquanto viagem no TMS é um frete concretizado e documentado.

Na Figura 6 temos uma exemplificação de um relacionamento entre tabelas, onde um motorista é destacado na tabela de cadastro de motoristas. Esse mesmo, é indicado como o condutor de um certo veículo. Quando um condutor tem a intenção de realizar determinado frete, ele dá um lance com o valor desejado para o serviço em uma oferta. Essa combinação de motorista e veículo, junto ao valor pretendido, está dentro da tabela de lances de ofertas, pois o motorista tentou negociar o valor daquele frete.

Figura 6 - Relacionamento entre tabelas

Fonte: Autor (2024).

O registro dessas informações é feito entre as interações entre o cliente motorista, utilizador do aplicativo para celular, e do cliente tomador, utilizador do portal web, o contratante do serviço.

O Quadro 2 sintetiza as tabelas disponibilizadas, identificando-as conforme sua nomenclatura, principais informações obtidas através delas e as quantidades de informações contidas.

Quadro 2 - Tabelas de dados disponibilizadas pela Bialog

Código	Nome	Quantidade de linhas	Quantidade de colunas	Principais informações contidas
DA3	Veículos	13.938	55	Placas, tipos de veículos, número de eixos
DA4	Motoristas	15.368	94	Nome, informações pessoais, endereço de residência
DT6	Documentos	174.654	127	CTEs gerados, valor de notas fiscais, peso de mercadoria, quantidade de produtos, valor de imposto
DTQ	Viagens	45.438	60	UF origem e destino, veículo, motorista, data de emissão
DTR	Veículos por viagem	1.757.793	14	UF origem e destino, ID de oferta, valor combinado, placas, motorista
SA1	Clientes	36.293	116	CNPJ, coordenadas, informações fiscais, tipos de produtos, parametrizações e características
ZF0	Solicitantes	52.222	37	Sobre o tomador do serviço: razão social, código, UF origem
ZF1	Origens	54.082	59	Sobre o remetente: razão social, código, coordenadas
ZF2	Destinos	80.504	52	Sobre os destinatários: razão social, código
ZF3	Lances	54.457	71	Quilometragem total da rota, valor inicial ofertado

Fonte: Autor (2024).

Ademais, foi indispensável a coleta de dados sobre os valores de diesel, que não são contidos no sistema da empresa. A ANP é o órgão regulador das atividades que integram as indústrias de petróleo e gás natural e biocombustíveis no Brasil,

sendo assim, é a responsável pela coleta e disponibilização, entre outros dados, sobre o preço do diesel ao longo dos anos nos estados brasileiros. Através do relatório de Levantamento de Preços de Combustíveis (LPC), foram tratados os valores de diesel, agrupando-os por mês, ano e estado, realizando a média entre os valores observados.

O relatório LPC faz o levantamento semanalmente em postos de gasolina de cada cidade, levantando os preços de revenda praticados. Não são todas as cidades que possuem o levantamento, porém todos os estados possuem coleta.

3.2.1 Método de precificação Bialog

A empresa possui um método para previsão de valores de fretes solicitados pelos motoristas, porém esse é pouco utilizado e normalmente o trabalho de levantamento de preços é feito a partir da demanda recebida, de atendimento a atendimento.

O método é utilizado normalmente como uma forma de validação de tabelas de frete pré-definidas por clientes, avaliando-se a aderência dos valores aos praticados no mercado.

Para a definição dos preços de fretes, são levados em consideração dois valores: um valor de diária mínima e um valor baseado no custo por quilometragem. Se o custo por quilometragem for inferior ao de diária mínima, esse será desconsiderado e o valor previsto será o da diária.

Para a definição dos custos mínimos, torna-se necessário definir para cada tipo de veículo, algumas informações que serão utilizadas.

A Figura 7 representa a tabela de informações por tipo de veículo utilizadas no método. Cada veículo possui capacidades e quantidade de eixos definidos. O consumo por litro é um dado que a empresa determinou de forma empírica através da sua gestão, estimando um valor médio para os veículos de cada tipo. O valor de diária será definido pelo produto entre a quantidade de eixos, a quantidade de horas ociosa e um fator determinado.

Figura 7 - Tabela de parâmetros Bialog

				MOTORISTA			
VEÍCULOS	EIXOS	CONSUMO/L	CAPACIDADE	CUSTO KM RODADO	OTORISTA KM RODAD	DIÁRIA	FREETIME
3/4	2	5,5	3,5	R$ 1,35	R$ 2,69	R$ 159,04	R$ 139,16
TRUCK	3	3,6	14	R$ 2,06	R$ 4,12	R$ 636,16	R$ 556,64
TRUCK	4	3,6	16	R$ 2,06	R$ 4,12	R$ 727,04	R$ 636,16
CARRETA 25 TON	5	2,2	25	R$ 3,37	R$ 6,74	R$ 1.136,00	R$ 994,00
CARRETA 32 TON	6	2,2	32	R$ 3,37	R$ 6,74	R$ 1.454,08	R$ 1.272,32
BITREM 48 TON	7	1,7	48	R$ 4,36	R$ 8,72	R$ 2.181,12	R$ 1.908,48
BITREM 65 TON	8	1,5	65	R$ 4,94	R$ 9,88	R$ 2.953,60	R$ 2.584,40
RODOTREM 74 TON	9	1,5	74	R$ 4,94	R$ 9,88	R$ 3.362,56	R$ 2.942,24

Fonte: Autor (2024).

O valor baseado no custo por quilometragem é formado através da multiplicação da quilometragem total da rota avaliada pelo custo por quilômetro rodado, sendo esse obtido através da divisão do valor de combustível pelo consumo por litro.

O método apresentado é extremamente útil, mesmo sendo formado de maneira empírica e baseado na experiência dos gestores, pois a empresa atualmente conta apenas com o método da ANTT como alternativa e ainda não dispõe de uma ferramenta mais adequada.

3.2 TRATAMENTO DE DADOS

O tratamento dos dados tem suma importância no desenvolvimento do estudo, pois a coesão das informações faz com que o modelo de previsão criado seja o mais confiável possível. Essa etapa é a que demanda mais custo de tempo e recursos computacionais.

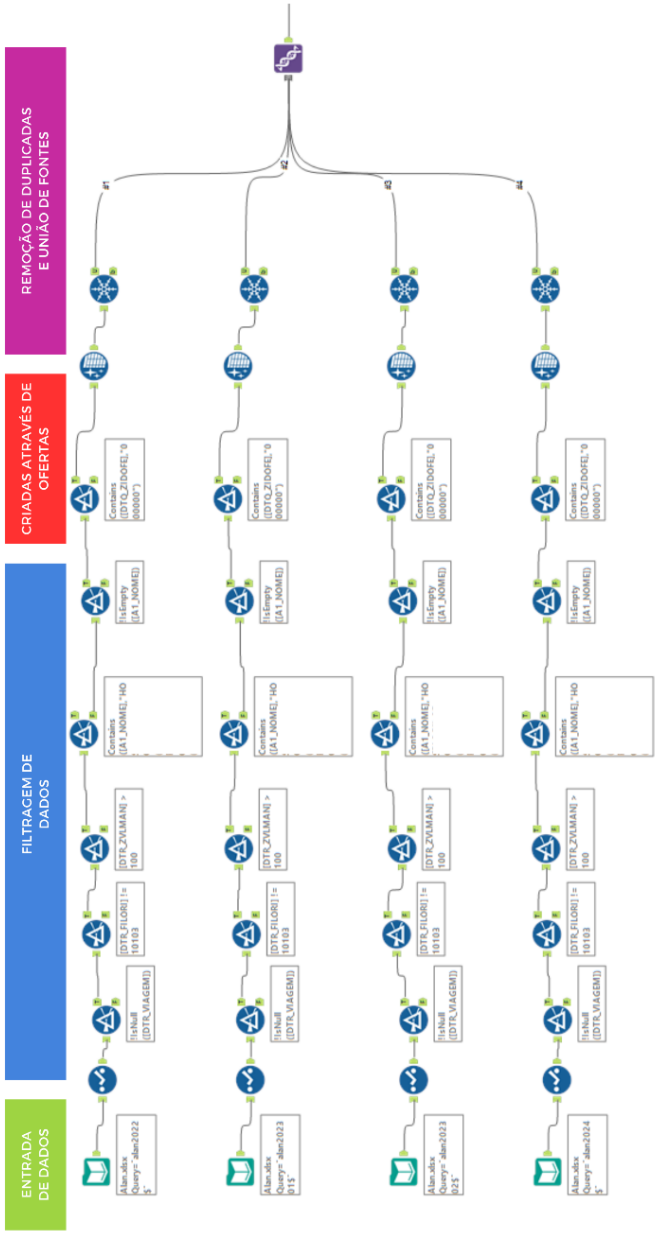
Nos dados fornecidos para o estudo, tem-se muitas particularidades que precisam ser tratadas ou ajustadas. O uso do Alteryx Designer potencializa a confiabilidade e qualidade dos tratamentos.

3.2.1 Tratamento das viagens

O início do tratamento se dá através da inclusão das planilhas de viagens, que são divididas em quatro arquivos. Primeiramente, são excluídos antecipadamente os dados que, devido a características específicas, não são relevantes para a análise. Através da ferramenta filtro, são excluídas as viagens que representam operações onde não há uma nova negociação a cada frete, seja identificando-os pela filial da

empresa ou pelo cliente atendido. São removidos também os dados onde o valor do frete combinado com o motorista é inferior a cem reais, pois baseando-se em procedimentos da empresa, viagens de correção de documentação ou problemas específicos são corrigidos com a geração de um novo ID de viagem, porém com valor combinado abaixo do padrão. Se mantidos esses dados, haveria duplicidade de informações em casos específicos com valores distorcidos da realidade, o que enviesaria o resultado. A Figura 8 mostra o primeiro tratamento dos dados no software.

Figura 8 - Tratamento inicial de viagens

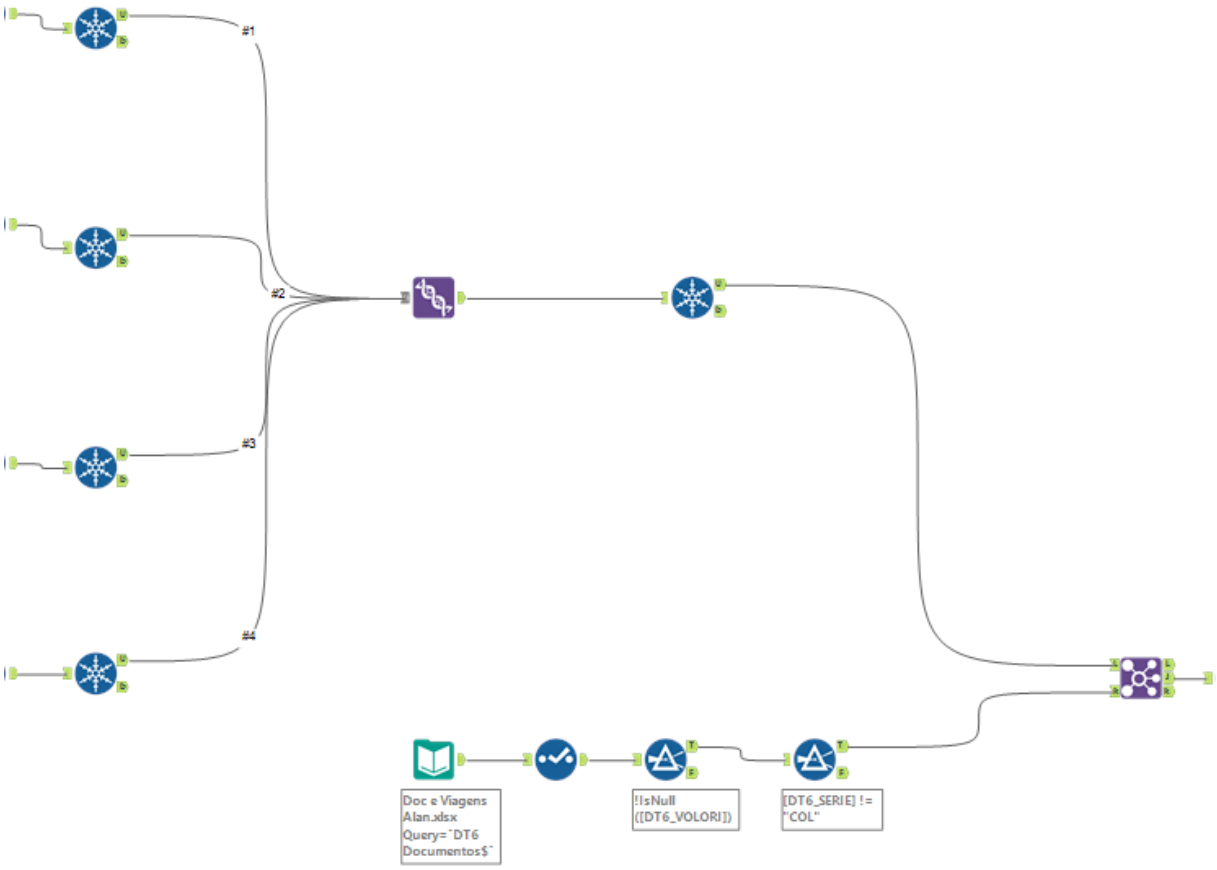


Fonte: Autor (2024).

O ID de viagem é único e representa um frete realizado, porém, nos dados exportados, temos que um mesmo ID de viagem pode ser repetido em uma nova linha em casos em que há mais de um documento dentro utilizado no frete. Assim, é necessário para a análise, considerarmos o número de viagens apenas uma única vez junto a um único número de CTe.

Para que sejam incluídos os dados sobre a carga, é incluída a entrada de dados dos documentos gerados, unindo essa entrada ao fluxo já tratado anteriormente, conforme observado na Figura 9.

Figura 9 - Inserção dos dados de documentos gerados



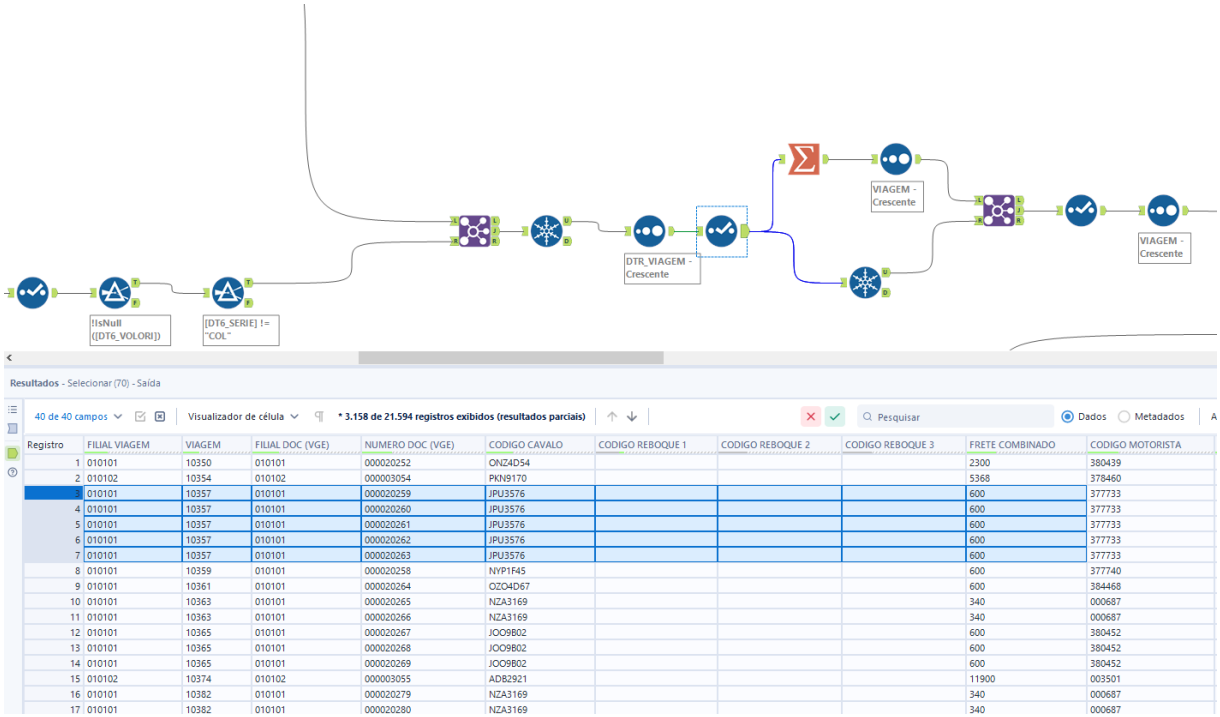
Fonte: Autor (2024).

O número de documento já aparecia dentro da viagem anteriormente e agora, será a chave a ser buscada na tabela de documentos, trazendo as informações de peso da carga, valor de nota fiscal, valor de impostos, entre outras informações. São

excluídos os dados que se referem a documentos auxiliares, que não representam diretamente um documento de um frete.

Dentro dos registros, encontram-se dados que são inseridos apenas uma vez dentro da viagem, como é o caso do frete combinado com o motorista, mas aparecem a cada vez que o ID de viagem aparece. Já outros dados são diferentes a cada ID de documento diferente. Dessa forma, a partir desse ponto do tratamento, algumas informações precisam ser sumarizadas e agrupadas, somadas ou calculadas através de médias. Na Figura 10, tem-se como exemplo a viagem 10357, onde existem cinco CTEs diferentes para a mesma.

Figura 10 - Exemplo de viagem com documentos múltiplos



Registro	FILIAL VIAGEM	VIAGEM	FILIAL DOC (VGE)	NUMERO DOC (VGE)	CODIGO CAVALO	CODIGO REBOQUE 1	CODIGO REBOQUE 2	CODIGO REBOQUE 3	FRETE COMBINADO	CODIGO MOTORISTA
1	010101	10350	010101	000020252	ONZ4D54				2300	380439
2	010102	10354	010102	000003054	PKN9170				5368	378460
3	010101	10357	010101	000020259	JPU3576				600	377733
4	010101	10357	010101	000020260	JPU3576				600	377733
5	010101	10357	010101	000020261	JPU3576				600	377733
6	010101	10357	010101	000020262	JPU3576				600	377733
7	010101	10357	010101	000020263	JPU3576				600	377733
8	010101	10359	010101	000020258	NYP1F45				600	377740
9	010101	10361	010101	000020264	OZO4D67				600	384468
10	010101	10363	010101	000020265	NZA3169				340	000687
11	010101	10363	010101	000020266	NZA3169				340	000687
12	010101	10365	010101	000020267	JOO9B02				600	380452
13	010101	10365	010101	000020268	JOO9B02				600	380452
14	010101	10365	010101	000020269	JOO9B02				600	380452
15	010102	10374	010102	000003055	ADB2921				11900	003501
16	010101	10382	010101	000020279	NZA3169				340	000687
17	010101	10382	010101	000020280	NZA3169				340	000687

Fonte: Autor (2024).

Para tratar esse problema, utiliza-se a ferramenta sumarizar. As viagens passam a ser agrupadas pelo seu ID e dessa forma, quando há mais de uma linha com ele, selecionam-se os campos que serão utilizados para algum tratamento. Na Figura 11, temos que os dados agrupados pelo número de viagem, serão somados. Assim, temos o valor total de cada campo, considerando todos os documentos presentes em uma mesma viagem.

Figura 11 - Ferramenta sumarizar

Fonte: Autor (2024).

Para a inclusão da quilometragem total do frete, necessita-se a inclusão do dado de entrada dos lances dos motoristas nas ofertas, pois essa informação é registrada sempre que um novo interessado em um frete é registrado. Além disso, é necessário considerar que, independentemente da quantidade de entregas, essa quilometragem é calculada com base na origem remetente, ao último destinatário. Ofertas com mais de um lance motorista registram a mesma informação de quilometragem, sendo assim, eliminamos a aparição de dois lances em mesma oferta.

A próxima etapa, é a inclusão dos dados sobre os veículos utilizados em cada frete. Para isso, é inserida a tabela de veículos, onde a chave de junção do fluxo à tabela será o código do veículo, que é a própria placa.

Foram identificados ao todo, 49 clientes únicos na base de dados tratada. Muitos desses não trabalham restritamente a um produto específico, mas sim a produtos de mesma característica. Para sintetizar o tipo de produto que cada cliente carrega, foram agrupados os produtos de forma a agrupá-los por segmento. Uma nova tabela foi criada, relacionando cada cliente ao seu produto e, posteriormente, novamente unida no fluxo de trabalho do Alteryx, para a inclusão dos produtos em cada viagem.

O Quadro 3 relaciona o tipo de produto ao seu código.

Quadro 3 - Produtos carregados

Produto	Código do produto
Material escolar	1
Plásticos	2
Diversos	3
Produtos de limpeza	4
Isopor	5
Bebidas	6
Painel solar	7
Isotérmicos	8
Insumos de leite	9
Móveis e eletrodomésticos	10
Material de construção	11
Aço	12
Grãos	13
Ração	14
Papel	15
Alimentos	16

Fonte: Autor (2024).

Pode-se observar que nos 49 clientes únicos atendidos, há 16 tipos de produtos diferentes sendo carregados. Os produtos mais comuns são os de bebidas e material de construção.

3.2.2 Tratamento dos valores de diesel

Um dos principais custos de frete é o valor do combustível, portanto é indispensável a coleta de dados sobre os valores de diesel, que não são contidos no sistema da empresa.

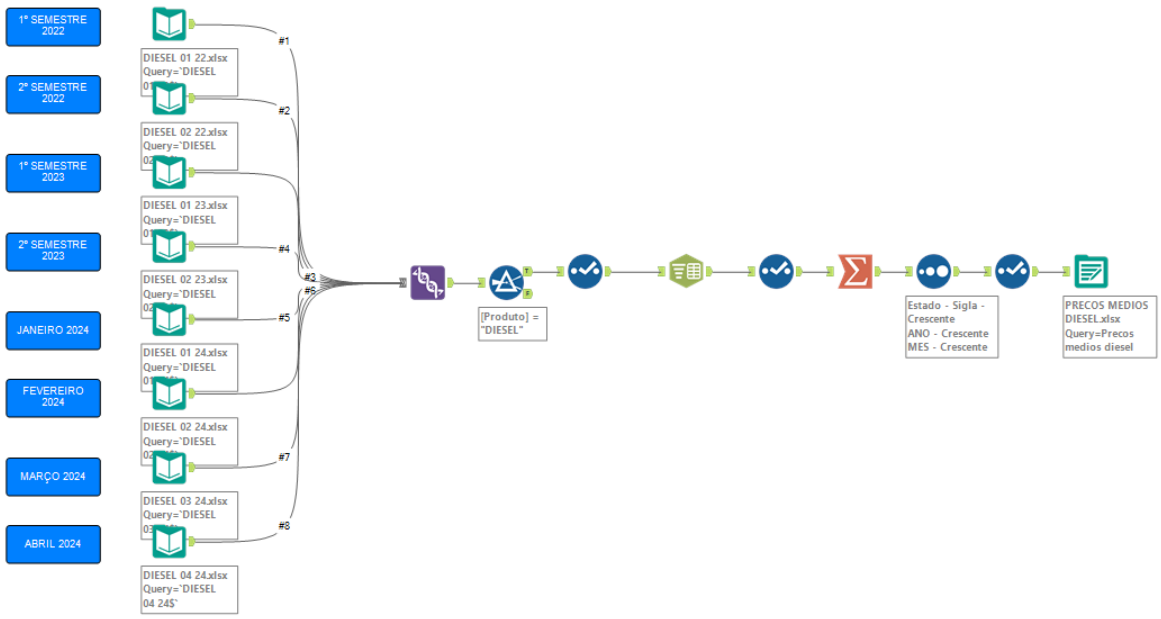
A ANP é o órgão regulador das atividades que integram as indústrias de petróleo e gás natural e biocombustíveis no Brasil, sendo assim, é a responsável pela coleta e disponibilização, entre outros dados, sobre o preço do diesel ao longo dos anos e estados brasileiros. Através do relatório LPC, foram tratados os valores de diesel, agrupando-os por mês, ano e estado, realizando a média entre os valores observados. O relatório LPC faz o levantamento semanalmente em postos de gasolina de cada cidade, levantando os preços de revenda praticados. Não são todas as cidades que possuem o levantamento, porém todos os estados possuem coleta.

Os dados são disponibilizados pela ANP através de tabelas. A série referente aos anos anteriores, é dividida entre semestres do ano, enquanto as séries do ano em

exercício é feita mês a mês. A coleta desses dados foi feita em maio de 2024, sendo assim até o momento, os dados utilizados foram divididos em oito planilhas, onde há duas para cada ano anterior e uma para cada mês do ano atual.

A Figura 12 a seguir ilustra o tratamento das planilhas de entrada dos dados.

Figura 12 - Módulo de tratamento dos dados de diesel



Fonte: Autor (2024).

Ao ser combinado um frete, é política da empresa realizar o adiantamento de no mínimo 70% do valor como forma de adiantamento para a prestação do serviço. De tal forma, presume-se que o principal abastecimento para a execução do serviço se dá na cidade de origem do carregamento da carga. No LPC não há levantamento em todas as cidades onde a empresa atua, dessa forma, optou-se por definir a média preço do combustível na UF de origem. Através da ferramenta sumarizar, foram agrupadas as informações da sigla de cada estado, para cada mês e ano. Sendo assim, são obtidos valores médios de diesel por data e localidade.

3.2.3 Correção de valores monetários

Os dados obtidos são defasados pois trata-se de valores monetários retroativos, sendo assim, é necessária a correção dos valores equiparando-os aos valores praticados atualmente. Para a atualização dos valores, utiliza-se o Índice

Nacional de Preços ao Consumidor Amplo (IPCA). Através desse, pode-se calcular o valor atualizado corrigido de acordo com a inflação.

A equação 7 aplica um fator de correção a um valor praticado anteriormente, resultando ao seu equivalente nos parâmetros atuais.

$$V_c = V_i \times \frac{F_{ATUAL}}{F_{ACUMULADO} \times (\%_{ANTERIOR} + 1)} \quad (8)$$

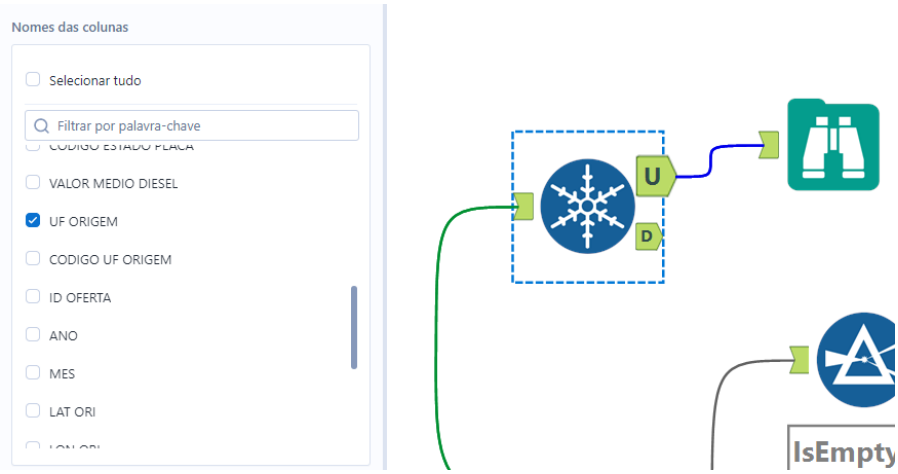
Para obter o valor corrigido V_c, utiliza-se do produto entre o valor inicial V_i e a razão entre o valor F_{ATUAL} que representa índice atual, com o produto do número $F_{ACUMULADO}$ que representa o índice acumulado até mês anterior à data observada.

Para a devida correção, todos os valores representados de forma monetária são calculados e ajustados.

3.3 DEFINIÇÃO DE VARIÁVEIS CANDIDATAS

Nos dados estão contidas as rotas atendidas desde o início de 2022 até o mês de abril de 2024, sendo esses fretes os que contém as informações tratadas anteriormente. Utilizando-se da ferramenta Exclusivo no Alteryx, pode-se definir quais unidades federativas foram atendidas, dividindo-as entre origens e destinos. Para a análise de origens únicas, o campo UF ORIGEM é marcado, enquanto para a análise de destinos únicos, o campo UF DESTINO é marcado. Na Figura 13, a configuração da ferramenta Exclusivo é mostrada.

Figura 13 - Exemplo de uso da ferramenta Exclusivo



Fonte: Autor (2024).

Após a execução do fluxo de trabalho, verifica-se que existem 17 origens e 26 destinos atendidos, das 27 unidades federativas brasileiras totais. Através da análise dos dados, observa-se que a empresa atende alguns clientes de forma esporádica, onde não há recorrência de carregamentos e, portanto, esses dados podem influenciar negativamente no resultado das predições. A fim de utilização de dados robustos, é tomada a decisão pela análise onde há recorrência, criando uma amostra de trabalho. Para defini-la, utilizou-se de ferramentas de SIG, como o QGIS para análise de distribuição geográfica e de recorrência.

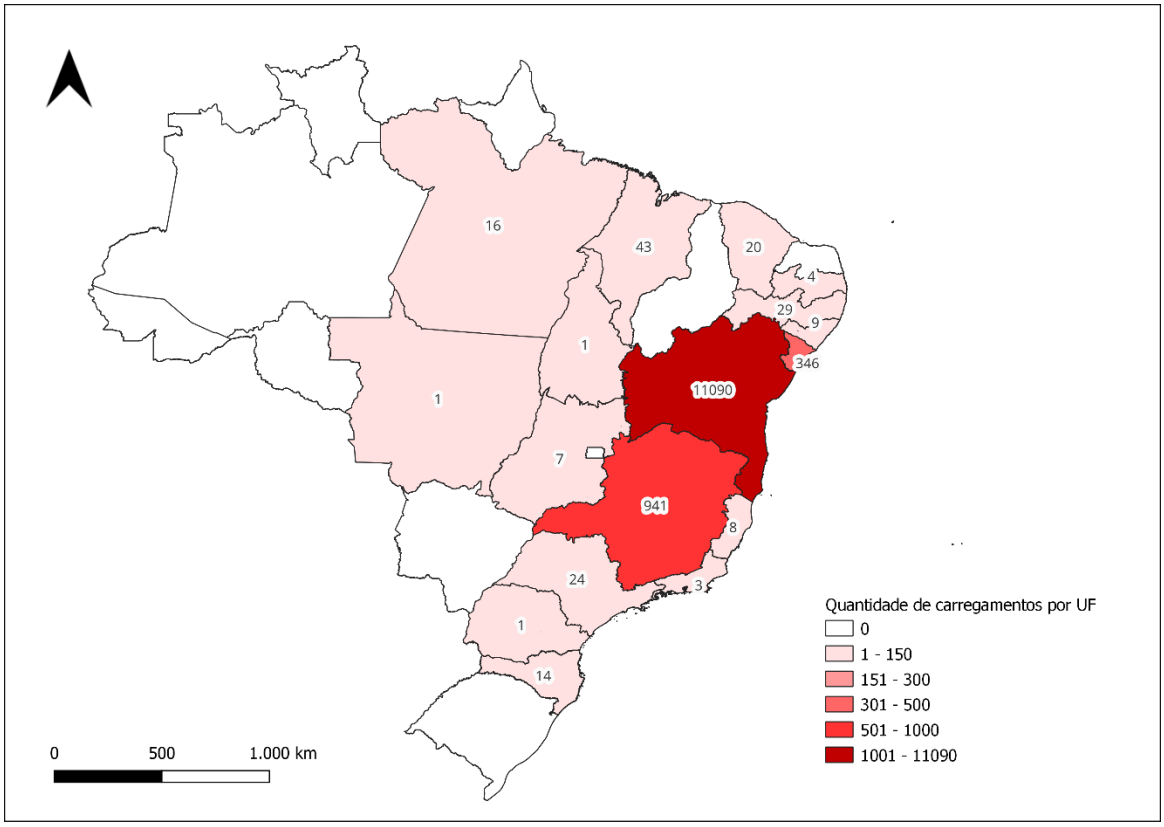
Os SIGs são programas que necessitam como entrada dados *shapefile* e dados referenciados através de localização. Para isso, todos os dados de origens e destinos necessitam do seu par de coordenadas de latitude e longitude.

Uma das dificuldades enfrentadas foi o tratamento do dado de geolocalização. Foi identificada uma falha de registro nos dados, sendo assim, alguns dos clientes finais não tinham a informação correta. A quilometragem total do frete não é impactada pela falha, pois seu cálculo vem de outra tabela fonte. Para evitar o descarte do frete, optou-se por abstrair a informação, fornecendo o par de localização referente à cidade de destino. Para tal, foi construído um script baseado em *Javascript*, para a coleta das coordenadas em lote de forma automatizada através de Planilhas Google. Ao fim da coleta, os novos dados obtidos foram inseridos novamente no fluxo, gravando a coordenada no campo de latitude e longitude do cliente destino.

Para a confecção dos mapas, foram utilizados os arquivos shapefile disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) através do site oficial do órgão.

Ao todo após o procedimento de tratamento dos dados, foram compilados 12.557 fretes. Para definição da amostra, optou-se por trabalhar com os clientes de maior volume. Essa análise é feita através da quantidade de fretes realizados para cada cliente. Através da representação na Figura 14, constata-se que os clientes com origem em Bahia e Minas Gerais são os de maior volume e compõem a amostra de trabalho.

Figura 14 - Distribuição de fretes geral



Fonte: Autor (2024).

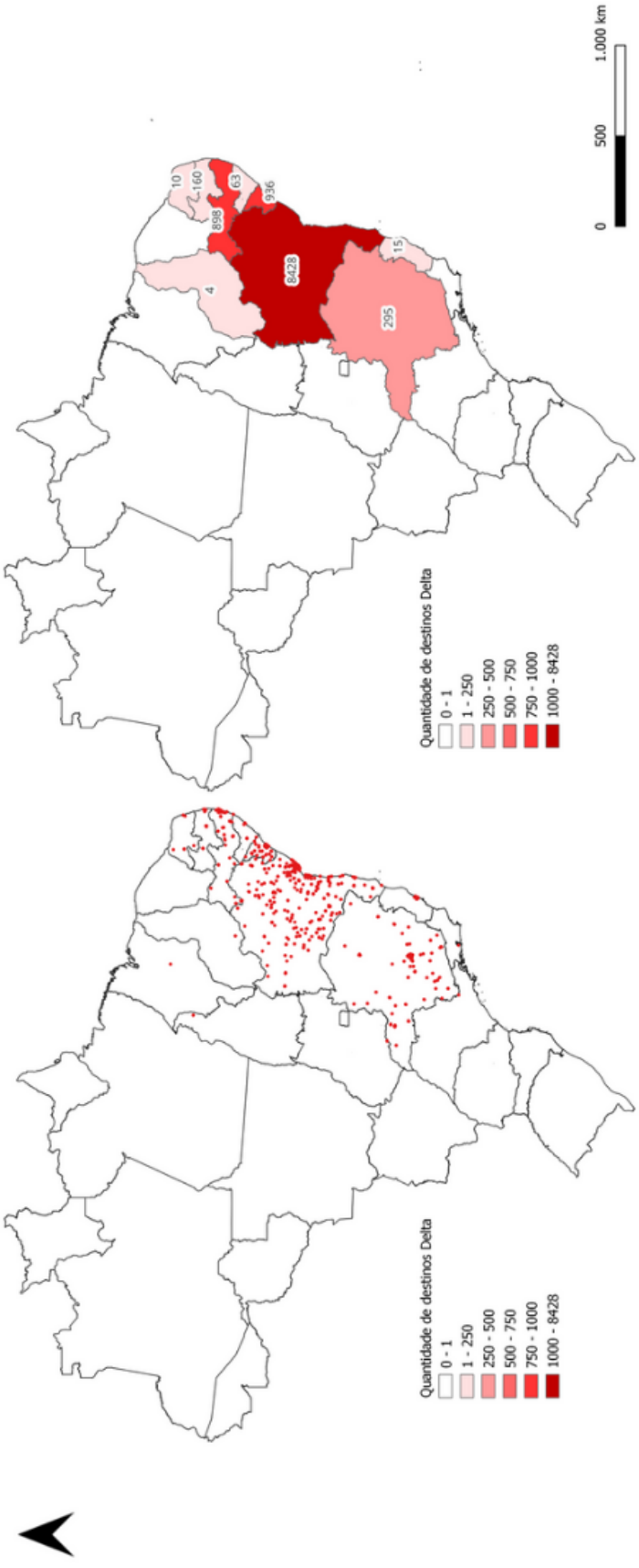
A Figura 15 representa o cliente com origem na Bahia, que é identificado pelo nome Delta. Esse cliente é o de maior recorrência da empresa, com motoristas considerados fidelizados e que já são em sua maioria, acostumados com a precificação. A principal área de atuação desse cliente que distribui água mineral é no Nordeste, seguido pelo Centro-Oeste.

Esse cliente possui rotas bem definidas, onde há um grande número de atendimentos em fretes intermunicipais dentro do próprio estado da Bahia, com mais de 8.400 fretes praticados.

O segundo maior cliente, identificado como Kappa, tem sua origem em Minas Gerais. Esse cliente atua com a produção de isotérmicos. Sua principal característica é a distribuição por rotas mais longas, interestaduais. Sua atuação é mais acentuada no centro-oeste, porém há uma boa distribuição pelo território nacional (Figura 16).

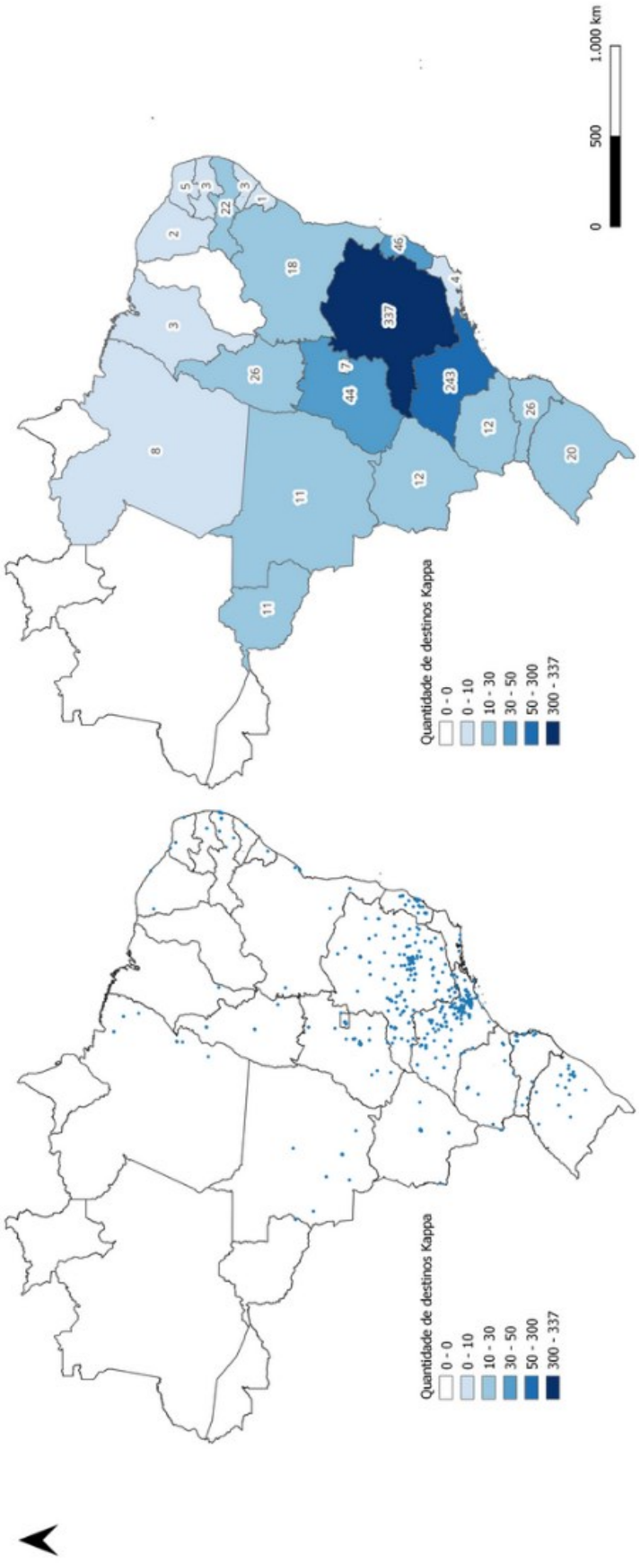
Dessa forma, os fretes realizados para os clientes Delta e Kappa são definidos como os dados que serão utilizados no estudo.

Figura 15 - Distribuição de fretes cliente Delta



Fonte: Autor (2024).

Figura 16 - Distribuição de fretes Kappa



Fonte: Autor (2024).

Para a criação dos modelos de predição, foram utilizados os dados padronizados da mesma base tratada simultaneamente. As variáveis candidatas utilizadas e a descrição das informações contidas estão dispostas no Quadro 4.

Quadro 4 - Campos e descrições

Nome do campo	Descrição da informação
CLIENTE	Razão social do cliente
CODIGO CLIENTE	Código do cliente
PRODUTO	Abstração do tipo de produto
CODIGO PRODUTO	Código do produto
KM TOTAL	Quilometragem total
FRETE COMBINADO	Valor do frete combinado com o TAC, essa será a variável alvo
TOTAL CTE	Total do serviço prestado pago pelo cliente empresa
IMPOSTO	Valor de imposto de ICMS
TOTAL NFS	Valor total dos produtos carregados
VOLUMES	Quantidade de volumes presente na carga
PESO	Peso da carga
N EIXOS	Número de eixos do veículo
ESTADO PLACA	Estado de licenciamento do veículo
CODIGO ESTADO PLACA	Código do estado de licenciamento
VALOR MEDIO DIESEL	Valor médio do diesel praticado na data e local
ID OFERTA	Número de identificação da oferta
ANO	Ano em que foi realizado o frete
MES	Mês em que foi realizado o frete
CODIGO CLIENTE ORIGEM	CNPJ do cliente origem
ENDERECO ORI	Endereço do remetente
BAIRRO ORI	Bairro do remetente
MUN ORI	Município do remetente
UF ORIGEM	UF do remetente
CODIGO UF ORIGEM	Código da UF remetente
LAT ORI	Latitude do remetente
LON ORI	Longitude do remetente
CODIGO CLIENTE DEST	CNPJ do cliente destinatário
ENDERECO DEST	Endereço do destinatário
BAIRRO DEST	Bairro do destinatário
MUN DEST	Município do destinatário
UF DEST	UF do destinatário
LAT DEST	Latitude do destinatário
LON DEST	Longitude do destinatário

Fonte: Autor (2024).

A entrada de dados para os modelos de predição foi definida através de um arquivo de planilha de 33 colunas e 11.675 linhas, chamado de ENTRADA REGRESSAO.

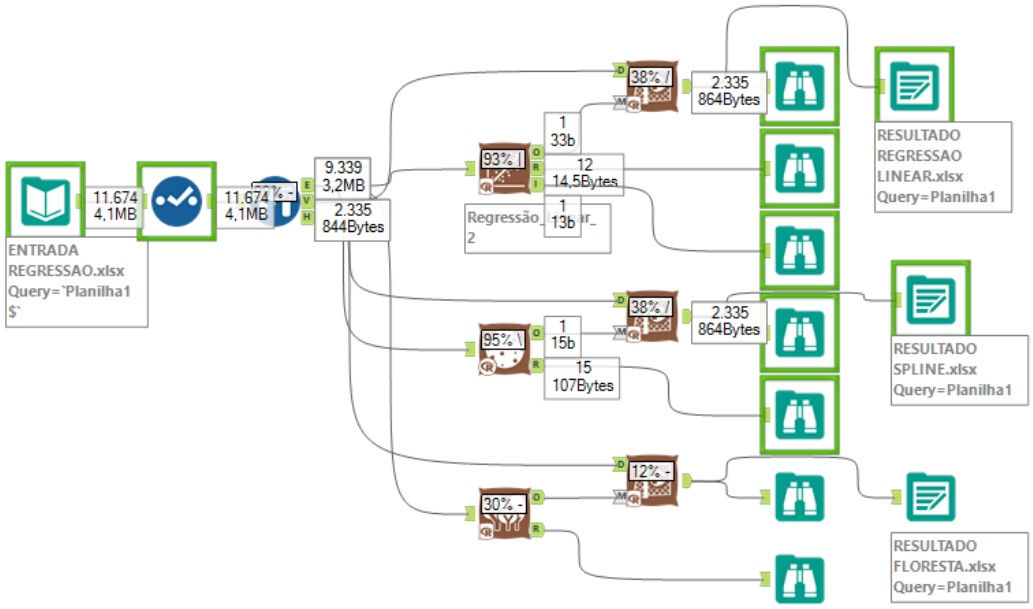
3.4 CALIBRAÇÃO E VALIDAÇÃO DOS MODELOS

Para a construção dos modelos de previsão de fretes, foi utilizado o módulo de predição do software Alteryx Designer. Para tal, os métodos testados foram calibrados através do uso das ferramentas de Regressão Linear, Modelo Spline e Modelo de Floresta.

Diferentes métodos foram testados para aferir a eficiência de cada um deles dada a base de dados selecionada.

O arquivo contendo as variáveis candidatas do grupo selecionado previamente é utilizado agora em um novo fluxo de trabalho dentro do programa. Para o uso das informações, primeiramente é utilizada a ferramenta de entrada de dados, vinculando a planilha como entrada. O fluxo em execução é mostrado na Figura 17.

Figura 17 - Fluxo de trabalho em execução



Fonte: Autor (2024).

Para que não existam problemas com conversão de valores, foi utilizada a ferramenta Selecionar, que permite que os tipos de dados sejam definidos optando-

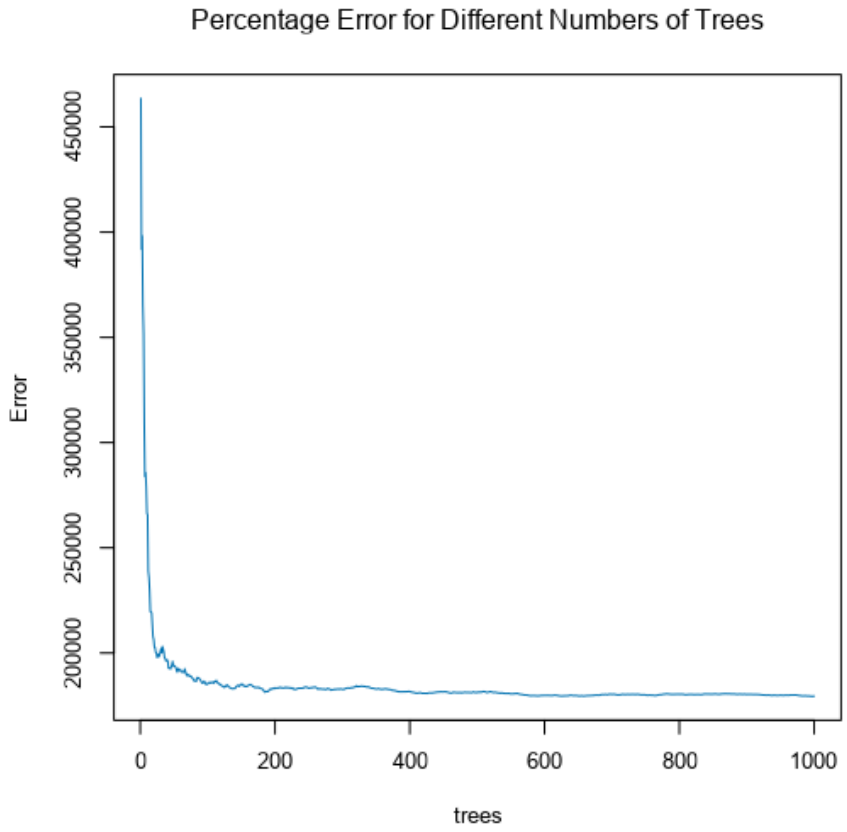
se por tipos numéricos, booleanos ou *strings*. Posteriormente, através da ferramenta Criar Amostra, os dados são divididos em duas amostras: de teste e de validação.

A amostra de teste foi definida em 80% e será a utilizada para a calibração dos modelos, enquanto os 20% restantes serão utilizados para a validação.

O Modelo de Floresta possui uma particularidade em relação aos demais, sendo necessário informar a quantidade de árvores a serem utilizadas no modelo. Baseando-se no baixo custo computacional e de tempo, empiricamente foi definida a quantidade de 1000 árvores. O número se torna além do necessário, mas garante uma estabilização ainda maior dos valores a um baixo custo.

A Figura 18 mostra graficamente a estabilização dos erros pela quantidade de árvores inseridas.

Figura 18 - Erro pela quantidade de árvores



Fonte: Autor (2024).

Para a validação de cada modelo, é utilizada a ferramenta browser no fluxo. Através dela, são gerados relatórios estatísticos que avaliam a confiabilidade de cada método.

Para a comparação dos modelos, são observados os valores de R^2 ou o percentual da variância explicada, pois esse valor indica o quão bem explicada é a variável dependente pelas variáveis independentes selecionadas. A correlação das variáveis analisadas se fez através da utilização da correlação de Pearson presente na Seção 2.3.1 para os campos que contém informações quantitativas sobre o frete (Figura 19).

Figura 19 – Correlação entre variáveis

	KM TOTAL	FRETE COMBINADO	TOTAL CTE	IMPOSTO	TOTAL NFS	VOLUMES	PESO	N EIXOS	VALOR MEDIO DIESEL	ANO	MES
KM TOTAL	1,0000	0,9132	0,8951	0,3812	0,5115	0,0075	0,4252	0,5884	-0,0779	0,1012	0,0004
FRETE COMBINADO	0,9132	1,0000	0,9786	0,3499	0,5962	-0,0148	0,5830	0,7480	-0,0940	0,1232	-0,0050
TOTAL CTE	0,8951	0,9786	1,0000	0,4066	0,5696	-0,0123	0,5960	0,7371	-0,1075	0,1311	-0,0158
IMPOSTO	0,3812	0,3499	0,4066	1,0000	0,0859	0,0057	0,3009	0,2148	-0,1258	0,2334	-0,0170
TOTAL NFS	0,5115	0,5962	0,5696	0,0859	1,0000	-0,0164	0,3094	0,6477	-0,0971	0,1275	0,0049
VOLUMES	0,0075	-0,0148	-0,0123	0,0057	-0,0164	1,0000	-0,0443	-0,0237	-0,0025	0,0059	-0,0035
PESO	0,4252	0,5830	0,5960	0,3009	0,3094	-0,0443	1,0000	0,6935	-0,0064	0,0348	-0,0003
N EIXOS	0,5884	0,7480	0,7371	0,2148	0,6477	-0,0237	0,6935	1,0000	-0,0299	0,0606	0,0206
VALOR MEDIO DIESEL	-0,0779	-0,0940	-0,1075	-0,1258	-0,0971	-0,0025	-0,0064	-0,0299	1,0000	-0,6821	0,1946
ANO	0,1012	0,1232	0,1311	0,2334	0,1275	0,0059	0,0348	0,0606	-0,6821	1,0000	-0,6311
MES	0,0004	-0,0050	-0,0158	-0,0170	0,0049	-0,0035	-0,0003	0,0206	0,1946	-0,6311	1,0000

Fonte: Autor (2024).

Algumas características da empresa também influenciam a tomada da escolha de variáveis a serem incluídas nos modelos: o valor definido na variável TOTAL CTE apesar de possuir grande correlação, não explica o valor de frete combinado. Esse valor é o que representa o valor final da prestação do serviço contratado pela Bialog ao cliente embarcador, no entanto, ele é formado proporcionalmente a partir do valor de frete combinado com os TACs, sendo assim está sempre correlacionado, mas sem agregar informação explicativa.

Os três modelos aplicados têm como variável alvo o FRETE COMBINADO, pois é esse campo que representa o valor pago ao motorista pelo serviço. Como variáveis preditoras, foram selecionados os campos KM TOTAL, IMPOSTO, TOTAL NFS, VOLUMES, N EIXOS, VALOR MEDIO DIESEL, ANO e MÊS.

A correlação das variáveis analisadas se fez através da utilização da correlação de Pearson para os campos que contém informações quantitativas sobre o frete.

Ao executar o fluxo de trabalho final, será criada uma coluna Score junto aos dados, onde está contido o valor previsto pelo modelo. Ao fim da comparação de resultados, apenas o método de melhor previsão será mantido.

4. ANÁLISE DE RESULTADOS

Com a calibração e validação dos modelos efetuada de forma satisfatória, a próxima etapa do estudo visa comparar os resultados obtidos. Após o melhor modelo ser definido, este será comparado aos métodos de precificação da ANTT e da Bialog para determinar qual é o mais eficaz para prever a precificação de fretes por motoristas autônomos.

Nesta seção, serão apresentados os principais resultados para cada método utilizado e suas principais características. O principal teste comparativo será feito através do teste estatístico R^2, avaliando-se os coeficientes de inclinação de reta e interceptos, além da estimativa de erros RMSE.

4.1 ESCOLHA DO MELHOR MODELO VALIDADO

Através do Alteryx Designer foram criados fluxos de trabalho para o tratamento e união de dados, bem como a calibração e validação de modelos de predição. Através da execução dos modelos, obtém-se resultados de previsão de valores gerados, que são exportados em planilhas de dados. O fluxo de trabalho dos métodos obteve os resultados em dois minutos e onze segundos de execução, sendo esse um tempo curto e de baixo custo computacional.

Toda a base de dados dos fretes realizados por clientes foi padronizada, sendo assim, todos possuem a mesma quantidade de dados e para a avaliação dos modelos, foram utilizadas as mesmas variáveis explicativas em todos os métodos. A principal variável explicativa identificada nos três métodos foi a KM.TOTAL, que indica a distância total do frete. De fato, esse era um resultado esperado, pois a precificação é baseada em custo e esse cresce conforme o tamanho da rota a ser praticada.

A segunda variável de maior impacto obtida foi a N.EIXOS, que indica a quantidade de eixos que o veículo possui. Essa variável abstrai em número de eixos os diferentes tipos de veículos que têm diferentes custos que compõem o seu frete.

Para o comparativo entre métodos, foram utilizados os indicadores do coeficiente R^2 e o erro RMSE. A tabela 1 exibe o comparativo entre os métodos.

Tabela 1 - Comparativo entre métodos preditivos

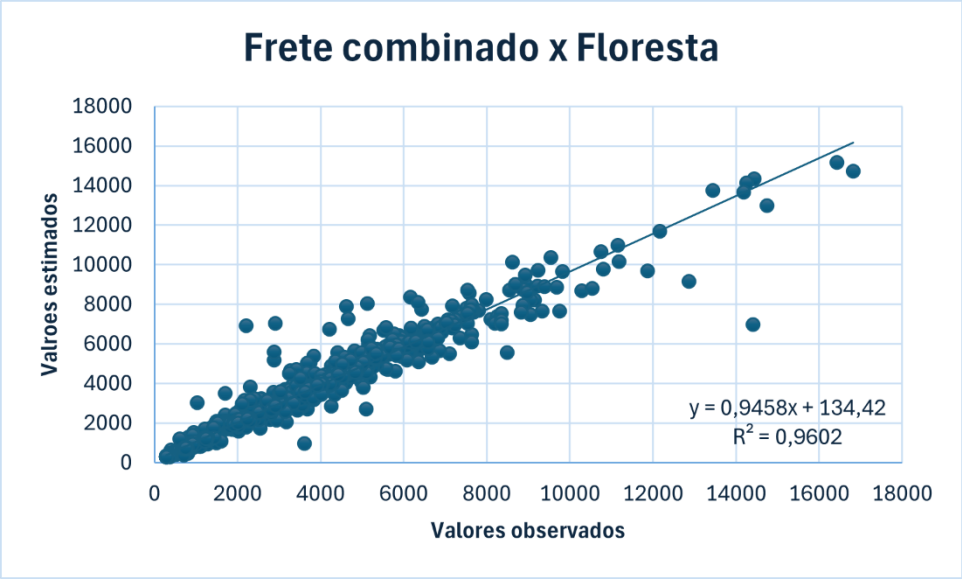
Modelos	R^2	α	β	RMSE
Regressão Linear	0,9243	200,08	0,9158	551,5627
Modelo Spline	0,9355	159,30	0,9329	523,1088
Modelo de Floresta	0,9602	134,42	0,9458	393,0512

Fonte: Autor (2024).

Os métodos resultaram em valores de R^2 próximos a 1, indicando que possuem uma modelagem onde as variáveis selecionadas explicam bem o fenômeno que busca ser explicado. Seguindo o disposto por Lopes (2010), os métodos apresentam também valores de α perto de 0, representando valores satisfatórios, além de β próximo de 1. O teste do RMSE indica que os erros obtidos foram baixos, dadas as dimensões de valores de fretes.

A Figura 20 ilustra os valores estimados para o frete combinado através do Método de Floresta, através de um gráfico de dispersão com eixo X representando o valor observado e Y os valores previstos. Com R^2 de 0,9602, α equivalente a 132,42 e β de 0,9458, além de RMSE de 393,0512, o método se destaca em relação aos demais e obtém o melhor resultado, pois em todos os indicadores ele é o mais bem avaliado.

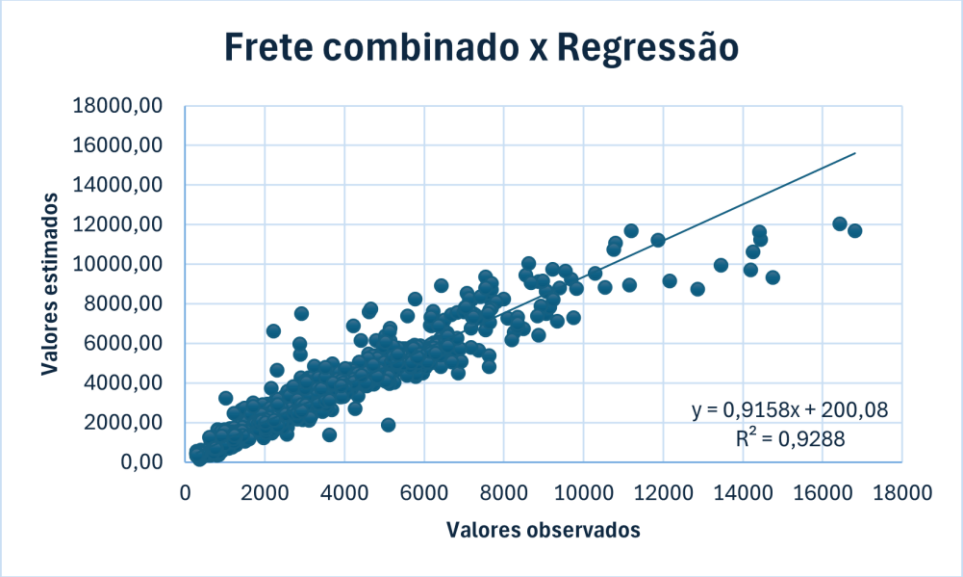
Figura 20 – Valores estimados pelo Método de Floresta



Fonte: Autor (2024).

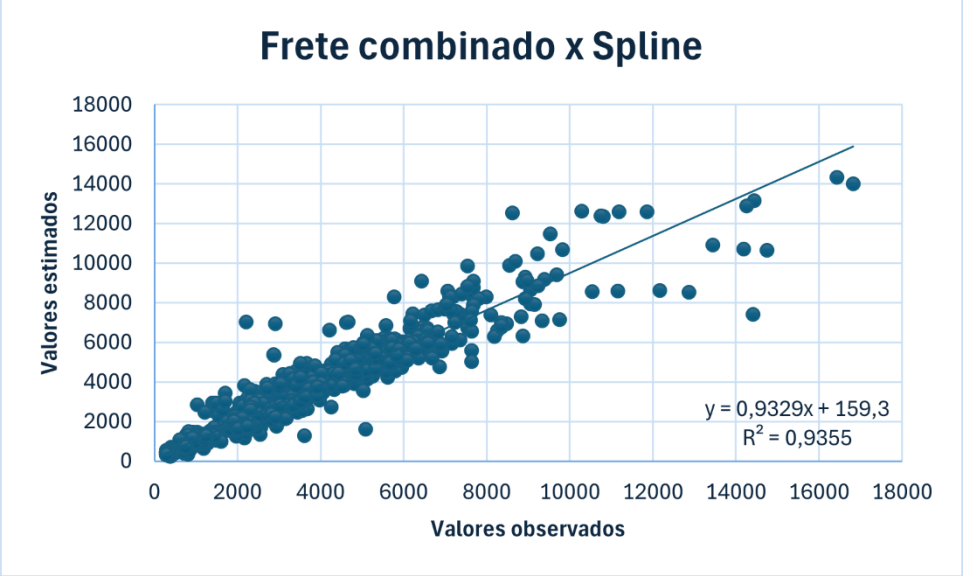
As Figuras 21 e 22, mostram, respectivamente, o desempenho dos métodos de Regressão e Spline, que obtiveram resultados satisfatórios, porém inferiores.

Figura 21 – Valores estimados pela Regressão Linear



Fonte: Autor (2024).

Figura 22 – Valores estimados pela Spline



Fonte: Autor (2024).

Através dos gráficos, pode-se identificar que quanto mais próximo à reta diagonal de 45º estiver localizado o ponto, mais próximo estarão um do outro os

valores estimados e observados. De mesma forma, quanto mais disperso um ponto em relação à reta, mais discrepante serão os valores observados e estimados, sendo assim há maior erro na previsão.

Os modelos tiveram performance positiva principalmente em casos em que há curta quilometragem a ser percorrida, que eram problemas nos métodos já existentes de precificação de fretes utilizados pela empresa.

4.2 COMPARATIVO COM OS MÉTODOS JÁ PRATICADOS

Com o melhor método preditivo definido, a próxima etapa do trabalho visa compará-lo aos métodos já utilizados pela empresa. O melhor resultado foi encontrado no Método de Floresta e de tal forma, esse será examinado em conjunto ao método de tabelas mínimas da ANTT e ao método Bialog.

Para o comparativo, precisa-se simular os valores de frete obtidos para cada um dos 2.335 fretes exportados como validação no Método de Floresta. Para isso, utilizou-se o Excel, parametrizando a tabela de fretes mínimos para carga lotação disponibilizada pela ANTT e a tabela de fretes da Bialog.

Um ponto importante a ser ressaltado, é de que os fretes a serem comparados foram escolhidos aleatoriamente pela ferramenta Criar Amostras no Alteryx Designer, dessa forma, tem-se fretes em diferentes datas. Como foi realizada a correção dos valores monetários através do IPCA, utilizamos também a tabela mais atual da ANTT.

O arquivo exportado do software traz a coluna AH nomeada como FLORESTA, onde o valor previsto pelo modelo criado é inserido. Semelhantemente, foi criada uma coluna AI nomeada ANTT onde estará presente o valor previsto através da tabela mínima de fretes da ANTT e a coluna AJ nomeada BIALOG onde estará presente o valor obtido através do método de precificação própria da empresa.

A Figura 23 mostra a comparação entre os valores obtidos por cada método.

Figura 23 – Recorte de resultados previstos por cada método

FRETE.COMBINAD	ÁRVORE	ANTT	BIALOG	ERRO ARVORE	ERRO ANTT	ERRO BIALOG	MELHOR MÉTOD
3914,82	3917,56	3915,65	2620,04	0,07	0,02	33,07	ANTT
3249,67	3222,77	4520,94	3250,07	0,83	39,12	0,01	BIALOG
3173,50	3109,96	3174,46	1761,72	2,00	0,03	44,49	ANTT
5720,96	5706,73	5737,17	5087,16	0,25	0,28	11,08	ÁRVORE
3817,35	3814,18	3806,45	2705,56	0,08	0,29	29,12	ÁRVORE
3868,43	3871,04	3880,70	2598,64	0,07	0,32	32,82	ÁRVORE
3852,25	3852,61	3832,65	2625,57	0,01	0,51	31,84	ÁRVORE
3852,25	3849,07	3832,65	2625,57	0,08	0,51	31,84	ÁRVORE
3826,13	3825,36	3845,76	2497,43	0,02	0,51	34,73	ÁRVORE
3071,13	3390,56	3073,15	2299,03	10,40	0,07	25,14	ANTT
3826,13	3816,64	3845,76	2497,43	0,25	0,51	34,73	ÁRVORE
5328,65	5366,87	5324,52	4259,08	0,72	0,08	20,07	ANTT
5742,15	5717,07	5773,58	5122,65	0,44	0,55	10,79	ÁRVORE
3921,08	3926,77	3898,17	2519,24	0,14	0,58	35,75	ÁRVORE
2786,30	2803,14	2804,05	1705,15	0,60	0,64	38,80	ÁRVORE
1910,27	1915,07	1923,09	930,03	0,25	0,67	51,31	ÁRVORE
3852,25	3849,79	3880,70	2663,17	0,06	0,74	30,87	ÁRVORE

Fonte: Autor (2024).

Para comparação em paralelo dos resultados, é utilizado o erro relativo como medida, verificando o erro para cada frete através de cada método em relação ao valor real. Para definir o melhor método então, é escolhido o que apresenta menor erro.

Dos 2.335 fretes, apenas em 5 casos o cálculo comparativo não pôde ser concluído, pois os valores dos métodos ANTT e Bialog resultaram em erro. Analisando esses dados, foi identificada que há uma informação incorreta inserida, que é a quantidade de eixos com o valor 0 atribuído. Esse erro vem do cadastro de veículos nos sistemas da empresa e impede o cálculo comparativo, sendo assim foram descartados da análise. A Tabela 2 a seguir dispõe os valores encontrados para cada indicador de desempenho dos métodos.

Tabela 2 - Comparativo entre método preferido e métodos já praticados

Método	Número de fretes	Percentual	R^2	α	β	RMSE
Método de Floresta	1949	83,65	0,9602	134,42	0,9458	393,0512
Método ANTT	290	12,45	0,9018	-207,16	1,0594	735,1231
Método Bialog	91	3,91	0,8795	-565,25	0,9148	1050,6751

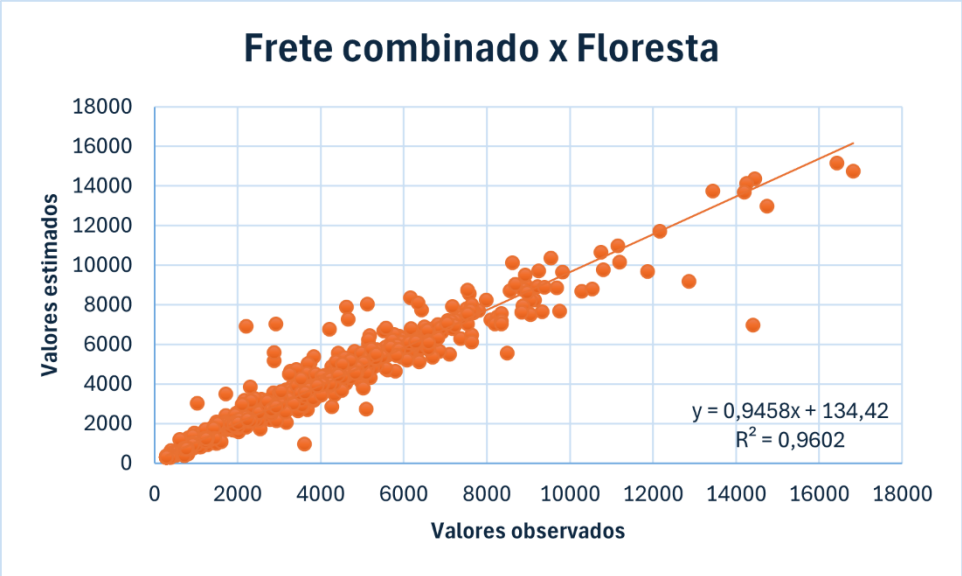
Fonte: Autor (2024).

O Método de Floresta calibrado e validado no trabalho foi o mais aproximado dos valores observados em 83,65% dos casos através da avaliação por erro relativo, isto é, das 2.330 avaliações, teve melhor desempenho em 1.949. O segundo melhor método foi o da ANTT, que foi o mais adequado em 290 ocasiões, representando 12,45%.

Em apenas 3,91% dos casos o método que a empresa possuía se saiu melhor, uma quantidade significativamente menor que os demais. Isso não necessariamente implica que o método da empresa seja totalmente ruim, apenas indica de que ele não é o melhor na comparação, mas pode estar dentro da taxa de assertividade que satisfazia a empresa anteriormente.

Através da análise de R^2, intercepto α, inclinação β e RMSE presentes na Tabela 2, o Método de Árvore se sobressai novamente aos demais. A Figura 24 representa os resultados previstos pelo método.

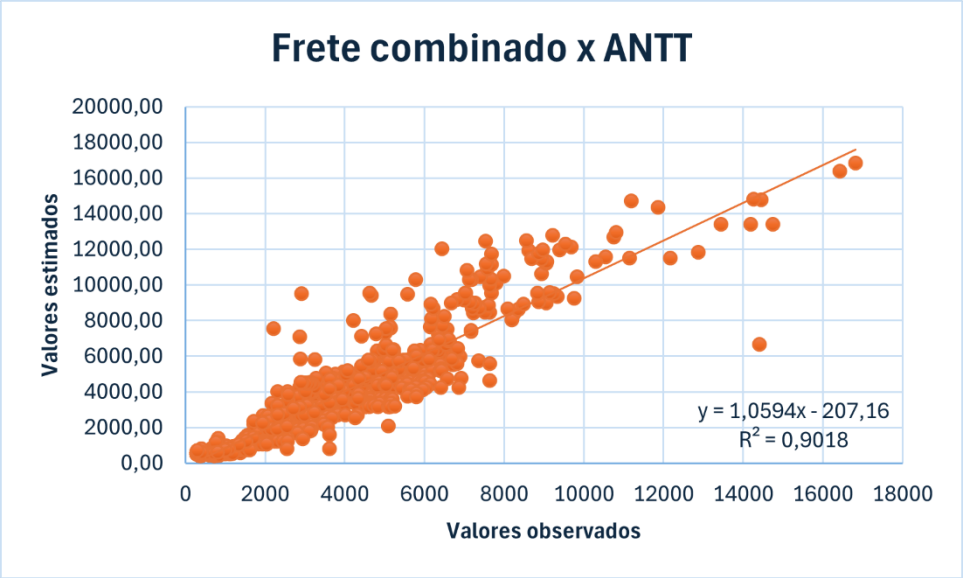
Figura 24 – Frete combinado versus Método de Floresta



Fonte: Autor (2024).

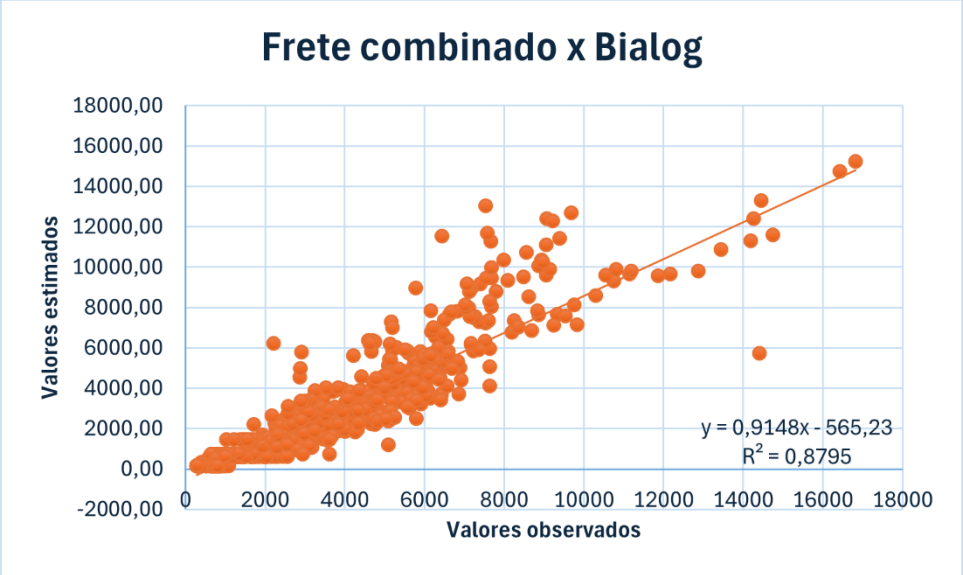
As Figuras 25 e 26, ilustram os valores previstas pelos métodos ANTT e Bialog, respectivamente. A dispersão dos dados em comparação à reta diagonal é maior, o que indica que os erros observados são maiores.

Figura 25 – Frete combinado versus ANTT



Fonte: Autor (2024).

Figura 26 – Frete combinado versus Bialog



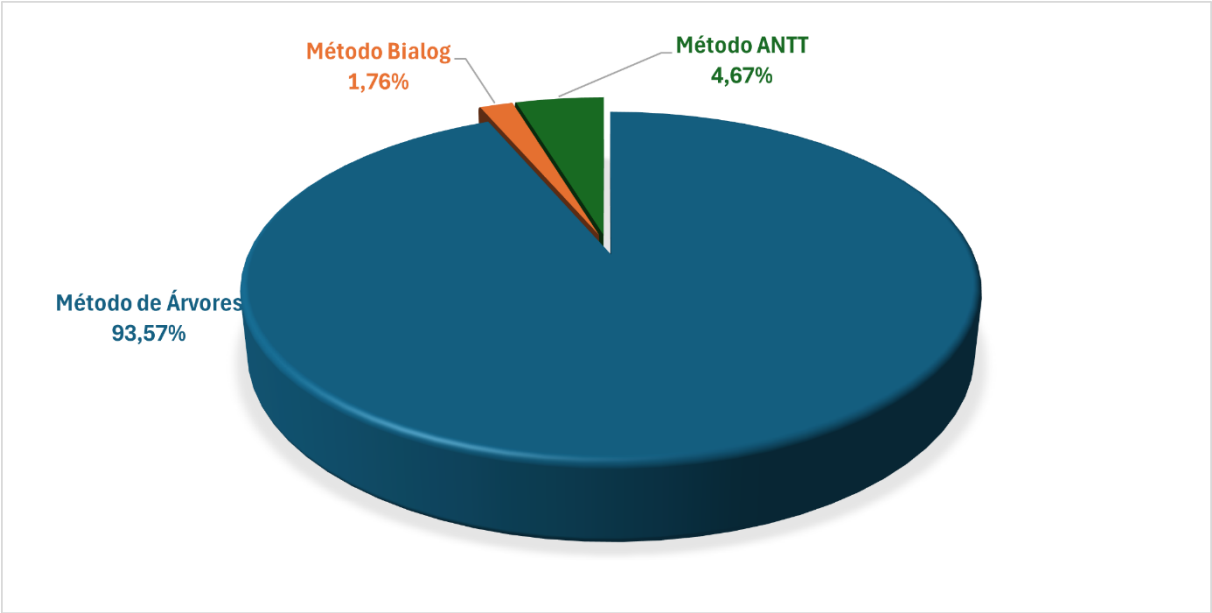
Fonte: Autor (2024).

Para uma análise um pouco mais detalhada, deve-se olhar os menores e maiores erros observados em cada método. Por exemplo, para o erro do Método de Floresta, tem-se predições exatas do valor que resultam em um erro de 0%, no entanto, se observados os métodos ANTT e Bialog, eles podem representar um valor distante do real, que resultam em erros grandes ou até mesmo um valor previsto próximo com pequenos erros.

Exemplificando um frete em que o valor combinado foi de R$ 3.921,13, pelo método desenvolvido o valor previsto foi praticamente o mesmo, enquanto o da ANTT foi de R$ 3.732,19 e o da Bialog foi de R$ 2.448,31. Os erros dos métodos são respectivamente de 0,1%, 4,82% e 37,56%, o que nos mostra que o método desenvolvido é melhor, mas não que necessariamente o método ANTT teve um desempenho ruim.

As dificuldades principais dos métodos já existentes encontram-se nos fretes onde há quilometragem baixa. Esses casos são limitados a valores mínimos de diária que não aderem ao mercado ou um valor mínimo muito pequeno e que não satisfaz a demanda dos transportadores. Com o novo método desenvolvido, os resultados são satisfatórios para essas ocasiões, conforme disposto na Figura 27.

Figura 27 – Métodos aplicados nas rotas curtas

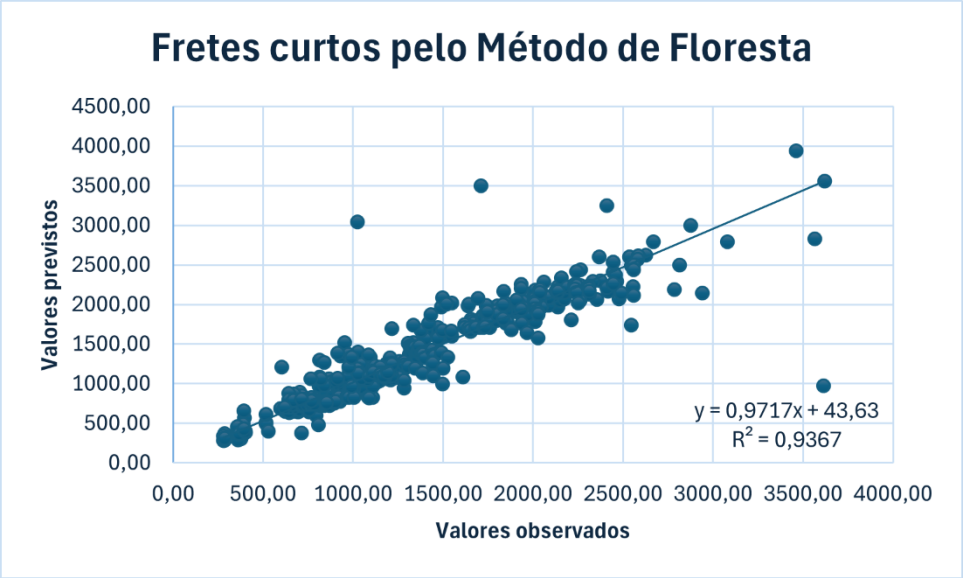


Fonte: Autor (2024).

Dos fretes validados, tem-se que 1.306 têm quilometragem inferior a 300 e são considerados fretes curtos. Desses, o melhor método foi o Método de Floresta desenvolvido em 93,56% dos casos. Dessa forma, há um aproveitamento grande para fretes com tais características.

O desempenho do método em fretes curtos é representado na Figura 28.

Figura 28 – Dispersão de valores previstos em rotas curtas



Fonte: Autor (2024).

O R^2 tem desempenho pouco inferior quando comparado com a regressão que inclui todos os fretes, no entanto, observa-se um melhor valor de inclinação de reta β de 0,9717, muito próximo de 1. O intercepto α também tem valor satisfatório de 43,63, se destacando pela proximidade a 0.

5. CONCLUSÃO

Precificar os fretes corretamente é uma tarefa extremamente importante e difícil. A principal parcela do custo de um frete é o valor a ser gasto entre pagamento do motorista e dos custos atrelados à viagem. Quando se trabalha com a contratação de TACs, alguns custos são repassados, porém ainda compõem a forma como o preço do frete será composto. Precificar com eficiência garante o melhor funcionamento da operação e o melhor acordo comercial com o cliente final.

Para alcançar o objetivo de criar um modelo de precificação eficiente, foi proposta a criação de um módulo em Alteryx Designer para tratamento de dados, calibração e validação dos modelos.

Visando obter e tratar dados, foi necessária a limpeza e união de diferentes entradas, garantindo a qualidade dos modelos a serem testados a partir dela. Para isso, as diferentes entradas de dados foram unidas formando um fluxo concentrado, onde 33 variáveis candidatas foram definidas.

O processo de calibração dos modelos foi a etapa seguinte do estudo, onde foram criados três modelos através do software Alteryx Designer, sendo os modelos de Regressão Linear, Método Spline e Método de Floresta. Todos os modelos tiveram resultados satisfatórios nos indicadores observados, provenientes da calibração bem ajustada.

As variáveis explicativas foram definidas como a quilometragem total do frete, o valor de impostos, o total das notas fiscais, a quantidade de volumes, o peso e número de eixos dos veículos, além de informações como mês e ano, juntamente ao valor médio de diesel praticado à época.

Na etapa de validação, os modelos testados tiveram desempenho satisfatório no parâmetro R^2, α, β e RMSE, porém houve uma leve vantagem do Método de Floresta sobre os demais, sem grande impacto em custo computacional ou de tempo. Dessa forma, esse método foi o preferido no comparativo.

O modelo preferido foi comparado aos métodos já conhecidos e utilizados pela empresa, que são o método de tabelas mínimas de frete da ANTT e o que foi denominado de método Bialog. Para analisar o método de melhor resultado, foi necessário aplicá-los na mesma base de validação dos métodos preditivos e em sequência, avaliar os resultados através da comparação de erros relativos entre o real

e o previsto. Seguindo a mesma linha de avaliação anteriormente exposta, os métodos foram também submetidos à comparação por R^2, α, β e RMSE.

O Método de Floresta modelado no trabalho se saiu superior novamente, sendo mais assertivo em 83,47% das observações, seguido pelo método da ANTT e Bialog, com 12,42% e 3,90%, respectivamente.

Uma segunda análise foi feita em um recorte dos fretes considerados de baixa quilometragem, onde os métodos utilizados pela empresa já não possuíam bom desempenho. Nesse cenário, o método desenvolvido se sobressai aos demais em 93,56% das observações.

De tal forma, o objetivo geral do trabalho é atingido em sua totalidade, pois foi estimado um modelo acurado de previsão e valores de fretes, que tem taxa de assertividade maior que os modelos já praticados.

As análises realizadas sobre o preço em fretes já realizados foram satisfatórias e são de grande valor financeiro e de tempo ao serem aplicados, mas não garantem que os valores previstos serão sempre alcançados. Como sugestão para trabalhos futuros, sugere-se desenvolver um método que seja baseado no preço solicitado pelos TACs e não no preço final combinado, a fim de avaliar as diferenças entre o desejado e o acordado.

A coleta dos dados e união em base concisa é a parte de maior custo de execução, porém o processo pode se tornar mais rápido e eficiente se integrado diretamente à base de dados da empresa, através de uma API (Application Programming Interface) que faça uma consulta SQL (Structured Query Language).

Por fim, sugere-se que o método seja aplicado ao sistema da Bialog para testar sua eficiência real e ser retroalimentado, para que o aprendizado seja constante e aperfeiçoado. Para resultados ainda mais satisfatórios, uma base externa de dados também pode aumentar a confiança do modelo, agregando dados e cenários diferentes dos aplicados na empresa.

REFERÊNCIAS

ALTERYX INC., **Ferramentas do Designer**. Alteryx Help. Disponível em: https://help.alteryx.com/current/pt/designer/tools.html. Acesso em: 23 jun. 2024.

ALTERYX INC.. **Alteryx:** Soluções para análise de dados. Disponível em: https://www.alteryx.com/. Acesso em: 21 maio 2023.

BALLOU, R. H. **Gerenciamento da cadeia de suprimentos/logística empresarial**. 5ª ed. Porto Alegre: Booksman, 2006

BARBETTA, P. A., REIS, M. M. & BORNIA, A. C. **Estatística para Cursos de Engenharia e Informática.** 2a edição, Editora Atlas, São Paulo, 2009. ISBN: 9788522449897.

BIAU, G. SCORNET, E. **A random forest guided tour.** An Official Journal of the Spanish Society of Statistics and Operations Research. 2016.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Springer Science+Business Media, LLC, 2006. ISBN-10: 0-387-31073-8.

BRASIL. Agência Nacional de Transportes Terrestres. **Como calcular o piso mínimo.** 2023. Disponível em: https://antt-hml.antt.gov.br/como-calcular-o-piso-minimo. Acesso em: 29 nov. 2023.

BRASIL. Agência Nacional de Transportes Terrestres. **Conheça a ANTT**. Disponível em: https://www.gov.br/antt/pt-br/acesso-a-informacao/institucional/conheca-a-antt. Acesso em: 14 jun. 2024.

BRASIL. Agência Nacional de Transportes Terrestres. **Lei nº 10.233, de 5 de junho de 2001.** Dispõe sobre a reestruturação dos transportes aquaviário e terrestre, cria o Conselho Nacional de Integração de Políticas de Transporte, a Agência Nacional de Transportes Terrestres, a Agência Nacional de Transportes Aquaviários e o Departamento Nacional de Infra-Estrutura de Transportes, e dá outras providências. Diário da União, 06 jun. 2001. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10233.htm#:~:text=Disp%C3%B5e%20sobre%20a%20reestrutura%C3%A7%C3%A3o%20dos,Transportes%2C%20e%20d%C3%A1%20outras%20provid%C3%AAncias. Acesso em: 14 jun. 2024.

BRASIL. Agência Nacional de Transportes Terrestres. **Política Nacional de Pisos Mínimos de Frete.** 2022. Disponível em: https://www.gov.br/antt/pt-br/assuntos/cargas/politica-nacional-de-pisos-minimos-de-frete. Acesso em: 23 nov. 2023.

BRASIL. Agência Nacional de Transportes Terrestres. **Resolução nº 5.976, de 7 de abril de 2022.** Altera a Resolução nº 5.867, de 14 de janeiro de 2020, em razão do disposto nos §§ 1º e 2º do art. 5º da Lei nº 13.703, de 8 de agosto de 2018. Disponível em: https://anttlegis.antt.gov.br/action/ActionDatalegis.php?acao=abrirTextoAto&link=S&ti

po=RES&numeroAto=00005976&seqAto=000&valorAno=2022&orgao=DG/ANTT/MI&cod_modulo=161&cod_menu=7796. Acesso em: 14 jun. 2024.

BRASIL. Agência Nacional de Transportes Terrestres. **Resolução nº 6.034, de 18 de janeiro de 2024.** Aprova o Regimento Interno da Agência Nacional de Transportes Terrestres. Disponível em: https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=19/01/2024&jornal=515&pagina=125. Acesso em: 14 jun. 2024.

BRASIL. **Lei nº 13.703 de 8 de agosto de 2018**. Institui a Política Nacional de Pisos Mínimos do Transporte Rodoviário de Cargas. Diário da União, 09 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13703.htm#:~:text=LEI%20N%C2%BA%2013.703%2C%20DE%208%20DE%20AGOSTO%20DE%202018.&text=Institui%20a%20Pol%C3%ADtica%20Nacional%20de,do%20Transporte%20Rodovi%C3%A1rio%20de%20Cargas. Acesso em: 14 jun. 2024.

BRASIL. Ministério dos Transportes. **PNL 2035.** 2023. Disponível em: https://www.gov.br/transportes/pt-br/assuntos/planejamento-integrado-de-transportes/copy_of_planejamento-de-transportes/pnl-2035. Acesso em: 23 nov. 2023.

BRASIL. Tribunal de Contas da União. **Transporte.** 2023. Disponível em: https://sites.tcu.gov.br/2025/transporte.html#:~:text=Segundo%20o%20Plano%20Nacional%20de,participam%20ambos%20com%20aproximadamente%2015%25. Acesso em: 23 nov. 2023.

BRUCE, Peter; BRUCE, Andrew. **Estatística prática para cientistas de dados:** 50 conceitos essenciais. 1ª ed. Rio de Janeiro — RJ: Alta Books, 2019.

CAMPOS, V. B. G. **Planejamento de Transportes:** Conceitos e Modelos. 1ª ed. Rio de Janeiro: Interciência, 2013.

CHAI, C. P. **The importance of data cleaning: Three visualization examples**. CHANCE, Taylor & Francis, v. 33, n. 1, p. 4–9, 2020.

CUNHA, M. C. C. **Métodos Numéricos.** 2ª ed. Campinas: Editora da UNICAMP, 2003

DEVORE, J. L. **Probabilidade e Estatísica para Engenharia e Ciências**. 6 ed. São Paulo: Cengage Learning, 2006.

FUNDAÇÃO R. **Sobre o R.** 2024. Disponível em: https://www.r-project.org/about.html. Acesso em: 20 jun. 2024.

GOMES et al. **Funções Splines aplicadas em dados de crescimento.** Colloquium Agrariae, vol. 13, n. Especial 2, Jan–Jun, 2017, p. 222-234. ISSN: 1809-8215. DOI: 10.5747/ca.2017.v13.nesp2.000229

LOPES, S. B. **Efeitos da dependência espacial em modelos de previsão de demanda por transporte**, 2005. Dissertação (Mestrado em Engenharia Civil) – Escola de Engenharia de São Carlos, Universidade de São Paulo, 2005.

LOPES, S. B. **Uma ferramenta para planejamento da mobilidade sustentável com base em modelo de uso de solo e transportes**, 2010. Tese (Doutorado em Ciências) – Escola de Engenharia de São Carlos, Universidade de São Paulo, 2010.

MICROSOFT. **Microsoft Excel.** Disponível em: https://www.microsoft.com/ptbr/microsoft-365/excel. Acesso em: 23 jun. 2024.

MONTGOMERY, D. C. & RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 4ª ed. Rio de Janeiro: LTC, 2009. ISBN: 9788521616641.

PESQUISA CNT DE PERFIL DOS CAMINHONEIROS 2019. **Confederação Nacional dos Transportes.** Disponível em: https://www.cnt.org.br/perfil-dos-caminhoneiros. Acesso em: 23 nov. 2023.

PESQUISA CNT: Perfil Empresarial 2021: Transporte de Cargas. **Confederação Nacional dos Transportes.** Disponível em: https://static.poder360.com.br/2022/04/pesquisa-cnt-perfil-empresarial-transporte-rodovia%CC%81rio-de-cargas.pdf. Acesso em: 23 nov. 2023.

POLYZOTIS, N. et al. **Data lifecycle challenges in production machine learning:** a survey. ACM SIGMOD Record, ACM New York, NY, USA, v. 47, n. 2, p. 17–28, 2018.

SILVEIRA, M. C. Desenvolvimento de um módulo em Alteryx Designer com base em dados de modelo de macrossimulação Emme para análise de alternativas em vias australianas. 2023. Trabalho de Conclusão de Curso – Universidade Federal de Santa Catarina.

SIMEONE, O. **A Brief Introduction to Machine Learning for Engineers.** 2018. Department of Informatics King's College London. 2018.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Practical machine learning tools and techniques**. Morgan Kaufmann, p. 578, 2005.