Thales César Giriboni de Mello e Silva

# *Big Data* aplicado à tarefa de categorização de níveis de severidade de vias férreas a partir de dados de veículos sensoriais

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Universidade de São Paulo

Escola Politécnica

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Brasil

2023

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Silva, Thales Cesar Giriboni de Mello
Big Data aplicado à tarefa de categorização de níveis de severidade de vias férreas a partir de dados de veículos sensoriais / T. C. G. M. Silva -- versão corr. -- São Paulo, 2023.
95 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Big Data 2.Aprendizado de máquina 3.Transporte ferroviário 4.Vias permanentes 5.Modelagem de dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

# Agradecimentos

Os agradecimentos principais são direcionados aos membros da Cátedra Underrail Angelo Samuel Junqueira, Jeaneth Machicao, Osvaldo Gogliano Sobrinho, Pedro Luiz Pizzigatti Corrêa, Rosângela Motta, Liedi L. B. Bernucci, Fernando Cossetin, Alan James Peixoto Calheiros, Wellinghton Dias Queiroz. Agradecimentos especiais aos Engenheiros Luciano Oliveira, Luciano Cassaro e Vitor Ohnesorge.

# Resumo

A manutenção da condição de vias férreas é de fundamental importância para esse setor de transporte, responsável por considerável porção do fluxo de exportação da economia brasileira. Entretanto, eventuais falhas e defeitos acarretam interrupções da via, o que pode afetar negativamente o seu tráfego. Assim, convém utilizar o estado da arte de engenharia de computação para, a partir de dados coletados por diversos sensores acoplados a Vagões Instrumentados e Carros Controle, estimar a deterioração dos componentes de uma determinada ferrovia. Para tanto, técnicas de *Big Data* e Ciência dos Dados são utilizadas para captar, modelar e armazenar esses registros, permitindo o emprego de técnicas de aprendizado de máquina supervisionado e não-supervisionado para a definição de níveis de severidade da ferrovia, isto é, a quantificação do estado dos componentes da via permanente em níveis discretos de operacionalidade, onde a maior severidade indica falha iminente e a menor, condições normais de operação. Esses resultados são apresentados em uma plataforma web capaz de ser acessada remotamente, possibilitando o acompanhamento *in loco* das predições dos modelos computacionais.

**Palavras-chaves**: Manutenção preditiva. Big Data Analytics. Ferrovias. Aprendizado de máquina.

# Abstract

Railroad maintenance is of critical importance for that transportation sector, which is responsible for a significant portion of Brazilian's exports. As such, occasional failures and defects can result in flow interruptions, which may negatively impact the output of the system. Therefore, it is desirable to use the state-of-the-art computer engineering's methods for developing a predictive algorithm for rail components deterioration estimation, using as input a variety of sensorial data gathered from instrumented ore wagons and track geometry control cars. For that goal, Big Data techniques are employed to capture, model and store these readings, paving the way for supervised and unsupervised machine learning algorithms to be used to train a model able to define a quantification of the rail component state into discrete operational levels, or severity levels for the railroad's condition, where the greatest severity would indicate an imminent geometry defect, and the lowest would point to normal conditions. These results are then presented through a web application that can be remotely accessed, allowing for the *in loco* follow-up of the computational models' predictions.

**Key-words**: Predictive maintenance. Big Data Analytics. Railtracks. Machine learning.

# Sumário

# 1 Introdução

A Estrada de Ferro Vitória-Minas (EFVM) é um importante ativo para a economia brasileira. Como tal, eventuais disrupções ao escoamento dos trens podem acarretar em grandes prejuízos. Dessa forma, a constante monitoração e busca por otimizações nas tarefas de manutenção são fundamentais para manter a competitividade do setor ferroviário.

Esse setor é regulamentado por várias normas de operacionalidade para as ferrovias. Uma série de medidas dos componentes da via permanente, como bitola e empeno, são utilizadas como parâmetro para o estabelecimento de limites de velocidade que os vagões devem obedecer. Esses limites variam em função de algumas variáveis, como, por exemplo, se a composição é utilizada para carga ou transporte de passageiros. As ferrovias também podem ser classificadas em classes de via, definidas pelas velocidades estipuladas e densidade de tráfego (RIBEIRO, 2017).

Nesse contexto de necessidade de intensa vigilância, equipamentos como o Carro Controle (CC) e Vagão Instrumentado (VI) são empregados em campo nos últimos anos (LINGAMANAIK et al., 2017). Com a maturidade da Internet das Coisas (IoT) (DAHLQVIST et al., 2019), fenômeno caracterizado pela interconexão entre objetos munidos de sensores e eletrônicos, é possível utilizar técnicas de Ciência dos Dados e *Big Data* para extrair, armazenar e modelar as leituras dos sensores desses veículos em sistemas computacionais. Com isso, torna-se viável a elaboração de painéis e centros de monitoramento, permitindo um acompanhamento das condições da via permanente em grande escala.

Para extrair conhecimento a partir desse volumoso conjunto de dados, a Ciência dos Dados é a disciplina cujos métodos e processos são empregados. Assim, com a integração das áreas de conhecimento de estatística, ciência da informação e engenharia da computação, cria-se um ciclo de trabalho, abrangendo as atividades de aquisição, análise e apresentação dos dados. Em relação ao processamento computacional de dados variados e volumosos, o termo *Big Data* também é utilizado para descrevê-lo.

Consequentemente, utilizando os modelos obtidos pelos processos de *Big Data*, torna-se adequado utilizar algoritmos de aprendizado de máquina para predizer as falhas na via, o que eventualmente anteciparia manutenções preventivas e evitaria visitas prematuras. Espera-se obter uma precisão adequada de previsões para que defeitos estruturais da via sejam rapidamente identificados e reparados antes que danos mais graves e custosos ocorram.

Para que os modelos preditivos possam efetivamente ser utilizados em campo,

é necessário visualizar esses resultados. Para tanto, o emprego de uma plataforma de visualização *web* é ideal, permitindo que os modelos sejam executados em servidores em computação em nuvem, e suas saídas sejam acessadas a partir de qualquer dispositivo com acesso à internet, como, por exemplo, um celular na ferrovia.

## 1.1 Justificativa

O transporte ferroviário brasileiro é de fundamental importância para a logística e o transporte de cargas para o país. Estima-se que esse setor de transportes contribui com cerca de 5% do PIB nacional (FALCÃO, 2013). A Estrada de Ferro Vitória-Minas (EFVM), por exemplo, foi responsável pelo transporte de mais de 110 milhões de toneladas em 2014, além de cerca de um milhão de passageiros entre os diversos municípios cobertos (RIBEIRO, 2017).

Nesse contexto, é de vital importância a determinação e detecção de defeitos e deteriorações da via permanente. Interrupções no tráfego de trens da ferrovia interferem negativamente o escoamento de produtos, o que pode provocar atrasos e perdas. Assim, a tarefa de viabilizar e sinalizar os necessários trabalhos de manutenção da via são extremamente relevantes (WESTON et al., 2015).

Tradicionalmente, empregam-se dois tipos de manutenção: corretiva e preventiva. A primeira é realizada após a constatação de uma avaria em algum ativo da via, como defeitos severos da via permanente. Já a manutenção preventiva é regida por normas e visa a se antecipar a falhas, utilizando, para isso, conceitos estatísticos de tempo médio entre falhas e tempo de vida útil de componentes. Entretanto, ambas as soluções apresentam desvantagens relevantes, como a interrupção do serviço no caso da corretiva e o comum descarte de componentes que ainda possuem operacionalidade aceitável, no caso da preventiva (FUMEO; ONETO; ANGUITA, 2015). Tem-se, então, um problema de otimização de esforços de manutenção.

Num cenário de crescente coleta de dados e maior capacidade computacional de armazená-los e processá-los, aliados à ascensão da chamada Indústria 4.0, em que se observa um significativo aumento de eficiência e produtividade das indústrias (SCHUMACHER; EROL; SIHN, 2016), convém estudar formas de predizer a ocorrência de um defeito de geometria da via permanente mediante dados coletados por Vagões Instrumentados (Figura 1) e Carros Controle (Figura 2). Através de dados coletados por esses veículos munidos dos mais diversos sensores digitais, é possível inferir o estado da via férrea para que se possa determinar níveis de severidade operacional da ferrovia, em que podem ser eventualmente utilizados para agendamento de visitas de inspeção aos locais em condição mais crítica.

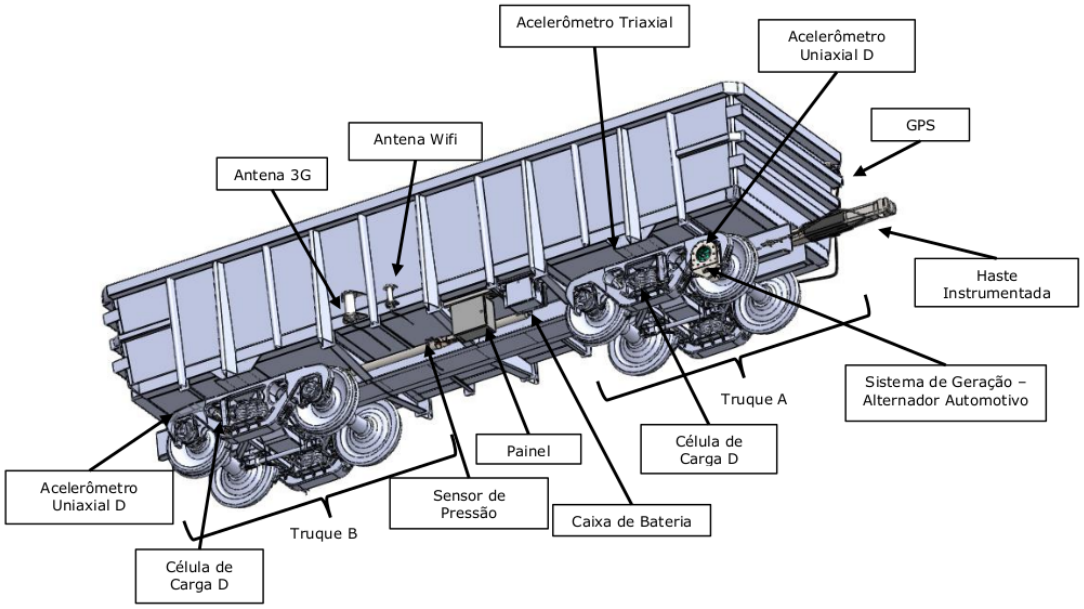


Figura 1 – Equipamentos de sensores de um Vagão Instrumentado (TUDEIA et al., 2019).

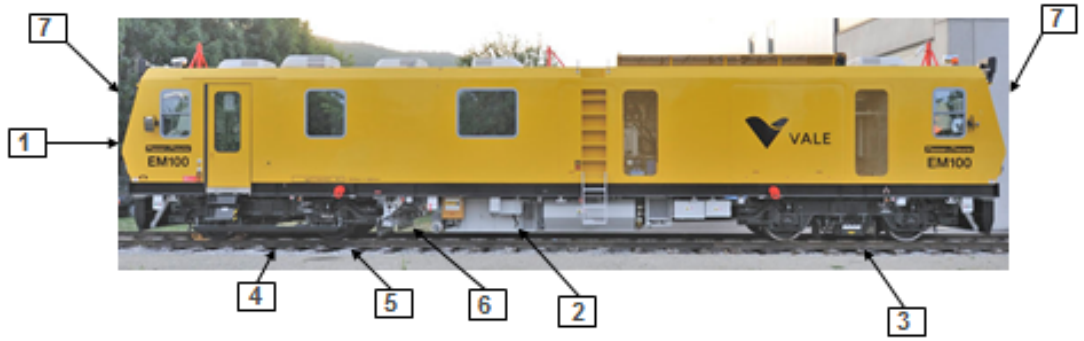


Figura 2 – Carro Controle modelo EM100 da Plasser, com indicações de seus sensores: 1) leitor de gabarito da via; 2) acelerômetros; 3) leitor de perfil do trilho e bitola; 4) leitor de bitola em 2 pontos; 5) unidade de medição inercial (leitor de geometria); 6) captura de imagem dos trilhos, dormentes e fixações; 7) captura de imagem panorâmicas da via (COSTA et al., 2016).

Esse tipo de trabalho já foi realizado com sucesso na literatura. Dados de aceleração dos eixos do vagão, por exemplo, alinhados com dados de localização de sistema de posicionamento global (GPS), foram utilizados para determinar trechos que requerem manutenção em ferrovias holandesas (NUNEZ et al., 2014). Também em vias holandesas foram realizados experimentos que conseguiram combinar análise de imagens e sensores de aceleração dos eixos para a detecção de defeitos e posterior dedução de fatores que contribuíram para o aparecimento da falha (JAMSHIDI et al., 2018).

Assim, existe ampla experiência acadêmica de detecção de defeitos e predição de manutenção de vias férreas (CARVALHO et al., 2019)(XU et al., 2019). Convém utilizar

uma maior diversidade de dados oriundos de diferentes veículos para melhorar a precisão desses modelos, bem como aplicar a cenários reais do Brasil.

## 1.2 Objetivos

Os objetivos do trabalho são divididos em um objetivo geral e dois específicos, a saber:

- **Geral**: Desenvolvimento de modelo computacional para predição de falhas na EFVM;
- **Específico 1**: Desenvolvimento de modelos de *Big Data* para representação dos dados coletados;
- **Específico 2**: Aprimoramento de plataforma de visualização de dados.

O desenvolvimento de um modelo computacional para a predição de falhas da EFVM, estabelecido com o objetivo geral, é possibilitado a partir de dados coletados por sensores instalados em Vagões Instrumentados e Carros Controle, bem como os registros de falhas e defeitos observados na via. Assim, a hipótese estabelecida é que é possível determinar os níveis de severidade para a via permanente a partir da análise desses dados.

Esses níveis de severidade são uma abstração em cima da ideia de discretização das condições operacionais da ferrovia. Inspirada pelos limites de operações que normas técnicas como a ABNT NBR 16387 definem (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2016), essa concepção envolve a premissa de que os componentes da via permanente degradam com o passar do tempo, partindo da situação de condições normais de operação, desenvolvendo deformações e defeitos gradativamente, acarretando em intervenções operacionais como restrição de velocidade, até atingir um estado em que a interdição da via e manutenção corretiva são necessárias. A princípio, cada nível de severidade a ser definido é bem característico e suficientemente distinto dos demais para sua correta identificação, então a quantidade exata de níveis a serem definidos não é conhecida *a priori*; ao contrário, faz parte do objetivo a definição desses níveis de forma a se obter um modelo computacional adequadamente preciso para utilizar esses indicadores para predizer a ocorrência de falhas na via.

As falhas e ocorrências a serem consideradas para a construção do modelo são referentes a defeitos geométricos observados em campo, tal como exemplificado pela Figura 3, e registrados pelas equipes de manutenção. Esses dados são, consequentemente, obtidos através dos relatórios gerados a partir dessas anotações, em acordo com a proposta de construção de um modelo que encontre as evidências dessas falhas a partir dos dados dos veículos.

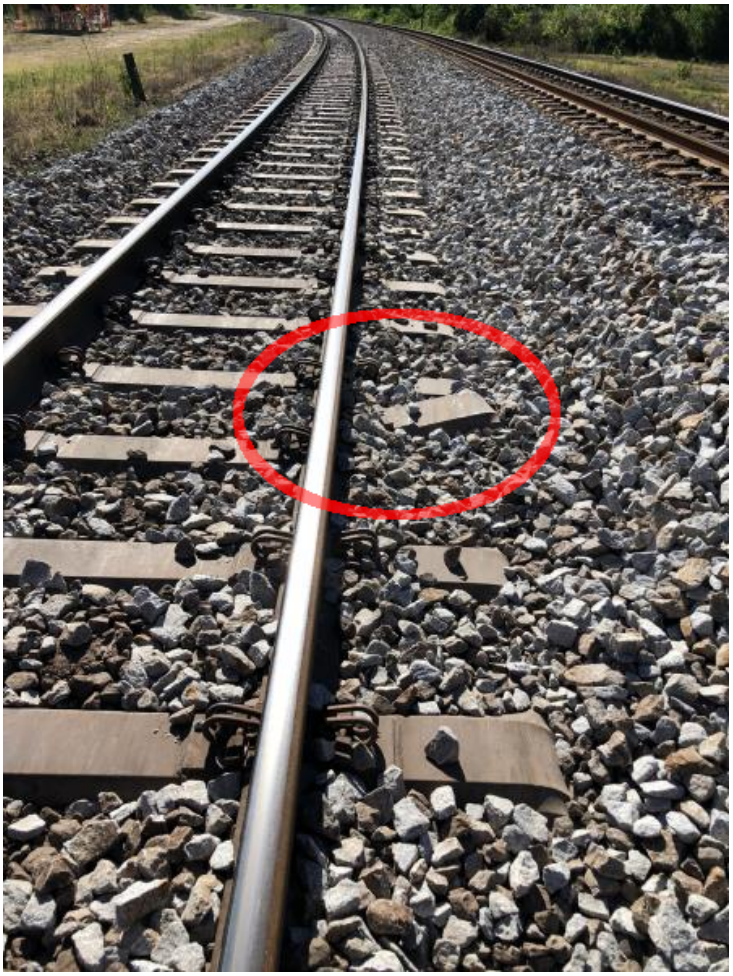

Figura 3 – Exemplo de falha na via permanente a ser considerado para o objetivo do trabalho (CASSARO; SILVA, 2021).

Os objetivos específicos do projeto são o desenvolvimento de modelos de *Big Data* para a representação e armazenamentos dos dados coletados, oriundos de diversas formas e formatos, e a disponibilização da visualização de dados, permitindo que os resultados dos modelos preditores possam ser acessados remotamente por eventuais equipes de manutenção e inspeção da via permanente.

### 1.2.1 Contribuições

Como consequência do cumprimento dos objetivos definidos, espera-se que o presente trabalho apresente contribuições relevantes para a manutenção de ferrovias. Uma primeira contribuição é a construção de modelos computacionais preditivos, agregando à literatura novas metodologias de construção e treinamento de classificadores com base nos dados de leituras de sensores veiculares e ocorrências em vias. Outro ponto importante é a definição de uma plataforma de *Big Data*, isto é, agregação dos dados em um modelo de armazenamento e consulta unificado, permitindo que novas análises sejam construídas para o conjunto de dados em questão, além de possibilitar a expansão com novas fontes de

dados. Por fim, também se espera que a ferramenta de visualização possa ser expandida para unificar as formas de exibição dos dados coletados, além de permitir acesso fácil aos níveis de severidade obtidos.

Assim, as contribuições esperadas variam desde adições ao estado-da-arte de predição de manutenção até aspectos bastante pontuais e importantes para operação da via.

### 1.2.2 Plano de trabalho

De forma geral, o plano de trabalho segue o método de ciclo de vida para execução dos experimentos de Ciência dos Dados e também o modelo proposto para gestão de dados baseado no ciclo de vida dos dados. O método envolve, numa etapa preliminar, agrupar os dados disponíveis que possam contribuir para o modelo preditivo de manutenção da via, gerados durante as viagens de Vagões Instrumentados, além das ocorrências de eventos e falhas registrados longo da utilização da via. Também foi planejado o desenvolvimento de modelos de aprendizado de máquina para predição de severidade de operacionalidade da via. Por fim, realiza-se uma apresentação de seus resultados utilizando a ferramenta de visualização desenvolvida e evoluída.

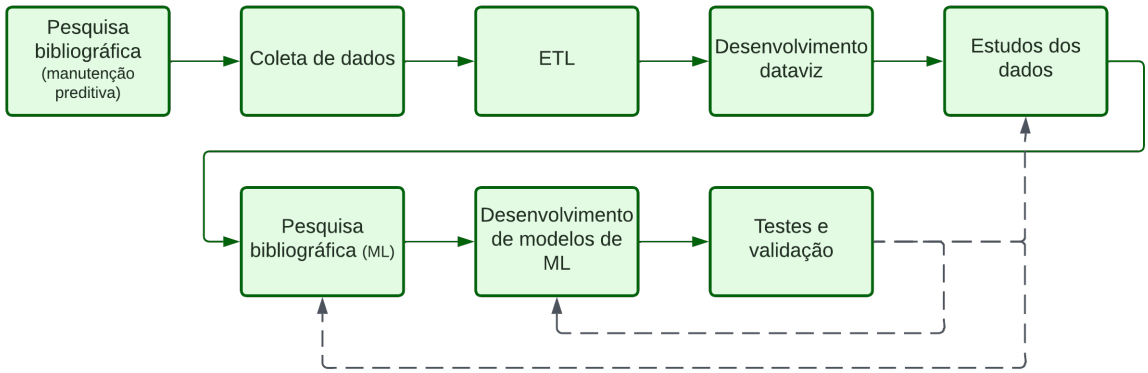


Figura 4 – Fluxograma simplificado do plano de trabalho, deixando nítida a característica circular de trabalho entre etapas de estabelecimento de hipóteses, experimentos, testes e novas hipóteses. Em particular, a etapa de testes e validação dos modelos preditivos desenvolvidos pode levar a novas rodadas de etapas anteriores, como exploração e estudo dos dados e revisão bibliográfica. Regressos a etapas anteriores são indicadas por setas acinzentadas.

A primeira etapa do projeto consiste na obtenção da maior quantidade possível de dados obtidos de Vagões Instrumentados e Carros Controle que trafegaram pela EFVM. Esses dados possibilitam que seja conduzido um amplo estudo a cerca das variáveis em questão, o comportamento delas ao longo da ferrovia e como podem se relacionar com

a ocorrência de defeitos. Assim, também é necessário a captação de dados de falhas observadas, a fim de possibilitar o relacionamento entre as ocorrências e as variações dos dados medidos pelos sensores. Essa é uma etapa contínua, uma vez que os dados em questão são disponibilizados em intervalos irregulares. Os resultados obtidos são apresentados na Seção 4.1.1.

A etapa seguinte à obtenção de dados é o pré-processamento e armazenamento dos dados. Seguindo as técnicas de *Big Data* de Extração-Transformação-Carregamento (do inglês ETL, *Extract-Transform-Load*)(SALIERNO et al., 2020), propõem-se algoritmos para o tratamento dos dados, oriundos nas mais diversas formas de representação utilizadas atualmente (planilhas, formatos abertos, etc). Todos os dados preparados são armazenados num banco de dados relacional, permitindo que as demais tarefas que consumam os dados não carreguem consigo a complexidade de lidar com dados em diferentes padrões. Os resultados dessa etapa são discutidos na Seção 4.1.

Na sequência, a interface da plataforma de visualização dos dados é atualizada com as funcionalidades necessárias para apoiar no entendimentos dos dados. Mapas e gráficos serão empregados para que se possa ter uma visualização (dataviz, *Data visualization*) adequada à natureza das leituras obtidas, bem como relações espaciais e temporais entre elas. Assim, para que seja possível essa incorporação, é necessário que a etapa anterior tenha disponibilizado os dados num formato único e acessível à plataforma. Esses resultados são detalhados na Seção 4.4.

A próxima etapa é realizar uma análise exploratória dos dados, de forma a procurar por padrões e entender o comportamento do conjunto de dados disponíveis. Nessa etapa, todas as variáveis são caracterizadas e são geradas análises mono-multivariáveis através de estatística descritiva. Os resultados dessa análises são apresentados na Seção 4.2.

As etapas seguintes envolvem o desenvolvimento de modelos de aprendizado de máquina (ML, *Machine Learning*) para procurar identificar dentre os algoritmos disponíveis aquele que tenha o melhor desempenho para extrair características que possam ser indicativos de avarias à via. Aqui, convém começar com os melhores algoritmos já validados na literatura e observar como eles se comportam com os dados em questão. A partir daí, espera-se que as deficiências de acurácia dos modelos possam apontar para as evoluções necessárias nos modelos para que se atinjam métricas de desempenho aceitáveis. Os resultados são apresentados na Seção 4.3.

Por fim, desenvolveu-se uma estratégia de validação entre todas as soluções de detecção e predição de manutenção criadas, detalhada na Seção 4.3.1.2, de forma a estabelecer com bastante certeza modelos adequados para a tarefa em questão. Como consequência do trabalho, criou-se um modelo computacional de aprendizado de máquina para a avaliação da condição da via e sua iminência de falha, bem com a possibilidade do emprego desse modelo em campo através da visualização dos seus resultados numa

ferramenta de visualização. Esta pesquisa desenvolveu um sistema computacional *web*, que pode ser acessado a partir de qualquer dispositivo com acesso à internet.

| Etapa | Prazo previsto |
|---|---|
| Pesquisa bibliográfica | Dezembro de 2021 |
| Coleta de dados | Contínuo |
| ETL | Janeiro de 2022 |
| Desenvolvimento dataviz | Março de 2022 |
| Estudos dos dados | Julho de 2022 |
| Desenvolvimento de modelos de ML | Dezembro de 2022 |
| Testes e validação | Dezembro de 2022 |

Tabela 1 – Cronograma estipulado para execução do projeto

# 2 Revisão bibliográfica

Dada a necessidade do emprego de métodos de Ciência dos Dados e de algoritmos de aprendizado de máquina, convém revisar o estado da arte desses tópicos. Também é fundamental levantar as experiências recentes com manutenção de ferrovias, incluindo uma visão geral sobre normas aplicadas, instrumentos utilizados e práticas consagradas.

## 2.1 Experimentos em Ciência dos Dados

A gestão de dados é um processo definido na disciplina de Ciência dos Dados que visa a coletar, armazenar e proteger os dados, além de garantir sua adequada utilização para agregar valor à base de conhecimento em questão. Para isso, introduz-se o conceito de ciclo de vida dos dados, caracterizado por uma série de etapas a serem planejadas e realizadas para o correto gerenciamento dos dados (DAMA INTERNATIONAL, 2017).

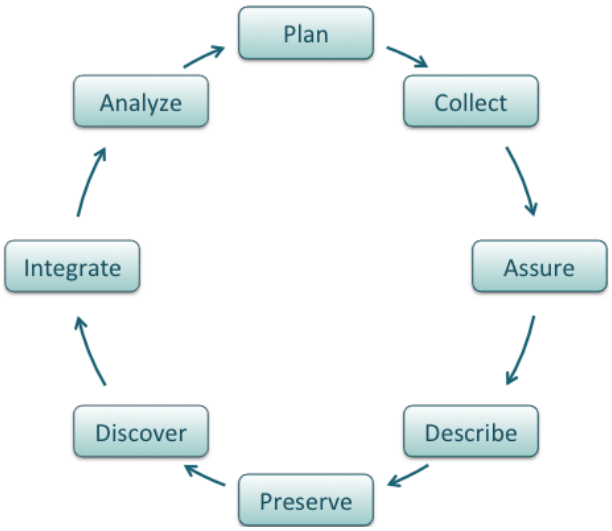


Figura 5 – Ciclo de vida dos dados aplicado em Ciência dos Dados (DATAONE, 2016).

O primeiro estágio do ciclo de vida é o planejamento. Neste momento, os objetivos do projeto são estabelecidos, a natureza dos dados é elaborada, a forma de armazenamento e organização é projetada e a maneira que os dados serão descritos (metadados) é desenvolvida. É neste momento também que os requisitos de compartilhamento e preservação dos dados são estruturados.

A etapa seguinte ao planejamento é a coleta, caracterizada pela determinação e execução das formas como os dados são coletados. Inconsistências das entradas são planejadas e minimizadas através das ferramentas utilizadas para a coleta, os formatos

a serem empregados são discutidos e implementados, e técnicas de integração dos dados são levantadas. Para atingir esses objetivos, uma das ferramentas mais empregadas é um banco de dados relacional, um sistema computadorizado no qual são armazenados dados de forma estruturada, permitindo aos seus usuários acessá-los e atualizá-los sempre que necessário (MATSUMOTO, 2006).

Na sequência, os dados passam por um controle e garantia de qualidade. São estabelecidos processos para revisão dos dados agregados, em que procuram-se por eventuais problemas na entrada das informações, bem como validações do domínio em questão. São analisados também erros de omissão, isto é, a ausência de dados esperados, ou sua incompletude.

As etapas de descrição, preservação e descoberta costumam ser discutidas em conjunto com o compartilhamento de dados. Dado o interesse em fornecer acesso aos dados ao público, comunidade científica e pesquisadores envolvidos, é de vital importância determinar a forma como os dados serão descritos, como eles serão armazenados no longo prazo, e de que forma os interessados poderão acessá-los no futuro.

Durante o estágio de integração, os metadados são revistos e eventualmente adaptados para garantir compatibilidade com novos contextos e métodos, documentando os pontos relevantes das etapas anteriores. Outro ponto executado nessa etapa é a documentação de todas as premissas implícitas assumidas para o conjunto de dados, e como isto poderia afetar um novo conjunto. Fluxos de trabalho reprodutíveis são alvo dessa etapa, a fim de garantir que todo o processo realizado possa ser replicado por terceiros.

Finalmente, com os dados armazenados, modelados e documentados, segue-se para a análise, na qual modelos estatísticos, gráficos, estimativas e demais métodos são empregados para extrair o conhecimento de interesse. Métodos como redução de dimensões, transformações de unidades, regressões lineares são técnicas comuns desta etapa (DATAONE, 2016).

## 2.2 Aprendizado de máquina aplicado à manutenção preditiva

Devido à relevância do problema de se definir com exatidão a melhor janela para se realizar manutenção de equipamentos e infraestrutura com o objetivo da redução de custos, observa-se o crescente interesse em soluções de manutenção preditiva através de aprendizado de máquina. Nos últimos dez anos, o número de publicações sobre o assunto aumentou drasticamente (CARVALHO et al., 2019). Além da significativa importância do assunto, a maior difusão de meios de coleta massiva de dados, possibilitada em parte pela disseminação da Internet das Coisas, aliada à capacidade de transmissão desses dados (KONDAKA et al., 2022), também contribui para a maior viabilidade dessa estratégia de identificar a necessidade de realizar manutenção (XU et al., 2019) (LIAO et al., 2022).

### 2.2.1 Técnicas de aprendizado de máquina utilizadas

Entre as técnicas de aprendizado de máquina mais utilizadas para a construção de modelos capazes de predizer a melhor janela de manutenção estão *random forests*, redes neurais artificiais, *Support Vector Machines*, modelos cinza, *fuzzy systems* e *K-means* (CARVALHO et al., 2019)(XU et al., 2019)(LIAO et al., 2022).

#### 2.2.1.1 Random forests

*Random forests* são, como o nome indica, um conjunto de várias árvores de decisão aleatorizadas, formando uma espécie de "floresta", onde o resultado final do modelo é obtido através de uma simples média dos resultados das árvores individuais. Por serem construídas a partir de árvores de decisão, tratam-se de um algoritmo de aprendizado supervisionado (BREIMAN, 2001). Uma das principais vantagens desse algoritmo é a usual boa performance quando envolve um grande número de variáveis, além de procurar evitar *overfitting* (sobre-ajuste, fenômeno observado quando o modelo se adapta em demasio aos dados de treino e se mostra incapaz de generalizar para outros dados) quando comparado a árvores de decisão tradicionais (CUTLER; CUTLER; STEVENS, 2012).

*Random forests* foram utilizadas em artigos para predição de reparos a vários componentes de veículos comerciais, construção de modelos preditivos para turbinas eólicas, predição de falhas de discos rígidos, detecção de curtos-circuitos em motores de indução e até detecção de problemas em sistemas de refrigeração e armazenamento a frio. Além de todas essa variedade de aplicações, *random forests* também são utilizadas em diversos outros trabalhos como modelos de aprendizado de máquinas utilizados como comparação para as soluções apresentadas nos projetos (CARVALHO et al., 2019).

#### 2.2.1.2 Redes neurais artificiais

Redes neurais artificiais são uma série de técnicas computacionais inspiradas em redes neurais biológicas em que uma rede de unidades de processamento (normalmente chamadas de nós ou neurônios) é formada. Cada unidade de processamento é associada a um peso, e é conectada a outras em suas entradas e saídas (SKANSI, 2018). A Figura 6 ilustra essa composição.

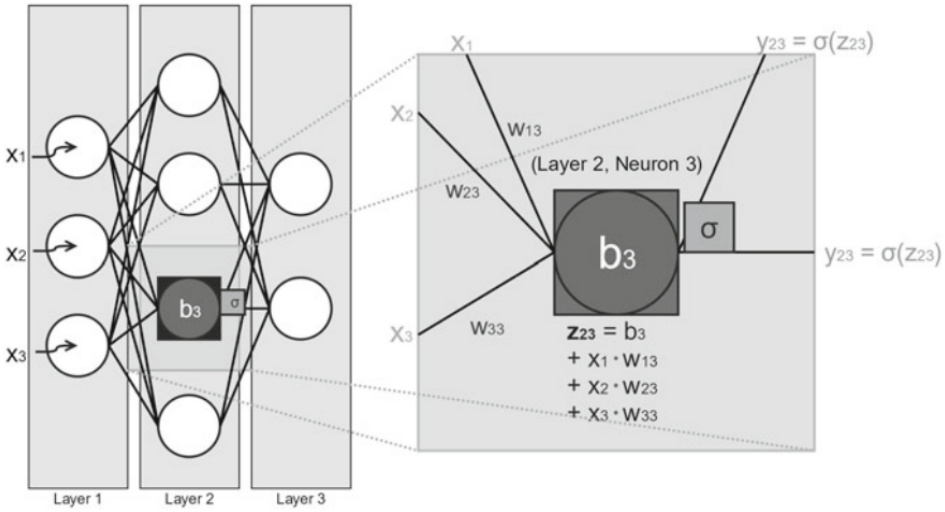


Figura 6 – Redes neurais artificiais são formadas por camadas, apresentada à esquerda, compostas por unidades de processamento, ilustrada à esquerda (SKANSI, 2018).

Essas técnicas de aprendizado de máquina, utilizadas em aprendizado supervisionado, possuem como vantagens a elevada robustez de seu aprendizado quando comparado a demais técnicas, a resistência a degradação de dados inconsistentes e a capacidade de ser implantada em ambientes produtivos e ser atualizada sem necessitar de uma mudança de arquitetura da rede de neurônios. Entretanto, como desvantagens estão o custo computacional mais elevado para treinamento e a necessidade de grande quantidade de dados para que uma rede neural artificial aprenda adequadamente (CARVALHO et al., 2019).

Redes neurais artificiais são uma das técnicas de aprendizado de máquina mais utilizadas para o problema de manutenção preditiva. Redes neurais artificiais foram utilizadas para a detecção de falhas críticas em componentes de turbinas eólicas, detecção de falhas em equipamentos em tempo real (através de redes *Long Short-Term Memory*, LSTM, e motor de processamento de dados Spark) e predição de falhas em sensores acústicos e painéis fotovoltaicos (através de redes *Convolutional Neural Network*) (CARVALHO et al., 2019). Também foram utilizadas para predição de tempo de vida útil restante de maquinário rotativo, motores aéreos (através de *Convolutional Neural Network*) e sistemas sob múltiplas condições operacionais (através de *Long Short-Term Memory bidirecional*) (XU et al., 2019). Já no setor ferroviário, redes neurais artificiais foram utilizadas para predição de defeitos estruturais em vias férreas (LIAO et al., 2022).

#### 2.2.1.3 Support Vector Machines

Outra técnica de aprendizado de máquina amplamente utilizada é as *Support Vector Machines* (SVMs). As SVMs são uma técnica de aprendizado supervisionado que procuram estabelecer um limiar entre duas regiões, num típico problema de classificação binário,

ilustrado pela Figura 7. Elas, porém, não são mais limitadas a esse cenário, podendo ser empregadas em problemas de múltiplas classes através da criação de múltiplos planos para cada grupo. Umas das vantagens mais aparentes dessa técnica é a elevada precisão de separação entre as diferentes classes (IGUAL; SEGUÍ, 2017)(CARVALHO et al., 2019).

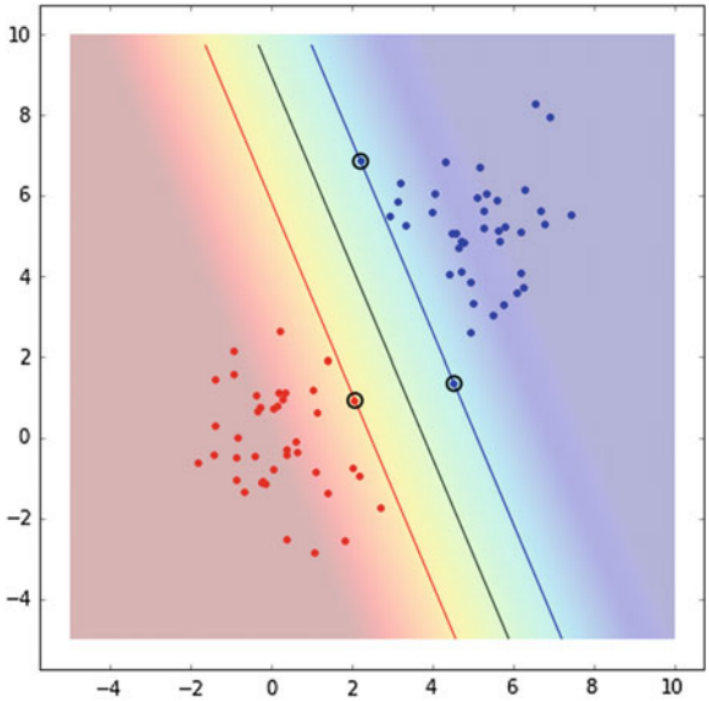


Figura 7 – Estabelecimento de limiar entre duas regiões através de SVMs (IGUAL; SEGUÍ, 2017).

Trabalhos acadêmicos utilizaram essa técnica para identificar falhas em caixas de transmissão de automóveis, construção de modelo preditivo de filamentos de íon, falhas de alarmes em via férrea, identificação de falhas em máquinas devido aos efeitos acumulativos de estresse e uso contínuo, predição de tempo de vida útil restante de máquinas e inclusive detecção de defeitos geométricos em vias férreas (CARVALHO et al., 2019)(XU et al., 2019)(LASISI; ATTOH-OKINE, 2018)(LIAO et al., 2022).

#### 2.2.1.4 Modelos cinzas

Criados a partir da Teoria de Sistemas Cinzas publicada em 1982 (JU-LONG, 1982), modelos cinzas são preditores de séries temporais para sistemas com informação incompleta. Apesar da pouca literatura disponível para esses modelos disponível em inglês (KAYACAN; ULUTAS; KAYNAK, 2010), derivações do modelo tradicional GM (1, 1) (*Grey Model*, com uma dimensão de sua equação diferencial e uma variável independente) têm sido desenvolvidas em tarefas de aprendizado de máquina voltadas à manutenção preditiva, como predição de degradação de ferrovia de linhas de passageiros e de linhas de serviços de carga pesada (LIAO et al., 2022).

#### 2.2.1.5 K-means

*K-means* é um algoritmo popular de aprendizado de máquina não-supervisionado, onde cada $K$ agrupamento é definido de forma iterativa, através da definição de centroides para cada grupo, e refinamento desses centroides a partir dos pontos mais próximos. Entre as principais vantagens desse algoritmo estão o relativo baixo custo computacional para sua execução em grandes conjuntos de dados, a capacidade de adaptação com a introdução de novos dados e a tendência de minimização da variância inter-classe. Como principal desvantagem, pode-se citar o problema de máximos locais, onde os grupos encontrados não são os melhores possíveis para o conjunto de dados, mas sim o melhor possível dado os centroides iniciais aleatórios.

Sistemas baseados em *K-means* para clusterização empregaram a técnica para analisar o comportamento de turbinas eólicas para classificar os tipos de falhas, extração de grupos de gases dissolvidos para encontrar os grupos que introduzem falhas para o sistema em questão, agrupamento de dados de ferramenta de derretimento a laser, detecção de falhas a partir de vibrações de exaustores e predição de falhas de motores (CARVALHO et al., 2019).

#### 2.2.1.6 Fuzzy systems

Lógica *fuzzy*, comumente traduzida como lógica difusa, é um sistema lógico introduzido no contexto de tratamento com incerteza, em que uma saída lógica não é um número discreto 0 ou 1, mas sim um número contínuo entre esses dois. Existem várias aplicações desse tipo de lógica nos mais variados algoritmos de aprendizado de máquina. *C-means*, por exemplo, é uma extensão do algoritmo *K-means*, onde cada ponto tem uma chance de pertencer a cada um dos grupos. De forma geral, a vantagem do emprego de sistemas *fuzzy* é a capacidade de incorporar ao modelo informações sobre a certeza (ou falta dela) de um resultado (ERTEL, 2018).

Sistema com lógica *fuzzy* foram utilizados para a tarefa de manutenção preditiva em alguns trabalhos. Modelos de Markov e sistemas *fuzzy* foram utilizados para estimar o tempo de vida útil remanescente a partir de sinais de vibração brutos, e sistemas *fuzzy C-means* foram empregados como comparação para análise de manutenção preditiva a partir de dados de vibração para exaustores (XU et al., 2019)(CARVALHO et al., 2019).

#### 2.2.1.7 PCA

*Principal Component Analysis* (PCA) é uma técnica comumente empregada na etapa de pré-processamento dos dados para a redução de dimensionalidade do *dataset* em questão. Esse algoritmo procura encontrar novas dimensões que concentrem a maior variância estatística possível das dimensões originais, preservando ao máximo as informações quantitativas do conjunto de dados em uma quantidade reduzida de dimensões (SKANSI,

2018). A Figura 8 a seguir ilustra esse processo, onde pontos de um sistema são transferidos para uma nova base do espaço vetorial de forma a aumentar a variância dos dados em uma dimensão (no caso, no eixo $y$), apresentada pelas barras paralelas aos eixos. Na sequência, a dimensão de menor variância, $x$, é descartada, reduzindo as dimensões totais dos dados.

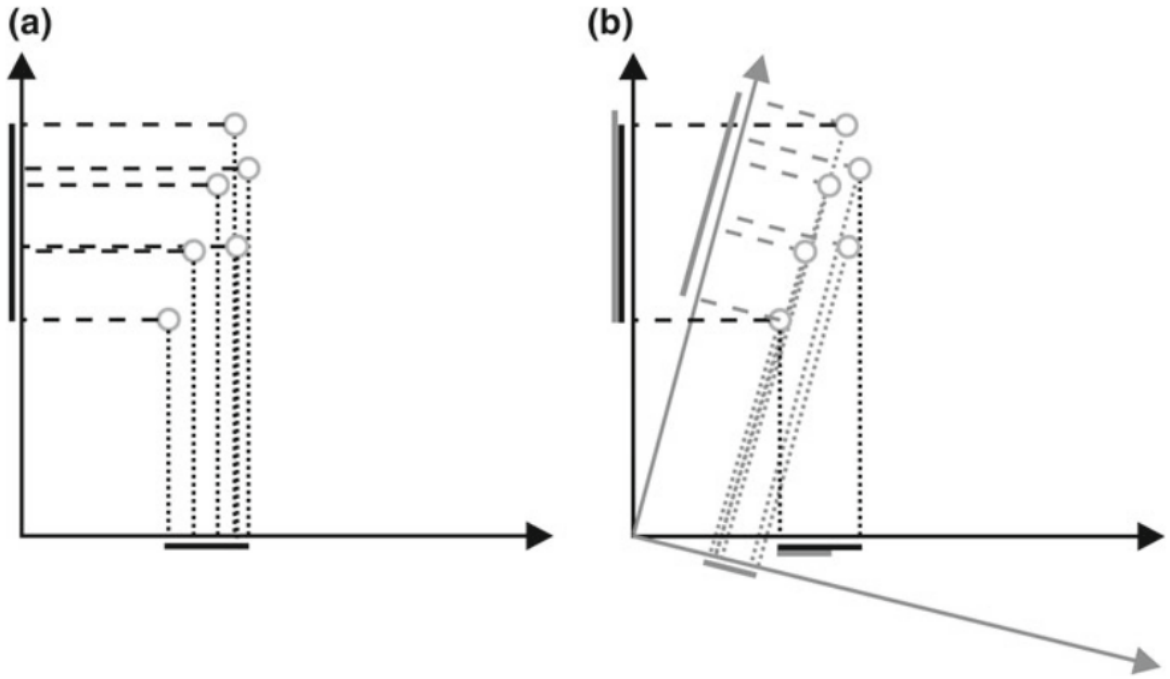


Figura 8 – Demonstração de aplicação de PCA em um sistema bidimensional. a) Dados na base inicial b) Resultado em uma nova base (SKANSI, 2018).

PCA foi utilizado em trabalhos onde se trabalhavam com dados de Carros Controle, coletando 11 dimensões após tratamento, e reduzindo para 3 após a aplicação dessa técnica, com um mínimo de 90% de variância representada (LASISI; ATTOH-OKINE, 2018).

### 2.2.2 Avaliação de desempenho de aprendizado supervisionado

Uma vez desenvolvidos os modelos de aprendizado de máquina supervisionados, é necessário avaliar os seus desempenhos. No contexto de aprendizado supervisionado, em que parte-se de um conjunto de dados pré-rotulados, pode-se comparar a classificação atribuída pelo modelo com os rótulos esperados. A partir daí, algumas métricas estatísticas são empregadas para avaliar os resultados obtidos (SKANSI, 2018).

Para tal, a partir do conjunto de dados completo, normalmente são realizadas subdivisões dos dados em subconjuntos para treinamento, validação e teste. O subconjunto de treinamento compõe a maior parte dos dados e é utilizado para treinar o modelo com as classes. Já o conjunto de validação é empregado para ajuste dos parâmetros internos

dos modelos, comumente empregado durante ensaios de validação cruzada. Por fim, o conjunto de testes é empregado para aferir a qualidade do aprendizado (IGUAL; SEGUÍ, 2017), e é em cima deste conjunto que as métricas de desempenho são realizadas.

Precisão é uma medição da probabilidade de um rótulo atribuído estar correto dado que ele foi classificado. É também uma forma de avaliar a proporção de falsos positivos que o modelo gera, em que entende-se por falso positivo todo aquele dado erroneamente classificado em um determinado rótulo. A fórmula da precisão é apresentada pela Equação 2.1, onde $TP$ simboliza verdadeiros positivos, que são os pontos que foram corretamente classificados na categoria esperada, e $FP$, falsos positivos.

Complementar à precisão, há a métrica de *recall*. Através dela se mede a probabilidade do modelo atribuir um certo rótulo a um dado pertencente a essa classe. A fórmula dessa métrica é a da Equação 2.2, onde $FN$ representa falsos negativos, que são os dados que erroneamente não foram categorizados no rótulo em questão.

Por fim, tem-se a métrica de acurácia, que mede uma visão geral do modelo, medindo a proporção de acertos sobre os dados totais, tal como apresentada pela Equação 2.3, onde $TN$ simboliza verdadeiros negativos, dados corretamente atribuídos a outra classe (SKANSI, 2018).

$$\frac{TP}{TP + FP} \tag{2.1}$$

$$\frac{TP}{TP + FN} \tag{2.2}$$

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{2.3}$$

### 2.2.3 Datasets públicos para validação

Com a ressalva de que o desafio de manutenção preditiva depende drasticamente do contexto ao qual se refere, foram desenvolvidos e publicados alguns *datasets* públicos para testes e validações de modelos para essa tarefa.

O *IEEE PHM2008 challenge dataset* é composto de dados de séries temporais multivariados coletados a partir de 218 motores turbofan. Existem 14 sinais de entrada e 58 sinais de saída, além de 13 parâmetros de saúde operacional e 21 variáveis dentro dos sinais de saída que podem medir a resposta do sistema mediante diferentes estados de saúde operacional.

O *Intelligent Maintenance Systems bearing dataset* foi construído a partir de dados de vibração coletados por acelerômetros. Existem 20480 pontos de coleta de dados na frequência de 20 kHz (XU et al., 2019).

Outros *datasets* construídos consistem de medidas de força e torque para detecção de falhas de robôs, dados de falhas de uma caixa de ferramentas, dados operacionais e de falha de um sistema de pressurização de um caminhão e dados de falha de um conjunto de 20 robôs com diferentes rodas (CARVALHO et al., 2019).

## 2.3 Manutenção de via férrea

Na área de manutenção de via férrea, atualmente é empregada uma gama de soluções de monitoramento. Os defeitos de geometria que são monitorados por tais sistemas são alinhamento lateral, alinhamentos verticais à direita e à esquerda, *gauge*, *twists*, juntas danificadas e ondulações de estrutura (WESTON et al., 2015). Neste contexto, são definidas normas do setor para controle e regulamentação das ferrovias. Entre elas, as normas europeia EN 13848-5 e a norte-americana AREMA definem limites de segurança em cima de parâmetros geométricos utilizados para a definição de irregularidades da via permanente. Por exemplo, a EN-13848-5 estabelece limites mínimos e máximos para a bitola, definida como a distância perpendicular entre os boletos dos trilhos, em função da velocidade de tráfego da ferrovia (COMITÉ EUROPÉEN DE NORMALISATION, 2017) (RIBEIRO, 2017).

Sobre sistemas de sensores de acompanhamentos, observam-se soluções desde sensores únicos até implementações completas com sistemas autônomos de medição de geometria. No Reino Unido, esse sistemas autônomos estão presentes nos mais recentes trens. Observa-se que essas soluções são mais robustas e evitam problemas de cálculo que surgem com sistemas mais simples e com menos sensores (WESTON et al., 2015). Tradicionalmente, aparelhos como Carro Controle para medições precisas de geometria de via são empregados em várias ferrovias ao redor do mundo, portando sensores sofisticados para medição de bitola, empeno e nivelamentos transversais e longitudinais (LINGAMANAIK et al., 2017).

Defeitos de geometria da via permanente, especialmente os de alinhamento e nivelamento, podem afetar o balanço de um vagão, especialmente quando este percorre a via em elevadas velocidades. Irregularidades na pista são transmitidas à massa do veículo através de seus truques, conforme exemplifica a Figura 9 (SILVA, 2019). Dessa forma, ao medir as forças atuantes nesse componente, tem-se uma noção razoável em relação à qualidade da via e à forma em que ela afeta os vagões. Por isso, normas como a AAR M976 definem parâmetros como folga de suspensão, rolagem de caixa e galope (JÚNIOR et al., 2020), medições que, somadas a anos de experiência do setor ferroviário com dinâmica de veículos ferroviários, tornaram-se base para a especificação de limites a serem observados para tomada de decisão, como restrição de velocidade da ferrovia (DARBY et al., 2003).

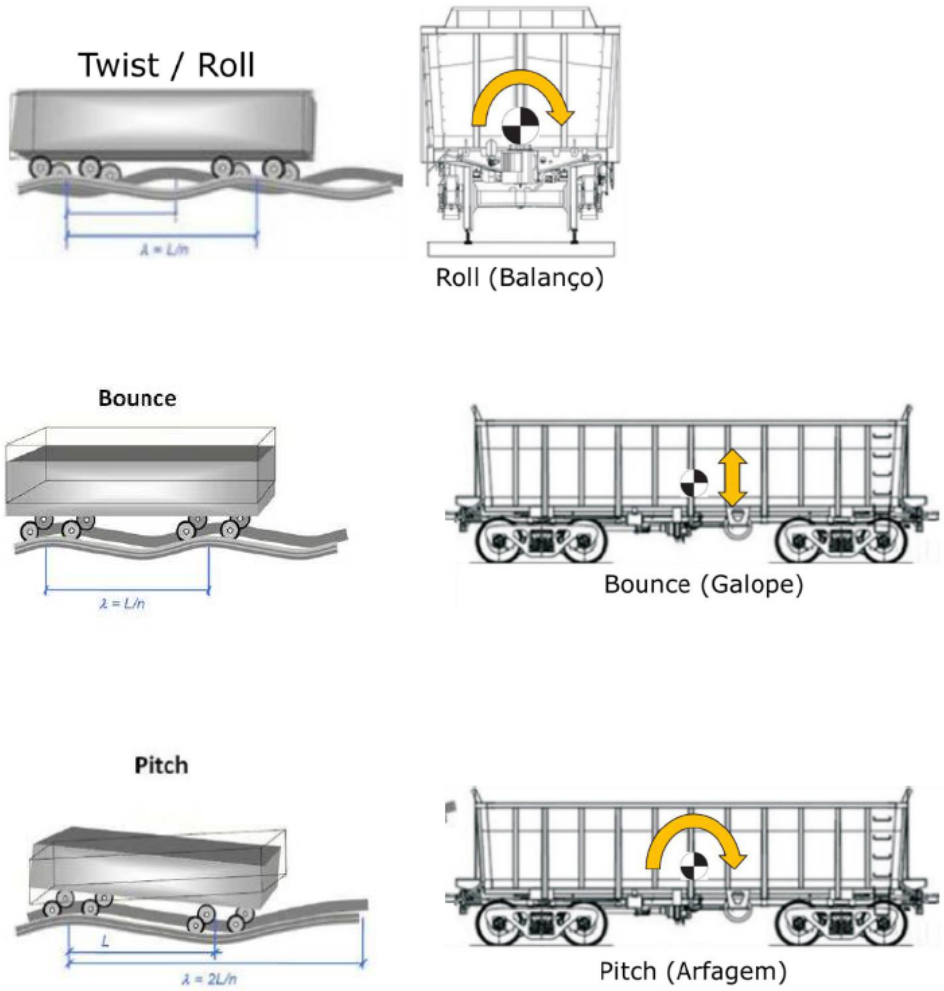


Figura 9 – Principais modos de oscilação de um veículo ferroviário em função dos tipos de deformação na via permanente (JÚNIOR et al., 2020).

Sobre sensores utilizados, são empregados tacômetros, acelerômetros, giroscópios, sensores de deslocamento, sensores ópticos 2D e sistemas de satélites de navegação global, além de outros menos comuns, como detetores de metal e gravação de vídeo. Geralmente, esses sensores são usados em conjunto de forma a melhorar a precisão das medições (WESTON et al., 2015), usualmente utilizados em Vagões Instrumentados para uma coleta mais frequente e econômica (DARBY et al., 2003).

No caso da EFVM em particular, Vagões Instrumentados utilizados para monitoramento são munidos de sensores como acelerômetros uniaxiais, acelerômetros triaxiais, sensor de pressão e células de carga, estruturados no veículo de forma a coletar uma série de medidas relevantes em relação à sua dinâmica na ferrovia. Dentre elas, estão algumas relacionadas ao movimento do vagão em si, como aceleração lateral no corpo do vagão, aceleração vertical em seus truques, pressão na sua linha de freio, e deslocamento nos quatro truques do veículo (TUDEIA et al., 2019). Para as grandezas em relação à

dinâmica da via, as métricas folga de suspensão, rolagem de caixa e galope estão bastante relacionadas a avarias na via permanente, o que as torna de particular importância no contexto de monitoramento de ativos de infraestrutura. Para calcular essas variáveis, uma das principais medidas utilizadas é o deslocamento vertical da suspensão secundária, que é mensurada a partir de uma célula de carga configurada da forma apresentada na Figura 10 (JÚNIOR et al., 2020).

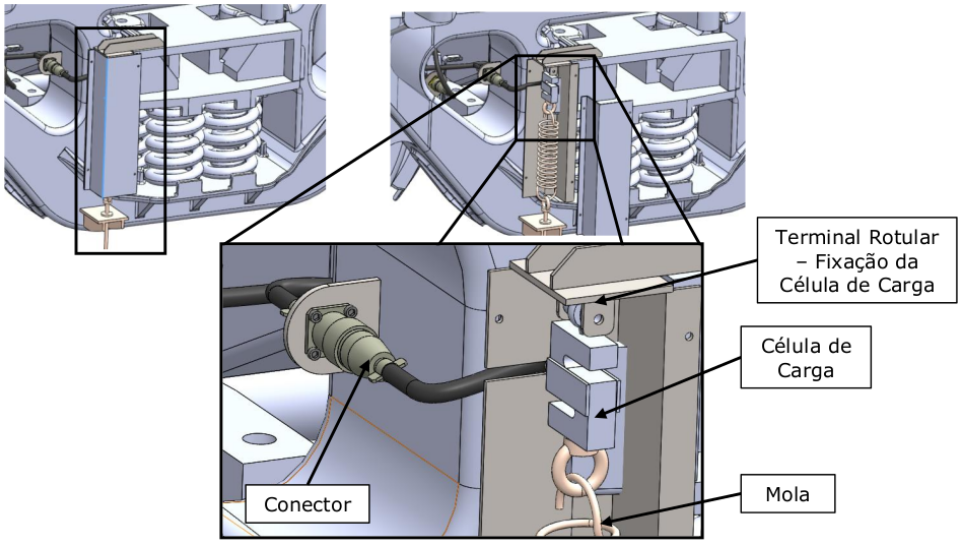


Figura 10 – Leiaute de instalação da célula de carga para medição do deslocamento vertical da suspensão secundária no VI (TUDEIA et al., 2019).

Por fim, é comum a transmissão de dados coletados pelos Vagões Instrumentados para uma localização central, onde o processamento pode ocorrer com maior poder computacional e com potenciais registros históricos para melhor análise (KONDAKA et al., 2022). Para tanto, a transmissão dos dados pode ser feita via redes móveis, como 3G ou 4G, ou via dispositivos físicos, como discos rígidos ou cartões SD. Por exemplo, no caso do VI da EFVM, um *modem* 3G de dois cartões SIM é utilizado em conjunto com armazenamento local limitado em cartões SD (TUDEIA et al., 2019).

Na sequência, alguns sistemas fazem uma validação dos dados capturados com mapas da via percorrida, de forma a complementar alguns tipos de informações que os sensores não conseguem capturar bem, como a exata linha do trajeto do veículo. Eventuais diferenças entre os dados e o trajeto esperado podem ser confrontadas e analisadas. Outra possibilidade é a comparação de desvios de geometria de diferentes viagens, assumindo que o servidor central mantenha um registro de viagens passadas. Por fim, outras fontes de dados sobre a condição da via férrea podem ser buscadas e comparadas com os registros coletados e armazenados no servidor central (WESTON et al., 2015).

## 2.4 Trabalhos relacionados

Na literatura, estão documentados vários casos de *Big Data* aplicados à manutenção preditiva de ferrovias. Os principais trabalhos apresentam significativa variedade de objetivos e soluções propostas dentro desse domínio.

NUNEZ et al. (2014) propôs métodos de se aplicar os 5Vs de *Big Data* (volume, velocidade, variedade, veracidade e valor) para facilitar a tomada de decisão de aplicabilidade de manutenção de vias férreas, e depois seguiu com um estudo de caso para vias holandesas envolvendo medidas de aceleração da caixa de eixo de vagões (ABA) e GPS. Para lidar com a grande quantidade de dados a serem processados e o poder computacional limitado, os autores propõem um método de redução de frequência de coleta de dados, em que trechos menos sensíveis são analisados mensalmente, e trechos mais críticos ou movimentados são analisados semanalmente ou mesmo diariamente. Como demonstração com dados reais, os autores escolheram dados a partir de uma das viagens de teste dos sensores ABA num trecho de dois quilômetros entre Assen e Groningen. A partir dos dados, foram identificados picos de energia, que podem sugerir locais que requerem manutenção. Analisando a densidade de picos, os autores sugerem que os quilômetros nos quais foram identificados múltiplos picos devem ter a frequência de coleta de dados intensificada. Na sequência, foi analisado um pico particularmente severo, em que dados visuais do local foram combinados para a sua classificação como defeito e que, seguindo a árvore de decisão proposta, deve ser remediado. Por fim, outro pico foi analisado, dessa vez um menos severo que, quando combinado com dados visuais, confirmou-se um pequeno defeito, cuja manutenção pode ser postergada para o próximo mês, evitando custos imediatos, e a frequência de coleta de dados dessa região deve ser aumentada. Apesar dos resultados positivos apresentados, é relevante ressaltar que o escopo do estudo de caso é mais limitado, deixando em aberto a questão da escalabilidade da solução para mais trechos da ferrovia.

JAMSHIDI et al. (2018) também utiliza vias holandesas para demonstrar o resultado do sistema proposto, particularmente a via Amersfoort-Weert. O projeto propõe uma metodologia para auxiliar a tomada de decisão de manutenção de vias férreas. Essa metodologia consiste de cinco grande etapas: monitoração inteligente das condições da via, conhecimento prévio da via, análise de interdependência entre os dois elementos anteriores, modelo difuso (*fuzzy*) de inferência e decisões de manutenção da via. O primeiro passo consiste basicamente de duas fontes de dados para a detecção de defeitos: ABA e imagens da via capturadas a partir de câmeras de alta cadência de quadros posicionadas em vagões. O estudo em questão foca-se na detecção de defeitos superficiais e utiliza-se de um modelo de rede neural profunda convolucional construída para a classificação de imagens em quatro níveis de severidade do defeito capturado. Na sequência, as imagens já rotuladas são pareadas com o sinal ABA capturado no mesmo instante para determinar a severidade final. Já o passo que envolve conhecimento prévio da via classifica sete fatores influenciadores de

defeitos de via em perfis de via, irregularidades de via e perfis de pesagem e velocidade dos vagões. No passo seguinte, cruzam-se as informações dos setes fatores de perfis de vias com a detecção de defeitos via os sinais ABA e imagens. A seguir, a interdependência é definida pelo cruzamento da severidade do defeito com os fatores influenciadores da via no local em questão. No passo do modelo de inferência, um sistema difuso é utilizado para criar as regras de avaliação de saúde da via com base nos fatores de influência e nas severidades dos defeitos encontrados. Por fim, o último passo se vale de um algoritmo para maximizar o aproveitamento do tempo das equipes de manutenção, de forma a priorizar os segmentos mais críticos e sugerir outros à medida que se sobra tempo para mais manutenções. Este é um trabalho que apresenta resultados significativos, mas difere da pesquisa em questão pelo emprego de câmeras e vídeos, algo não disponível no presente projeto.

SALIERNO et al. (2020) propõe uma arquitetura de quatro camadas de Big Data para o gerenciamento de dados coletados de vias férreas especificamente para o problema de manutenção preditiva, seguindo com uma implementação de rede LSTM e demonstração do projeto com dados reais da via italiana Milano-Monza-Chiasso. Para isso, o projeto faz uso de dados de arquivos de registro de um sistema de intertravamento ferroviário. Esse sistema computacional garante que nenhum trem ocupe uma via onde outro veículo já esteja trafegando, evitando acidentes. Assim, esse sistema produz registros sobre os comandos executados para a troca de trilhos e seus efeitos. Com isso em mente, a arquitetura é composta por quatro camadas: armazenamento, processamento (responsável pelas manipulações e transformações dos dados), serviço (camada que contém ferramentas para análise dos usuários finais) e ingestão (responsável pela agregação dos dados a partir de fontes externas). Para a implementação da arquitetura, foi utilizado o ecossistema Hadoop. Assim, a camada de armazenamento foi implementada através do *Hadoop Filesystem* (HDFS), e a camada de processamento utilizou arquivos XML especificando as *feaures* mais relevantes dos dados e tabelas HIVE para importação dos arquivos. Já a camada de serviço foi implementada através de *Jupyter notebooks*, e, por fim, a camada de ingestão foi implementada através do Apache NiFi. Por fim, o projeto treinou uma rede LSTM com os dados coletados de um ponto específico da via italiana para detectar falhas, em especial dados de potência fornecida, voltagem e tempo de movimento da peça. Assim, esse trabalho apresenta uma implementação do ciclo de vida dos dados discutido anteriormente de grande relevância, apesar do domínio em questão ser diferente desta pesquisa.

SHARMA et al. (2018) desenvolve uma política de manutenção baseada em dados para a inspeção e manutenção da geometria da via. Para isso, o projeto adota um índice de qualidade de via (TQI) como um indicador geral da habilidade da via férrea de garantir os movimentos esperados dos vagões com segurança. Para a tomada de decisão, foi desenvolvido um modelo de processo de Markov em conjunto com uma simulação de Monte Carlo. Na sequência, foram realizados estudos de caso a partir de dados coletados de 50 milhas de uma via férrea Classe 1, de março de 2009 a dezembro de 2011. As variáveis

utilizadas pelo modelo de predição foram o TQI, tonelagem, limite de velocidade do carro de carga, número de dias entre as inspeções das vias e indicadores sobre falhas encontradas durantes as inspeções atuais e mais recentes. Para fazer a análise dos dados, primeiro foi calculado o TQI de cada região da via e, na sequência, utilizaram-se *random forest* e SVM para a previsão de defeitos geométricos. Na sequência, discretizaram-se os estados para a cadeia de Markov ao dividir o TQI em cinco níveis de percentil. Como resultado, probabilidades da transição de um estado do modelo de Markov para outro (representando a degradação da região da via) foram calculadas, apresentando um resultado tido como satisfatório. Como limitação para o escopo da pesquisa, cita-se a ausência de um TQI brasileiro bem estabelecido, além do emprego de dados indisponíveis, como tonelagem do vagão.

## 2.5 Conclusões parciais

Com a disciplina de Ciência dos Dados estabelecendo práticas consagradas para extração de conhecimento a partir de grande volume de dados, aliado à grande variedade de algoritmos de aprendizado de máquina sendo empregados com êxito em tarefas semelhantes, evidencia-se a viabilidade da presente pesquisa. Somado ainda aos empreendimentos recentes do setor ferroviários com os instrumentos Vagão Instrumentado e Carro Controle, em que se verifica o alinhamento de normas técnicas com as leituras provenientes dos seus sensores, conclui-se que o objetivo proposto demonstra ser factível de ser alcançado.

Embora já houve trabalhos na literatura dentro do tema de manutenção preditiva com veículos munidos de sensores, não se verificam projetos com requisitos e condições próximas dos atuais. Grande parte da literatura foca em apenas um dos tipos de veículos, além de serem raras as integrações com soluções para apresentação dos resultados dos modelos, aspecto fundamental para a eventual implantação da solução atingida.

# 3 Metodologia

## 3.1 Modelagem de dados

Os dados a serem trabalhados no projeto são as leituras provenientes do Vagão Instrumentado e Carro Controle, além dos registros de ocorrências na EFVM. Entretanto, é necessário estabelecer uma modelagem desses dados para que as operações corriqueiras sejam facilitadas, bem como facilitar a incorporação de todas as diversas fontes de dados num modelo único e centralizado. Além disso, os dados brutos carecem de abstrações necessárias para a manipulação e entendimento das amostragens, fazendo-se necessário a incorporação dessas concepções na construção da modelagem dos dados.

A Tabela 2 apresenta as variáveis do VI, e a Tabela 3, as do CC. Ambas as Tabelas apresentam, além dos dados brutos dos sensores, as abstrações introduzidas pela modelagem dos dados.

| Variável | Tipo | Variável | Tipo |
|---|---|---|---|
| Data e hora | Datetime | Medidor de tensão 1 | Racional |
| Viagem | Inteiro | Medidor de tensão 2 | Racional |
| Km inicial | Inteiro | Acelerômetro eixo X | Racional |
| Km final | Inteiro | Acelerômetro eixo Y | Racional |
| Latitude | Racional | Acelerômetro eixo Z | Racional |
| Longitude | Racional | Acelerômetro dianteiro direito | Racional |
| LRS | Racional | Acelerômetro dianteiro esquerdo | Racional |
| EH/RH | Texto | Acelerômetro traseiro direito | Racional |
| Linha | Inteiro | Acelerômetro traseiro esquerdo | Racional |
| Aceleração | Racional | Deslocamento dianteiro direito | Racional |
| Folga de suspensão | Racional | Deslocamento dianteiro esquerdo | Racional |
| Rolagem de caixa | Racional | Deslocamento traseiro direito | Racional |
| Galope | Racional | Deslocamento traseiro esquerdo | Racional |
| Velocidade | Racional | Deslocamento dianteiro direito filtrado | Racional |
| Pressão | Racional | Deslocamento dianteiro esquerdo filtrado | Racional |
| Tensão | Racional | Deslocamento traseiro direito filtrado | Racional |
| | | Deslocamento traseiro esquerdo filtrado | Racional |

Tabela 2 – Descrição dos parâmetros dos dados do Vagão Instrumentado.

| Variável | Tipo | Variável | Tipo |
|---|---|---|---|
| Data | Date | Raio inverso | Racional |
| KM e metro | Racional | Empeno (corda 5,5m) | Racional |
| LRS | Racional | Empeno (corda 10,0m) | Racional |
| Linha | Inteiro | Empeno (corda 1,7m) | Racional |
| EH/RH | Texto | Empeno (corda 2,0m) | Racional |
| Elemento | Texto | Nivelamento esquerdo | Racional |
| Bitola traseira | Racional | Nivelamento direito | Racional |
| Bitola frontal | Racional | Nivelamento longitudinal | Racional |
| Variação da bitola | Racional | Alinhamento (corda 3m) | Racional |
| Suspensão | Racional | Alinhamento (corda 10m) | Racional |
| Acelerômetro X | Racional | Variação alinhamento (corda 3m) | Racional |
| Acelerômetro Y | Racional | Variação alinhamento (corda 10m) | Racional |
| Acelerômetro Z | Racional | Inclinação trilho esquerdo | Racional |
| | | Inclinação trilho direito | Racional |

Tabela 3 – Descrição dos parâmetros dos dados do Carro Controle.

### 3.1.1 Sistema linear de referência

Para o problema de localização especial de elementos e dados num espaço unidimensional, como é o caso do trabalho com ferrovias, é interessante utilizar algum sistema linear de referência, de forma que cada elemento pode ser localizado por um número escalar. Dessa forma, para operações como filtros e cruzamento de localizações, basta se conhecer os valores escalares nesses sistema (CURTIN; TURNER, 2019).

Um benefício do emprego de um sistema linear de referência (LRS, do inglês *linear referencing system*) de localização é a construção de filtros adequados. Dadas as curvas da ferrovia, a implementação de seletores com base apenas nas coordenadas GPS pode resultar em recortes incorretos. A Figura 11 a) mostra um exemplo de aplicação levando em conta apenas os mínimos e máximos dos valores de GPS: observa-se que há uma região considerável que foge a esses limites, resultando numa seção sendo ignorada pelo filtro. Já a Figura 11 b) ilustra a aplicação do seletor para a mesma localização, mas dessa vez utilizando os mínimos e máximos do LRS. O que se observa é o resultado esperado, obtido através de um simples filtro escalar entre dois números.


Figura 11 – a) Filtro para uma região da EFVM utilizando apenas coordenadas GPS. b) Mesmo filtro utilizando o LRS. Pontos em azuis representam leituras do VI e pontos em amarelo, *Relay Houses* (RHs). Mapa construído com dados de OpenStreetMap (2017).

Na prática, o sistema linear utilizado em campo é a referência por quilometragem. Essa marcação utilizada para a EFVM, porém, não é respeitada constantemente pelas anotações. Não apenas existem pontos na via onde a marcação zera, necessitando de mais informações complementares para a correta demarcação espacial de um ponto na via, como também existem leituras do Vagão Instrumentado que não são consistentes com a quilometragem estabelecida.


Figura 12 – Filtro dos dados do VI entre as RHs 35 e 46 utilizando como parâmetro geoespacial a quilometragem. A existência de dois segmentos de elementos, um deles (em vermelho) não coincidindo com a real localização desejada (em azul), demonstra que a referência utilizada para marcação de quilometragem não é constante. Mapa construído com dados de OpenStreetMap (2017).

A Figura 12 resultante acima demonstra bem o problema de inconsistência de quilometragem entre diferentes viagens. A partir de um *dataset* de referência dos elementos da ferrovia, são obtidos os quilômetros referentes às *Relay Houses* (RHs) em questão. Com base nesses valores, faz-se um filtro simples pelo *dataset* de dados de viagens do VI. O esperado seria apenas um único trecho contínuo, estando totalmente contido pelas RHs definidas. Porém, o que observa são dois trechos desconexos, com um deles estando em um segmento totalmente diferente daquele desejado, a quilômetros de distância do imaginado. Isso ocorre devido a uma das viagens do VI não estar consistente com a quilometragem estabelecida. Um exemplo dessa situação é apresentada pela Figura 13, em que se observa que registros do VI de um mesmo trecho da EFVM apresentam marcações de quilometragem inconsistentes entre diferentes viagens.

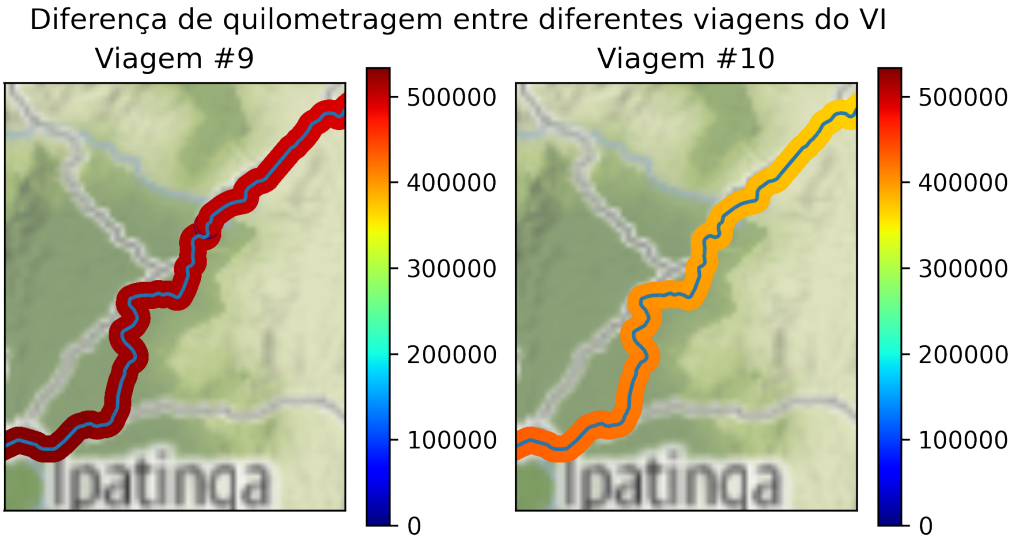


Figura 13 – Mapas de algumas viagens do VI ilustrando a quilometragem marcada. A marcação registrada em uma não é consistente com a marcação adotada em outra, conforme a diferença de tonalidade da escala mostra. Mapa construído com dados de OpenStreetMap (2017).

Assim, fica evidente a necessidade do emprego de um novo sistema linear de referência a pontos na ferrovia. O objetivo desse sistema é a criação de uma função inversível de uma coordenada geográfica para um número escalar, representando uma posição na EFVM, conforme Equação 3.1. Assim, por exemplo, poderia-se traduzir as coordenadas $-20.246650, -40.258098$ para a posição 0.270561, e vice-versa.

$$f(lat, lng) \longleftrightarrow lrs \tag{3.1}$$

Para isso, unificam-se os segmentos de reta que compõe a EFVM, cujos pontos são formados por coordenadas de GPS, numa base de referência. A partir dessa base, todo novo ponto georreferenciado é projetado nessa base, que, por ser unidimensional, retorna um elemento de uma única dimensão, satisfazendo o requisito de um LRS. Esse processo é apresentado pela Figura 14.

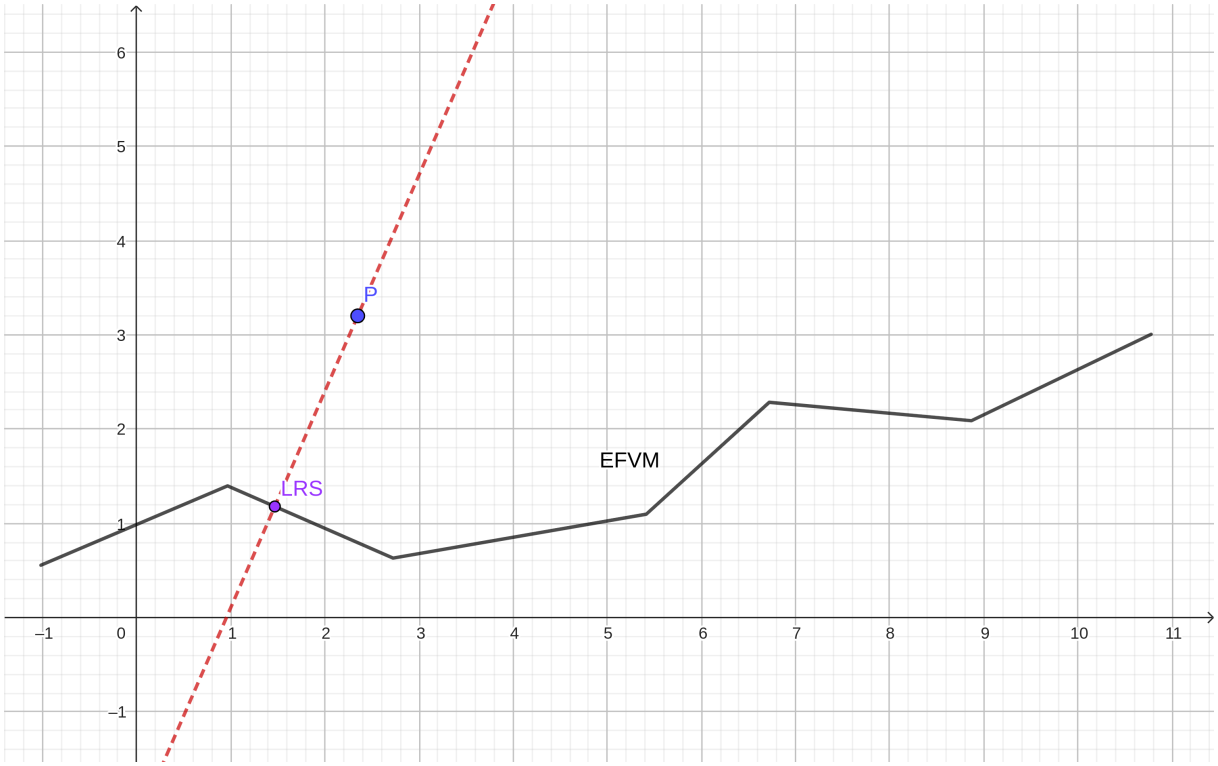


Figura 14 – Transformação de uma posição $P$ de GPS para o sistema linear de referência do projeto, resultando no ponto $LRS$, cuja posição em relação à referência $EFVM$ é representada por um número escalar.

#### 3.1.1.1 Inferência de coordenadas do VI com base no LRS

As viagens do Vagão Instrumentado anteriores a Agosto de 2020 não têm dados de GPS. Com isso, não apenas não é possível ilustrar os pontos num mapa, como também não se pode utilizá-los nos experimentos de aprendizado de máquina devido ao problema de inconsistência da quilometragem discutido anteriormente, impossibilitando a devida geolocalização desses dados e, assim, impedindo a execução da estratégia de validação estabelecida na Seção 3.2. Porém, desprezar esses dados reduz significativamente o *dataset* de viagens do VI. Assim, é interessante de alguma forma inferir a posição geográfica das leituras do VI com base em outros valores conhecidos.

A estratégia para inferir as coordenadas das viagens anterior a Agosto é a seguinte: a partir da quilometragem que for consistente com aquela especificada nos dados de elementos da via, cria-se uma função polinomial linear para converter a quilometragem para o LRS do projeto, de forma a possibilitar o cálculo de longitude e latitude de cada ponto.

Primeiro, são verificadas quais viagens apresentam dados de quilometragem consistentes com os dados de elementos da via. Essa verificação consiste em analisar as quilometragens de cada RH e comparar com as referências. Caso haja alguma divergência,

a viagem é marcada como inválida.

Na sequência, para se converter da quilometragem para o LRS, basta encontrar o polinômio de primeiro grau que represente bem a relação entre ambas as grandezas. Com isso, tem-se uma função polinomial de conversão de quilometragem para LRS. De posse dessa função, basta preencher as colunas de LRS, longitude e latitude do VI com base nos resultados calculados e persisti-los de forma a serem reaproveitados em outros estudos.

Finalmente, os dados de GPS podem ser obtidos através da função inversora da função de cálculo de LRS, apresentação na Equação 3.1.

### 3.1.2 Variáveis calculadas do VI

O Vagão Instrumentado é composto por múltiplos sensores, como acelerômetros uniaxiais, acelerômetros triaxiais, sensor de pressão e células de carga, cada um efetuando suas próprias medições. Com base nas leituras brutas desse conjunto de sensores, é importante o estabelecimento de certas variáveis capazes de representar bem a situação geral de uma leitura com base numa agregação de múltiplas leituras (JÚNIOR et al., 2020).

Dentre as medições captadas pelo Vagão Instrumentado, conforme discutido anteriormente na Seção 2.3, três variáveis são de particular importância para a análise e entendimento da situação da via, frequentemente utilizadas em campo: folga de suspensão (*suspension travel*), rolagem de caixa (*bodyrock*) e galope (*bounce*). Folga de suspensão é o deslocamento de umas das suspensões em relação às demais, galope é a média do desvio da mola dianteira e traseira, muitas vezes relacionada ao movimento de arfagem e, por fim, rolagem de caixa é a medição da rotação em torno do eixo horizontal longitudinal do trem (TUDEIA et al., 2019).

JÚNIOR et al. (2020) define algoritmos para o cálculo dessas três relevantes variáveis a partir das leituras brutas dos sensores do Vagão Instrumentado, além de uma adicional para a aceleração combinada das três dimensões. Dada a importância dessas variáveis nas análises de qualidade da via, é desejável que esses algoritmos sejam integrados ao modelo de representação dessas capturas. A Tabela 4 a seguir ilustra quais leituras são necessárias para o cálculo dessas variáveis.

| Variável | Dependências |
|---|---|
| Aceleração | Aceleração vertical dos quatro acelerômetros da travessa lateral |
| Folga de suspensão | Deslocamento vertical da suspensão secundária |
| Rolagem de caixa | Deslocamento vertical da suspensão secundária |
| Galope | Deslocamento vertical da suspensão secundária |

Tabela 4 – As quatro variáveis de interesse e suas dependências de leituras brutas do Vagão Instrumentado.

### 3.1.3 Modelo multidimensional

Os dados de leituras do VI e CC possuem algumas características distintas que devem ser levadas em conta na construção dos modelos para os representarem.

Todo registro obtido é composto por uma data e hora, informações de localização, e valores observados pelos sensores. Esse conjunto de variáveis pode ser divido em três grandes dimensões: tempo, espaço e valores. A dimensão tempo é fundamental para que se possa observar as mais diversas relações de evolução das condições da via permanente aferidas pelos vagões. Já a dimensão espaço é crítica para o correto posicionamento dessas leituras nas regiões corretas, diferenciando pontos semelhantes de localidades diversas. Por fim, a dimensão de valores diz respeito às leituras brutas dos sensores, sendo estes os principais números que de fato medem o que o veículo observa ao trafegar na via. Serão esses os valores que constroem as matrizes de dados a serem alimentadas pelos modelos de aprendizado de máquina.

Além do aspecto multidimensional, os modelos para os dados do VI e CC também serão cruzados com os conjuntos de elementos da via e falhas. A partir da dimensão espacial de um registro de veículo, representado pelo sistema linear de referência detalhado anteriormente na Seção 3.1.1, a informação equivalente dos demais conjuntos é extraída de forma a se estabelecer um relacionamento entre esses dados. Assim, um registro de veículo é expandido com informações sobre qual o elemento de via ele se encontra, permitindo que sejam realizados filtros como obtenção de registros realizados apenas em um determinado tipo de elemento (curvas, tangentes, pontes, etc), entre duas RHs escolhidas, contidos em um determinado elemento da via, entre outros.

A característica multidimensional é levada em conta quando da construção do relacionamento dos registros dos veículos com falhas. Aqui, não apenas é necessário o cruzamento com a dimensão espacial, tal como é realizada para o *dataset* de referência da EFVM, como também é essencial a correta correlação com o tempo. Assim, ambas as dimensões são cruzadas com as suas respectivas do conjunto de falhas para se relacionar um registro do VI e CC com um defeito registrado na mesma região, e em uma certa distância temporal entre as duas ocorrências.

A Figura 15 a seguir ilustra o modelo de dados para os dados do VI e CC, evidenciando sua multidimensionalidade e capacidade de relacionamento com os *datasets* de falha e elementos da EFVM através de suas dimensões.

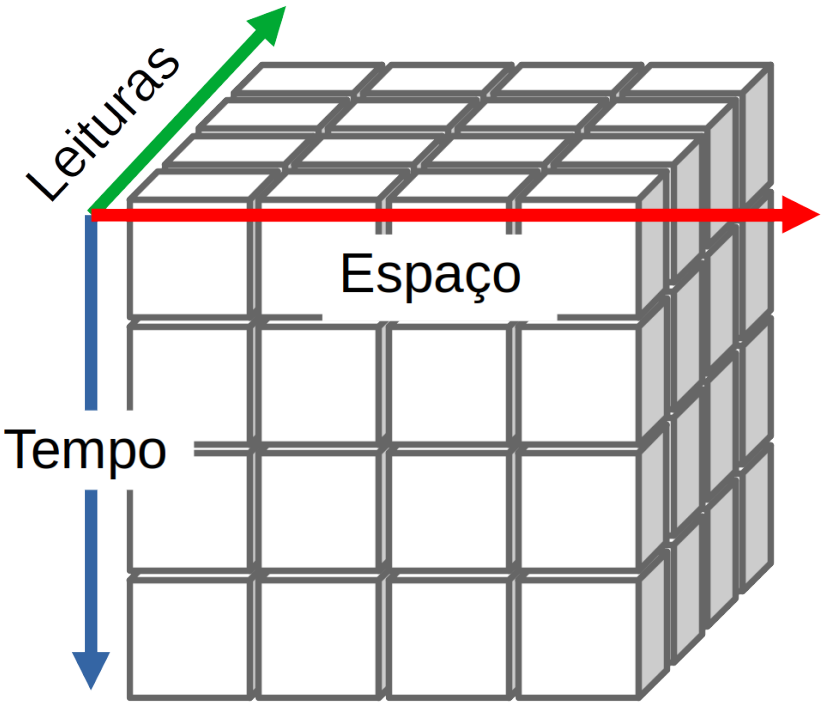


Figura 15 – Modelo de dados de veículos de 3 dimensões: espaço, tempo e leituras dos sensores, com o qual podem ser relacionados os dados de registro de ocorrências e de elementos da EFVM. Cada cubo representa um ponto de dado distinto, descrito pelas três dimensões do modelo.

Não são apenas os registros dos veículos que carecem de modelagem. Os já mencionados conjuntos de dados de referência da via e falhas registradas também serão devidamente modelados seguindo as dimensões espaciais e temporais propostas. Para o conjunto de defeitos, porém, é necessário uma extensão da dimensão espacial em alguns valores possíveis. Isso porque esses dados podem ser espacialmente identificados utilizando apenas a referência ao elemento de via em que se foi observada a falha, ou em uma quilometragem mais específica e precisa de onde foi realizada a manutenção. Essas duas informações se mostram independentes quando realizada a ingestão desse conjunto de dados, ora estando ambas presentes, ora apenas uma delas, então é importante adaptar o modelo para que essa dimensão seja adequada à realidade dos dados. Assim, tem-se que a dimensão espacial dos dados de falha possui diferentes escalas. Como consequência, ao cruzar as ocorrências de falhas com os conjuntos do VI e CC, a escala da localização que se deseja fazer o relacionamento espacial deve ser especificada.

## 3.2 Modelagem de aprendizado de máquina

É sabido que os dados de leituras do VI em particular se comportam de forma diferente dependendo do estado da ferrovia em que o transporte circula, conforme discutido na Seção 2.3. Assim, para o estabelecimento de níveis de severidade para a construção de modelos preditores de manutenção, propõe-se a utilização de algoritmos de aprendizado de máquina não-supervisionado. Esses algoritmos têm como objetivo a determinação de grupos de dados, ou *clusters*, tais que dados similares são agrupados em grupos comuns e dados diversos, em grupos distintos. Dessa forma, espera-se que esse agrupamento (ou *clustering*) reflita a distinção existente entre registros em regiões deterioradas (alta severidade) e regiões em condições normais de operação (baixa severidade).

Um ponto importante para esse processo é a necessidade de se filtrarem dados do VI em apenas certos elementos da via: curvas e tangentes. A explicação para esse cuidado se dá na dinâmica do vagão ao entrar e sair de componentes tal como pontes, Juntas Isoladas Coladas (JICs) e Aparelhos de Mudança de Via (AMVs), em que se espera alguma vibração natural como consequência; vibração que afeta as leituras dos sensores e, fora de contexto, poderia ser indicativo de falha quando, recontextualizada, é característico de passagem de elemento de via.

Outro aspecto fundamental a ser levado em conta para o estudo é a influência que eventuais *outliers* das variáveis de sensores têm sobre os resultados. Valores muito distantes em relação ao desvio padrão da média das leituras podem significativamente impactar o modelo, imprimindo um viés errôneo em algumas leituras atípicas. Assim, um algoritmo para a remoção desses valores inusuais deve ser empregado para verificar essa hipótese. Para essa tarefa, o método amplitude interquartil (*interquartile range*, IQR) foi escolhido, em que cada variável é filtrada de forma que um valor seu $x$ permanecesse no conjunto de dados apenas se a condição $Q1 - 1,5IQR < x < Q3 + 1,5IQR$ for satisfeita, onde $IQR = Q3 - Q1$, $Q1$ é o primeiro quartil da variável (25%) e $Q3$ é o terceiro quartil (75%). O IQR, além de ser comum na literatura para esse tipo de função, é computacionalmente barato de ser executado, uma característica desejada, já que os experimentos são executados repetidas vezes em configurações distintas de quais variáveis (se alguma) terão seus *outliers* tratados.

Com os *clusters* definidos, a estratégia de validação a ser empregada para esses resultados é definir intervalos de distância em dias de uma falha a dados do VI na mesma região, e comparar com a presença dos *clusters* nesses recortes de dados. Afinal, é esperado que um adequado nível de severidade elevada se manifeste apenas quando há iminência de uma falha na via. De forma similar, um nível de severidade intermediário só estaria presente num intervalo de tempo maior até a ocorrência do defeito, e assim sucessivamente para as menores severidades.

Dessa forma, é importante, para cada ponto do VI agrupado, encontrar a próxima falha registrada no mesmo setor da EFVM. Depois, definida uma janela de tempo em que se espera que os níveis de severidades comecem a se manifestar antes da ocorrência da avaria, filtram-se apenas os pontos cujo setor tenha um registro de falha dentro dessa janela de tempo. Esse procedimento é ilustrado pela Figura 16.

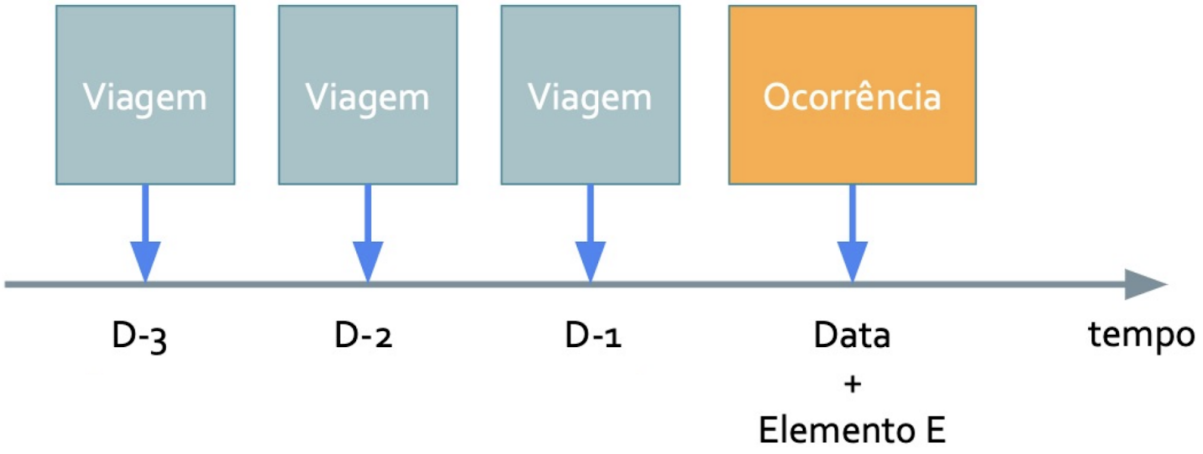


Figura 16 – Visualização da metodologia de filtro de dados do VI com base em data de ocorrências na via.

De posse dos registros do VI em regiões com falhas dentro de um intervalo de tempo, calcula-se a porcentagem de pontos agrupados no *cluster* com menos elementos no total. A justificativa é que, quanto maior a severidade, menor a sua frequência de ocorrência ao longo do tempo. Assim, conta-se a quantidade de elementos para cada *cluster* na totalidade da base de dados, e depois o filtro de falhas é aplicado. Com isso, o nível de maior severidade é avaliado de acordo com a proporção dos pontos do VI agrupados com o *cluster* menos popular. Um exemplo dessa metodologia é apresentado na Figura 17.

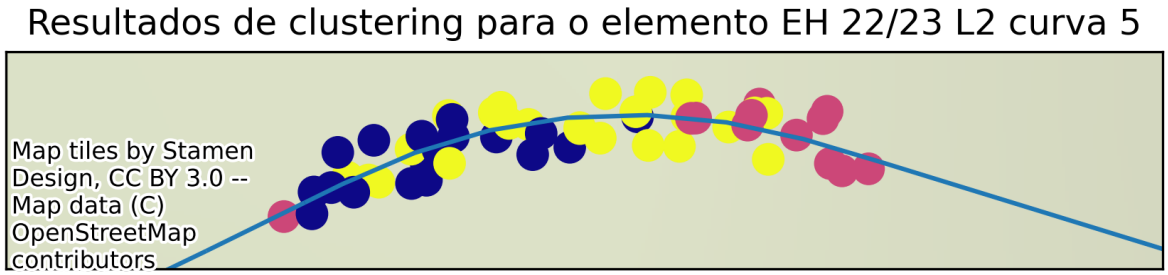


Figura 17 – Visualização da estratégia de validação para uma ocorrência registrada no elemento curva 23 L1 EH 81/82, em que cada ponto é uma leitura do VI, cuja coloração indica o *cluster* ao qual o registro foi agrupado. Neste exemplo, o *cluster* de cor rosa representa o nível de maior severidade, enquanto que as demais representam severidades mais amenas. Observam-se que apenas os dados do VI dentro desse elemento foram filtrados para a análise, e apenas aqueles dentro da janela de tempo definida (3 dias neste exemplo).

Na sequência, os resultados obtidos da estratégia de validação são comparados com algum nível de severidade previamente utilizado, para se medir se houve alguma melhoria em relação a um *benchmark* relevante. No início dos trabalhos da cátedra, havia sido estabelecido um critério simples e puramente estatístico projetado através de um método utilizado para a Estrada de Ferro Carajás pela Universidade de Monash para se definirem os níveis de severidade, denominados Níveis estatísticos (CÁTEDRA UNDERRAIL, 2021a). Esse é a base de comparação utilizada para essa etapa de avaliação do modelo.

Depois de encontrados níveis de severidades satisfatórios, o relacionamento entre a distância no tempo dos registros à data de defeitos é analisada com o emprego de algoritmos de aprendizado de máquina supervisionado. O objetivo é procurar obter mais diretamente um relacionamento entre as leituras brutas do VI e CC com a deterioração da via através da reutilização da estratégia de validação utilizada pelo *clustering* anteriormente. Assim, dados dos vagões em uma região de ocorrência anteriores ao fato dentro de uma determinada janela de dias são atribuídos a uma severidade; aumenta-se a janela de dias e outra severidade é atribuída; o processo se repete até que todos os níveis de severidade desejados tenham sido rotulados. Desta forma, espera-se que o classificador tenha bastante utilidade prática, na medida em que, se for bem sucedido, será teoricamente capaz de predizer a deterioração de uma região da via.

Nesta etapa, é interessante analisar dois aspectos do modelo treinado: a qualidade do treinamento e a capacidade de generalização. Para isso, adaptando o que foi discutido na Seção 2.2.2, após selecionado um conjunto de dados para treino, dois outros conjuntos de dados são captados: um para teste e outro para validação. O conjunto de testes será um semelhante à natureza do conjunto de treinamento, contendo dados de regiões e períodos

de tempo semelhantes ao de treinamento. Já o conjunto de validação traz dados mais diversos em relação ao *dataset* de treinamento, como em períodos ou regiões inéditas. É importante ressaltar que técnicas como validação cruzada são empregadas sempre no conjunto de treinamento, sempre preservando o ineditismo do conjunto de testes para apenas o término da etapa de treinamento. Esse processo é apresentado em detalhes pela Figura 18.

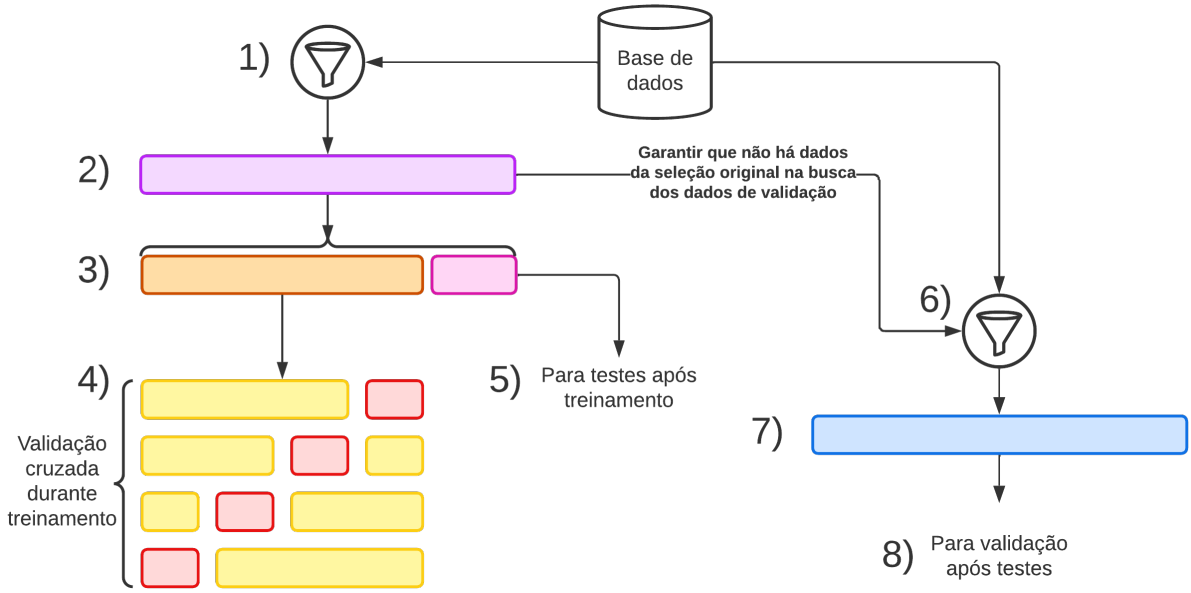


Figura 18 – Metodologia para separação de dados entre conjunto de treinamento, teste e validação. Primeiro, realiza-se um filtro em cima do banco de dados, normalmente utilizando as dimensões tempo e espaço (1). Depois, a partir desses dados filtrados pela busca (2), uma amostragem aleatória é separada (3), e o restante é utilizado para treinamento do modelo, com o emprego de técnica de validação cruzada (4). Uma vez treinado o modelo, a amostragem separada anteriormente é utilizada para testar o modelo (5). Por fim, uma nova busca é realizada no banco, dessa vez com parâmetros distintos daqueles da busca original (6), onde um novo conjunto de dados (7) é utilizado para validar o modelo (8).

Para exemplificar, se num dado experimento, os dados do CC dos meses de Janeiro a Abril entre as RH 20 e 65 forem escolhidos para treinar o modelo, então uma amostragem aleatória desses dados é separada para testes, e uma nova busca por dados do CC para a mesma região, mas entre os meses de Maio a Agosto são escolhidos para validação. Dessa forma, tem-se uma avaliação ampla do modelo em relação a seu aprendizado e eventual *overfitting*.


Figura 19 – Situação temporal dos dados em estudo, em que quanto mais escuro o ponto, maior a quantidade de dados registrados naquele momento. Regiões em cinza representam ausência de dados.

Conforme a Figura 19 ilustra, os dados disponibilizados para o projeto cobrem uma faixa de tempo consideravelmente estreita para o domínio em questão, em que defeitos geométricos levam anos para se formar, o que restringe a utilização de modelos restritos espacialmente. Ou seja, devido à quantidade e variedade dos dados, não é possível o desenvolvimento de modelos especializados em pequenas regiões da via, sendo necessário a busca por grande generalizações para curvas e tangentes.

Como prova final da busca pela definição de um relacionamento entre falhas e leituras, experimentos de aprendizado de máquina com algoritmos de regressão serão empregados para os dados do CC, uma vez que, diferente do VI, o CC faz leituras diretas sobre parâmetros de qualidade dos componentes da ferrovia em si. Esse modelo deverá predizer diretamente a distância em dias da ocorrência de uma falha dado as leituras dos sensores, virtualmente tornando contínuo os níveis discretos de severidade. A contribuição de cada variável para a regressão calculada será avaliada, bem como a capacidade do modelo de generalizar diferentes defeitos de via.

## 3.3 Visualização de dados

A cátedra de estudos dos dados do Vagão Instrumentado e Carro Controle se propôs a construir uma ferramenta unificada de gerenciamento e visualização dos dados do projeto, denominada Datamap (CÁTEDRA UNDERRAIL, 2021b). O protótipo criado foi fruto de um Trabalho de Conclusão de Curso (SAITO; RIBEIRO, 2021), em que os autores foram responsáveis pelo desenvolvimento de uma aplicação *web* capaz de exibir a região da EFVM, bem como seu trajeto, relevo e imagens de satélite. Além disso, a

plataforma também apresenta simulações de gráficos de métricas de variáveis importantes sobre os dados capturados pelos vagões, embora tenha utilizado dados aleatórios para apresentar como prova-de-conceito.

Consequentemente, tem-se uma base de uma plataforma própria para visualização dos resultados de experimentos de aprendizado de máquina, tornando possível a construção de interações mais complexas totalmente adaptáveis às necessidades vigentes. Assim, como parte da metodologia do presente trabalho, tem-se a elaboração de novas funcionalidades para que o Datamap possa efetivamente cumprir com as necessidades das investigações da cátedra. Entre elas, destacam-se:

- Integração com dados do Vagão Instrumentado;
- Integração com dados do Carro Controle;
- Integração com dados de falhas e ocorrências;
- Exibição clara dos níveis de severidade obtidos;
- Construção de gráficos das principais variáveis medidas por ambos os veículos.

É importante ressaltar a relevância da plataforma enquanto solução de visualização dos resultados dos modelos de aprendizado de máquina. Para que um resultado obtido seja verdadeiramente útil, é necessário que ele seja acessível aos impactados por ele. No caso, uma vez que se propõe a construção de modelos computacionais para a definição de níveis de severidades a partir dos dados do VI e CC, é fundamental que os operadores de campo possam ter uma maneira de acessar esses resultados para as inspeções *in loco*. Assim, a construção de uma ferramenta *web*, que pode ser acessada remotamente a partir de um dispositivo móvel com acesso à internet, como um celular, por exemplo, vem também com o propósito de auxiliar nesse requisito. Esse processo é apresentado na Figura 20.

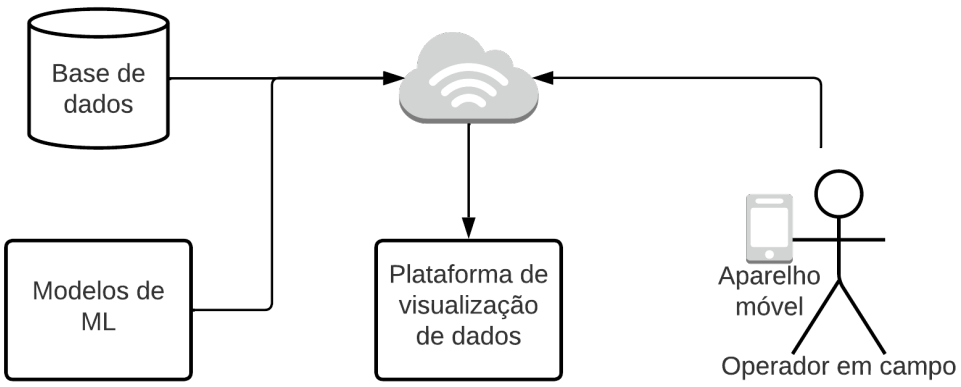


Figura 20 – Papel da plataforma de visualização de dados enquanto meio de acesso aos resultados dos modelos de aprendizado de máquina dos operadores em campo.

# 4 Resultados

## 4.1 Gestão de dados

No contexto dos experimentos de Ciência dos Dados que utilizam técnicas de aprendizado de máquina para identificar limites de severidade da ferrovia através do comportamento dos dados coletados pelos veículos de controle na via férrea, faz-se necessário a construção de um modelo de dados que abstraia algumas características relevantes, bem como forneça uma interface unificada independente de sua estrutura original. Conforme discutido na Seção 2.1, a importância e serventia dos dados dependem diretamente da forma em que são organizados, armazenados e acessados. Determinadas circunstâncias do estudo, discutidas a seguir, demandaram a implementação de um banco de dados.

A disponibilidade dos dados do VI e CC não ocorreu de forma contínua ou uniforme, de forma que houve a necessidade de realizar a integração com dados em diferentes formatos, disponibilizados em diferentes períodos do projeto. Sem uma abstração criada para a modelagem de dados do projeto, todo e qualquer novo dado necessitaria de uma atualização de todo o sistema, o que inviabiliza o perfil de uma aplicação de *Big Data*. Além disso, o cálculo das variáveis importantes para negócio, dados intermediários de cálculo como o LRS, a representação geométrica da EFVM e eventuais resultados de classificação dos estudos se mostraram tarefas computacionalmente caras de serem realizadas sob demanda, além de serem comuns de serem aplicadas a todo e qualquer novo pedaço de informação nova. Outro motivo importante que leva à construção do banco de dados é os formatos de entradas dos dados. Arquivos KML e XLSX, embora bastante amigáveis para uso humano, são computacionalmente mais intensivos para a leitura, o que também prejudica a constante execução dos algoritmos que os consomem. Por fim, outros experimentos com diferentes finalidades podem reutilizar esse conjunto de dados.

Assim, elaborou-se um *pipeline* de absorção e transformação dos dados no seu formato bruto, passando para o modelo gerado no banco de dados. Com isso, todo e qualquer novo experimento, seja ele de aprendizado de máquina ou apenas de análise exploratória dos *datasets*, passa a consumir um conjunto de dados já tratado e uniformizado, com todas as variáveis mapeadas e conhecidas. A plataforma de visualização dos dados também se beneficia desse banco de dados, passando a depender do esquema de dados definido para a implementação das visualizações das variáveis. A Figura 21 ilustra esse processo.

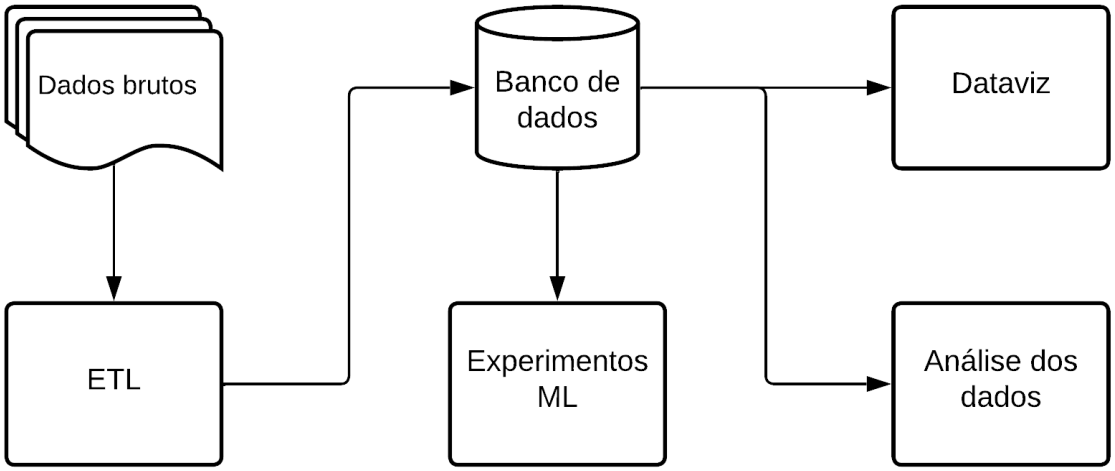


Figura 21 – Processo de ingestão, transformação e consumo dos dados simplificado.

### 4.1.1 Ingestão de dados

Uma das primeiras e fundamentais etapas de um projeto de Ciência dos Dados é a ingestão dos dados. No projeto, os *datasets* trabalhados foram os obtidos pelos Vagão Instrumentado, Carro Controle, os registros de falhas e defeitos da via, além de um cadastro dos elementos da EFVM e a representação geométrica da ferrovia. É importante salientar que esses dados de monitoramento têm processos e períodos distintos de aquisição de dados.

#### 4.1.1.1 Vagão Instrumentado

Os dados do Vagão Instrumentado foram disponibilizados em diferentes períodos ao longo da pesquisa. De início, havia um conjunto restrito e tratado *a priori*, contendo registros de alguns meses de 2020. Assim, nesse primeiro momento, foi realizada a caracterização de cada atributo medido, bem como a presença de cada uma quanto ao conjunto total de dados. Com isso, estabeleceu-se um formato padrão e base para a modelagem desse conjunto de dados.

Ao longo da pesquisa, novos dados de diferentes viagens do Vagão Instrumentado foram disponibilizados. Cada novo conjunto foi fornecido em diferentes formatos de arquivos, de forma que cada processamento precisou ser especializado para cada entrada.

O primeiro conjunto de dados foi disponibilizado em um arquivo *SQLite*, em que se foi necessário fazer uma busca pela tabela daquele banco de dados com os dados desejados. Comparando as variáveis disponíveis com o *dataset* original do VI, evidenciou-se a ausência da informação de linha férrea na dimensão de localização, assim como também de leituras

de deslocamento filtrado de sensores. Esse primeiro é um dado fundamental para a correta localização de um registro de uma leitura do VI, uma vez que tanto a quilometragem quanto o LRS a ser calculado apenas indicam a altura que o vagão se encontra na via, mas nada determinam sobre qual das duas (ou três, em alguns pontos específicos da EFVM) linhas paralelas ele se encontra. Isso se dá porque o registro de posicionamento via GPS não tem precisão suficiente para fazer essa distinção com confiança, já que as linhas têm distância uma da outra de cerca de 3 metros. Já os deslocamentos filtrados, variáveis pré-processadas em cima das leituras brutas de deslocamento, embora até então percebidos como dados obrigatórios, podem ser ignorados sem muito prejuízo da integridade do conjunto de dados, especialmente dado que os deslocamentos brutos estão presentes nesse primeiro conjunto.

O segundo conjunto de dados fornecidos se encontrara na forma de um conjunto comprimido de arquivos CSV, necessitando primeiro descompactá-los e, depois, ler arquivo a arquivo. Comparando os nomes das variáveis encontradas nos arquivos, observou-se que houve mudanças de nomes que necessitaram de confirmação para se ter certeza de que se tratavam ainda das mesmas grandezas. Além disso, diferente dos demais dados do VI até então, esse novo conjunto foi coletado por dois veículos distintos, o que é relevante ter em mente ao se agregar esses novos dados com os anteriores.

Independente da origem de cada novo conjunto de dados, alguns procedimentos de consistência de dados foram executados. Primeiro, a localização de cada registro foi verificada, de forma a garantir que apenas os dados referentes à EFVM fossem coletados. A Figura 22 ilustra um exemplo em que se observou que dados da Estrada de Ferro Carajás encontravam-se entre os dados brutos de um dos *datasets*.

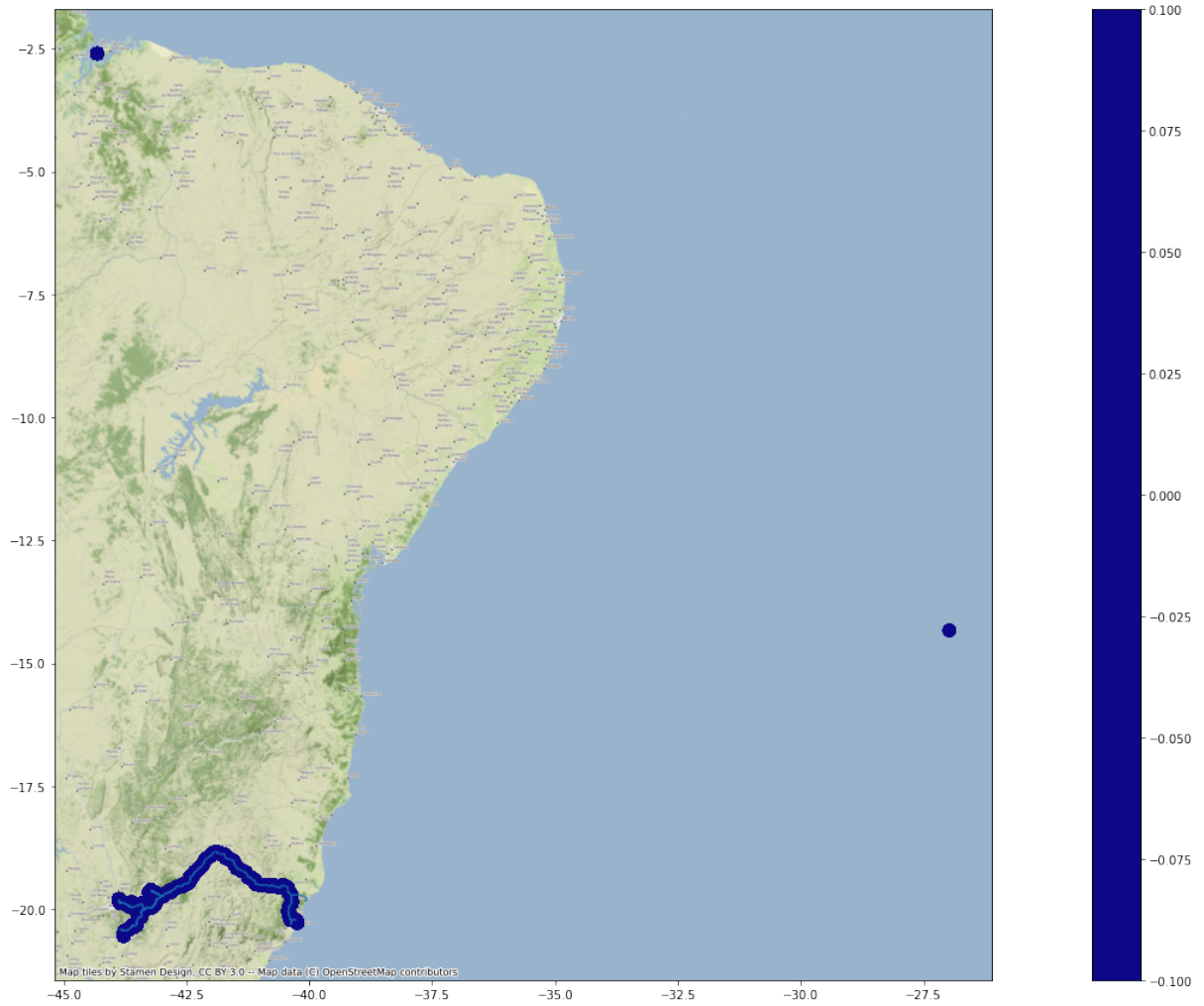


Figura 22 – Localização de novos registros do VI de um dos novos conjuntos de dados fornecido durante o decorrer do projeto. Observa-se que há pontos que não se referem à EFVM, bem como localizações inválidas.

Outra dimensão importante a ser avaliada é a temporal. Cada novo conjunto teve suas datas de registros analisadas, de forma a eliminar eventuais duplicidades com algum conjunto de dados já disponibilizado e tratado anteriormente. Como resultado dessa etapa, observou-se que o primeiro conjunto de dados contém dados de alguns meses do ano de 2019, e o segundo, dos cinco primeiros de 2022. Esta última observação explicaria a razão pela qual novos equipamentos foram utilizados para a coleta dos dados em campo, dado que o instrumento anterior teria sido utilizado nos anos de 2019 e 2020, sendo substituído depois de 2021.

Por fim, cada variável de cada novo conjunto de dados disponibilizado foi caracterizada através de estatística descritiva e comparada com seus dados anteriores, a fim de se ter certeza de que as leituras coletadas permanecem coerentes com suas grandezas. Eventuais discrepâncias observadas foram pontuais e descartadas. A Figura 23 ilustra um

exemplo desse procedimento feito para uma das variáveis.

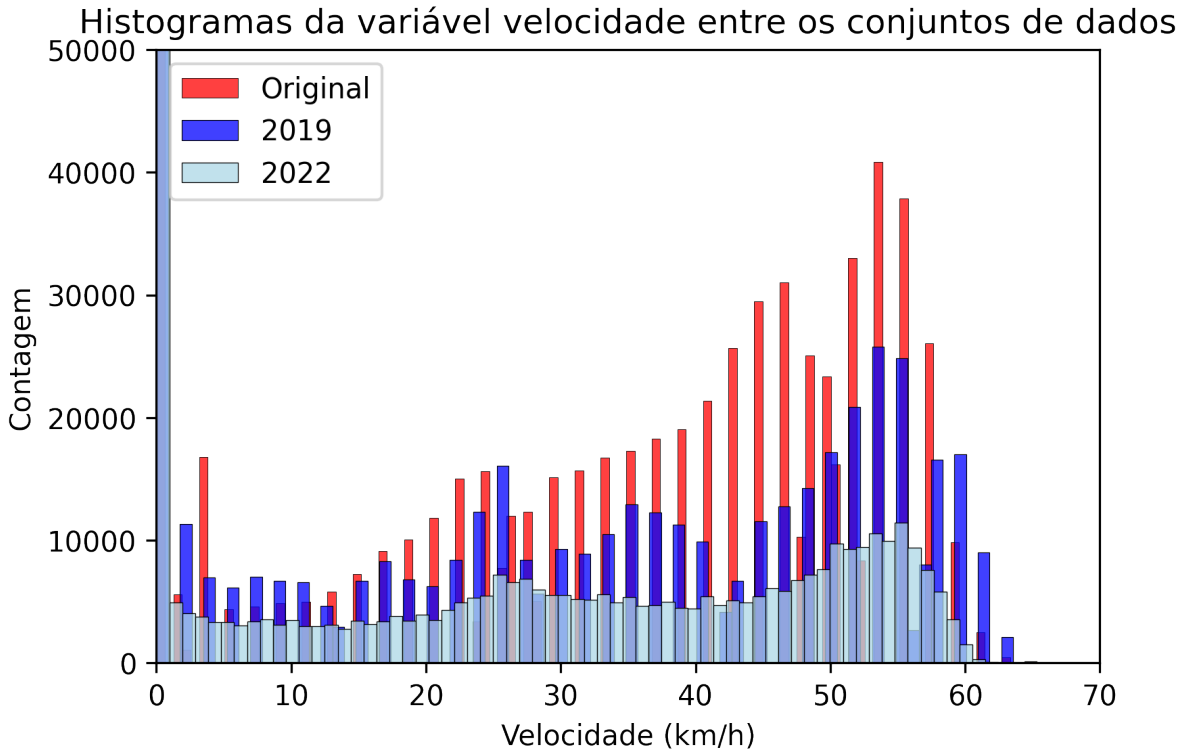


Figura 23 – Comparação de histograma de uma das leituras de sensores do VI entre o conjunto original dos dados e os novos *datasets* fornecidos. Para esse exemplo, conclui-se que os novos registros se adéquam ao comportamento esperado.

Uma vez validados os dados, para efetuar o carregamento deles no banco de dados com o modelo definido, foram gerados identificadores únicos para para registro com base nos seus atributos. Com isso, no caso de ainda haver algum registro duplicado, esse procedimento detectaria essas ocorrências e evitaria o carregamento repetido. Como resultado, foram processados 11.388.197 registros, resultando em aproximadamente 4 GB de dados.

#### 4.1.1.2 Carro Controle

Os dados do Carro Controle foram disponibilizados através de uma série de arquivos XLSX comprimidos. Para processá-los, foi necessário descompactar o arquivo raiz e processar cada arquivo individualmente, uma vez que alguns arquivos brutos abertos ao mesmo tempo consomem demasiada memória computacional, inviabilizando sua execução em computadores pessoais, ou eventualmente acarretando custos extras em ambientes de nuvem. Devido à quantidade de dados e aos cálculos necessários para a adequação ao modelo, várias horas de processamento foram necessários. Dessa forma, para evitar perdas em caso de interrupção imprevista, um sistema de controle de marcação de progresso foi

implementado para que se possa retomar do último ponto não persistido no banco de dados.

Assim como os dados do Vagão Instrumentado, também foi necessário validar as dimensões de espaço e tempo dos dados. Havia um pequeno lote de dados datados de 2005, e, como se sabe que o Carro Controle não existia nessa época, esses dados foram descartados. Também observou-se outro subconjunto de dados cuja localização registrada não se encontrava dentro do trajeto da EFVM, conforme ilustra a Figura 24.

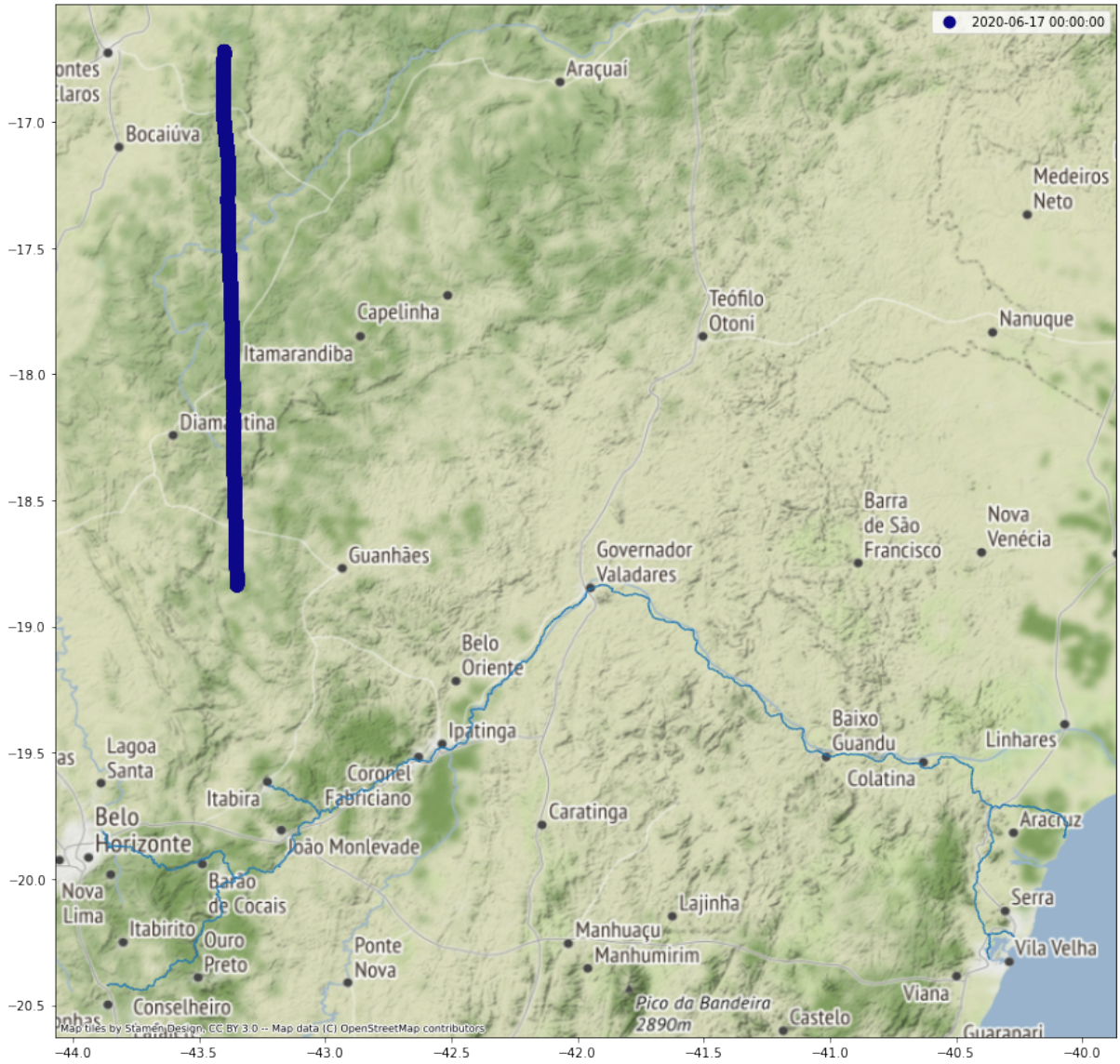


Figura 24 – Localização de registros do CC de um lote específico, após a aplicação de um filtro para selecionar dados muito distantes da EFVM.

Como resultado da ingestão de dados do Carro Controle, foram armazenados 35.707.947 registros, resultando em cerca de 30 GB de dados.

#### 4.1.1.3 Falhas

Os dados de defeitos de via foram obtidos através de uma planilha de ocorrências disponibilizada. Esta planilha não contém apenas dados relevantes aos estudos de falhas geométricas, incluindo também registros de vandalismo, operações de apoio e outras diversas rotinas de inspeção. Assim, um primeiro passo para a ingestão desse conjunto de dados foi o cruzamento com uma segunda planilha fornecida de filtro, que contém os identificadores dos registros da planilha original que dizem respeito aos defeitos de interesse. Esse filtro, porém, foi construído a partir de palavras-chave e alguns padrões de registros observados por especialistas da área, então é interessante manter todos os registros de ocorrências, mas deixando a distinção clara entre ter sido filtrado ou não. Dessa forma, uma coluna de classificação bruta entre "falha"e "outros"foi criada para permitir eventuais novos tratamentos futuros.

Os registros das falhas contêm três escalas de localização da ocorrência: entre *house* (EH), elemento de via e quilometragem. A Figura 25 ilustra cada uma dessas diferentes escalas. Cada EH contém alguns elementos de via, e as quilometragens registradas são sempre menores em comprimento do que um elemento individual. Tem-se, então, uma variação do nível de escala de precisão de localização dos registros. Porém, a disponibilidade de cada uma dessas escalas não é regular; pelo contrário, vários registros têm apenas uma ou duas disponíveis, sem nenhuma preferência entre elas. Além disso, foi observado que há casos de escalas de localização conflitantes, ou seja, a informação de uma escala não é condizente com a localização de outra escala menor. Um exemplo de situação observada é apresentada na Figura 26. Portanto, também foi incorporado ao processo de ingestão desses dados uma verificação de coerência entre essas informações, dando preferência àquelas apontadas pelos especialistas como a mais confiável.


Figura 25 – Todas as diferentes escalas de localização de falhas em cores: EH em vermelho, elemento de via em laranja, e quilometragem em amarelo. Os pontos azuis indicam RHs.

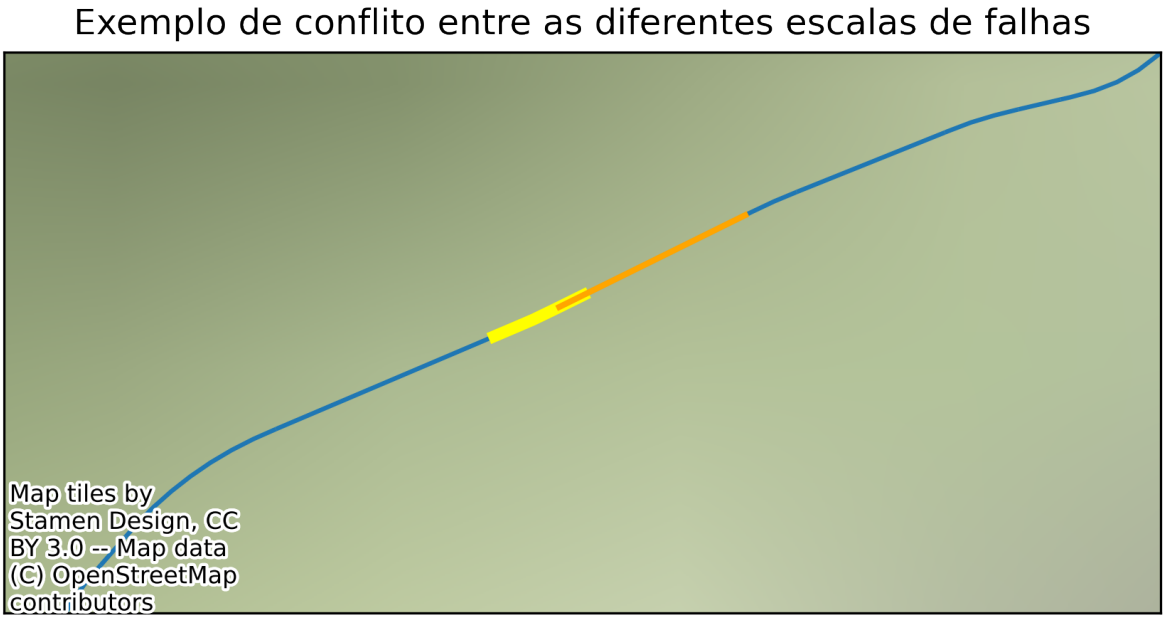


Figura 26 – Exemplo de conflito entre as escalas de geolocalização de uma falha. Em laranja, a escala de elemento de via e em amarelo, quilometragem. Há uma região de convergência, mas grande parte das duas informações não coincidem.

Para a uniformização das informações de localização através das diferentes escalas, o LRS foi empregado: para elementos e EHs, o conjunto de dados seguinte de referência da via foi utilizado e, para quilometragem, a função transformadora de quilometragem para LRS construída na Seção 4.1.2.1 foi utilizada.

#### 4.1.1.4 Elementos e representação geométrica da EFVM

Como dados de referência para cruzamento de informações entre os diferentes *datasets* principais descritos anteriormente, foi disponibilizado uma planilha de elementos da EFVM. Essa planilha enumera todos as curvas, tangentes, pontes, JICs e RHs da via, acompanhados de suas coordenadas GPS.

Além disso, também foi fornecido um arquivo KML com a representação geométrica da EFVM em coordenadas geolocalizadas, isto é, os mais diversos segmentos de reta que compõe a ferrovia. Esse desenho é utilizado como base do LRS, configurado como base de referência para todos os pontos nesse sistema.

Como resultados da ingestão, têm-se os dados representados pelo Modelo Entidade-Relacionado apresentado pela Figura 27.

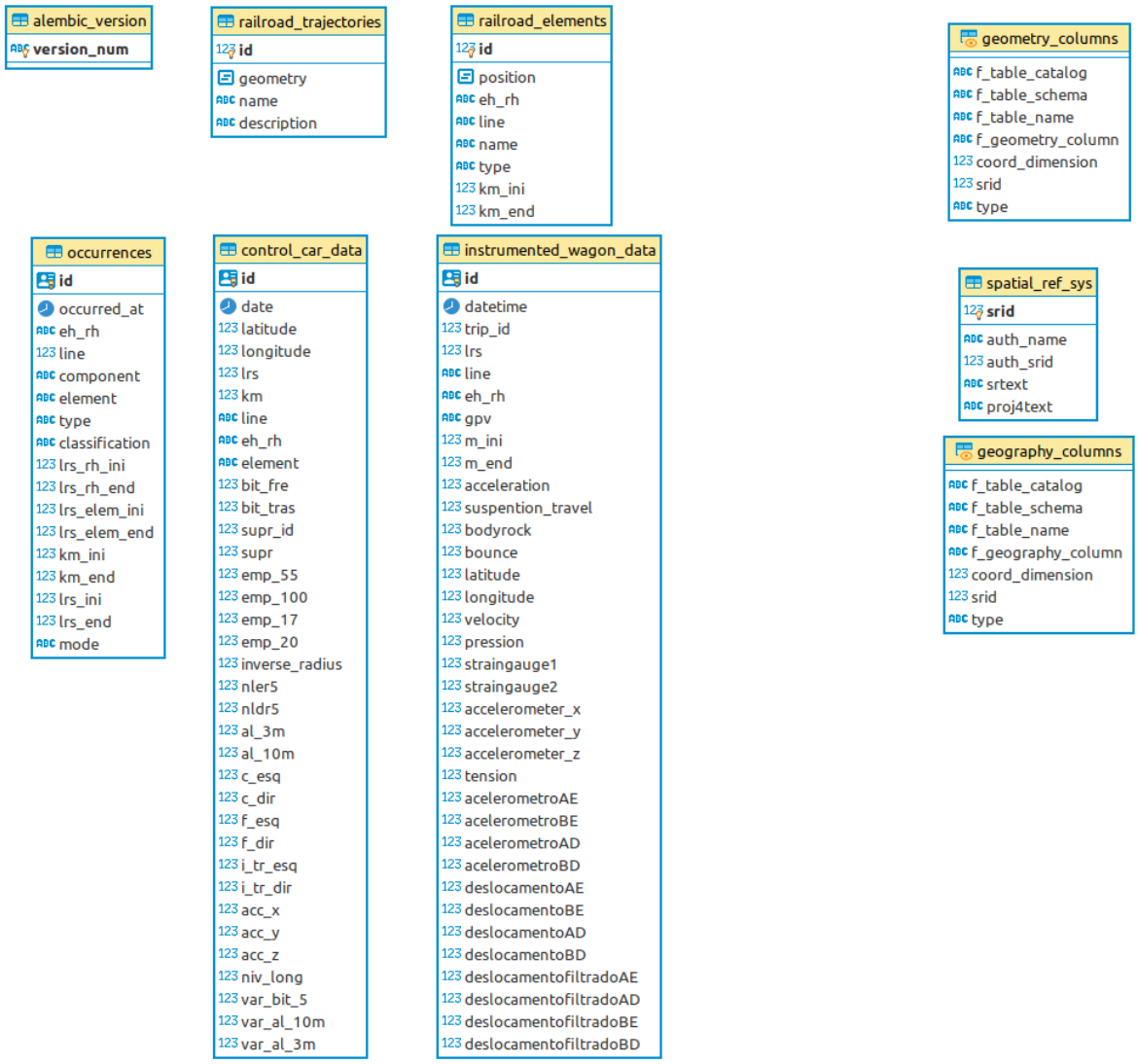


Figura 27 – Modelo Entidade-Relacionamento criado para representar todos os dados do projeto no banco de dados.

### 4.1.2 Sistema linear de referência

#### 4.1.2.1 Inferência de coordenadas do VI com base no LRS

Para a inferência de coordenadas de dados do Vagão Instrumentado carentes de georreferenciamento, aplicando a estratégia apresentada na Seção 3.1.1.1, obtém-se que apenas 3 viagens, dentro um total de 14 viagens em 2020, apresentam quilometragem inconsistente.

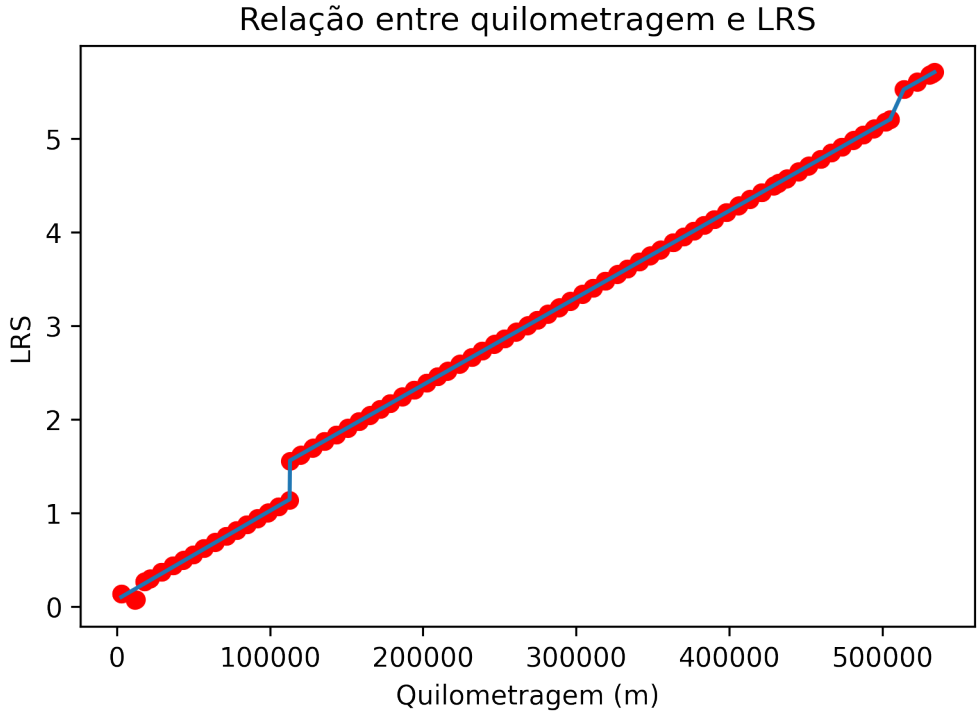


Figura 28 – Polinômio construído para conversão de quilometragem para LRS. No eixo $x$, tem-se a quilometragem e no $y$, LRS.

Pode-se ver na Figura 28 que de fato há uma clara relação linear entre a quilometragem e o LRS. Não é de se surpreender essa relação, dado que a quilometragem nada mais é do que um sistema linear próprio.

Vale destacar aqui as descontinuidades observadas na relação quilometragem-LRS mostrada anteriormente. Isso ocorre devido às ramificações da EFVM, em que o LRS cobre a totalidade da via, incluindo, portanto, essas ramificações. Como a quilometragem não as cobre, e apenas segue direto, então é necessário fazer esse ajuste. Para resolver esse problema, construíram-se funções polinomiais em cada intervalo contínuo, formando uma função final descontínua, conforme ilustrado pelo gráfico da Figura 28.

De posse da função de conversão de quilometragem para LRS, os dados de GPS dos registros dessas viagens do Vagão Instrumentado foram inferidos com êxito, podendo agora fazer parte da base de dados para os estudos.

É importante ressaltar que, como subproduto desse processo, tem-se agora um conversor de quilometragem para LRS. Esse conversor pode ser utilizado como um mecanismo unificador de novos dados, apresentando-se como uma ferramenta valiosa para o tratamento de dados com sistemas de referência distintos.

#### 4.1.2.2 Análise de sobreposição de ramificações

Por conta das ramificações da EFVM, o LRS pode ocasionar em alguns filtros problemáticos quando existir sobreposição de ramos; ou seja, quando uma ramificação da linha começar junto com o "trajeto principal". Nesses casos, a precisão do GPS não é suficiente para garantir que a posição do VI fique consistentemente num mesmo ramo, e essa flutuação entre um ramo e outro ocasiona em saltos no LRS, fazendo com que eventuais filtros não funcionem como esperado.

Na Figura 29 a), observa-se que o início da ramificação que segue a leste começa junto com o ramo que segue a oeste, e, por uma pequena porção do trajeto, ambos os percursos coexistem bem próximos um do outro. Assim, qualquer localização de GPS nesse trecho com os dois ramos pode facilmente flutuar entre estar no trecho à direita ou à esquerda. Isso não é ideal porque os valores do LRS em cada um dos trechos serão distintos.

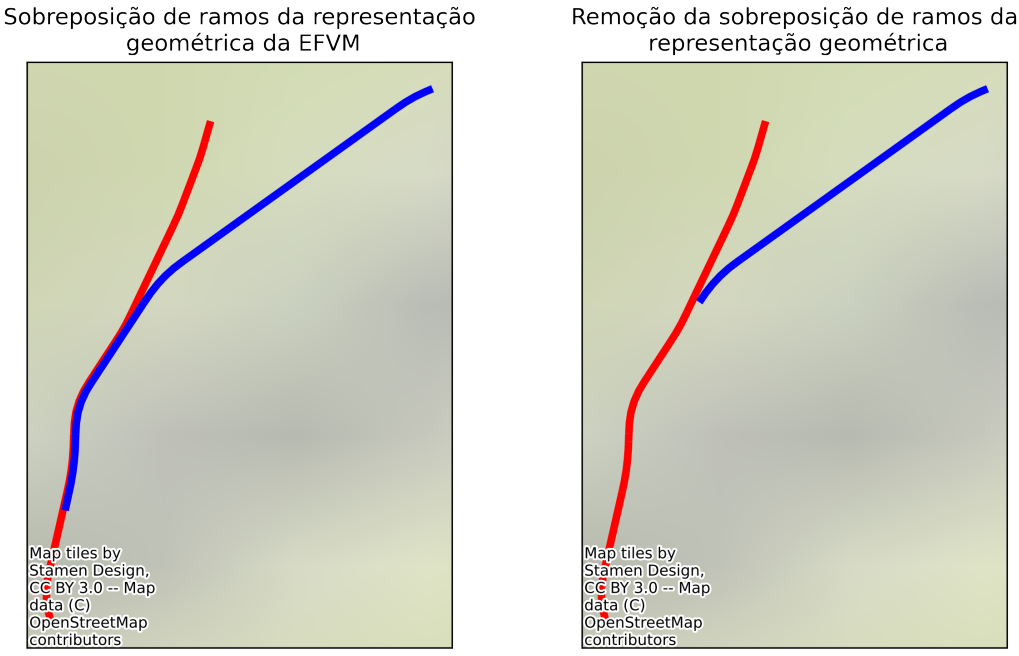


Figura 29 – a) Uma das ocorrências de sobreposição de ramificações da EFVM no modelo do LRS. b) Mesmo trecho da EFVM com a sobreposição de ramos removida.

Para resolver esse problema, todas as ramificações da EFVM foram analisadas, e os segmentos de reta do modelo que compõe cada sobreposição foram removidos, de forma a se obter um resultado tal qual ilustrado pela Figura 29 b).

## 4.2 Análise exploratória dos dados

### 4.2.1 Vagão Instrumentado

O primeiro passo da análise exploratória dos dados do VI é conhecer todas suas variáveis: quais os tipos de dados de cada uma, suas distribuições e comportamentos. A Figura 30 ilustra alguns histogramas das variáveis do VI oriundas dos sensores, apresentadas na Tabela 2. Com isso, tem-se um entendimento mais aprofundado do conjunto de dados em questão, determinando a primeira dimensão do modelo multidimensional discutido na Seção 3.1.3.

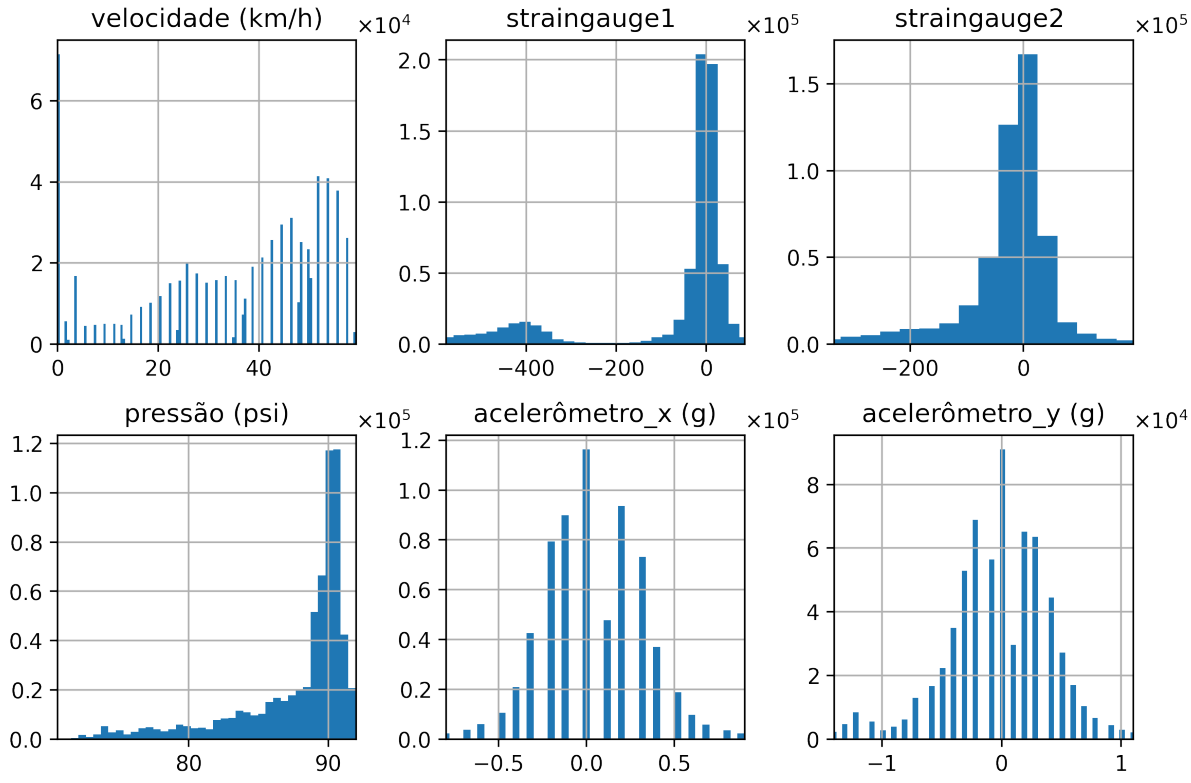


Figura 30 – Gráficos de distribuição das variáveis de interesse do *dataset* do VI.

Na sequência, é fundamental conhecer as demais dimensões do modelo multidimensional: o espaço e o tempo dos dados. Esse aspecto é importante na medida em que não apenas se tem um conhecimento importante acerca do *dataset* em questão, como também se estabelecem as possíveis relações com os demais conjuntos de dados. Dessa forma, explora-se a cobertura temporal e espacial dos dados do VI, averiguando a parcela da EFVM que o VI atua, e em que meses e anos estão disponíveis. A Figura 31 ilustra a cobertura espacial e a Figura 32, a espacial.


Figura 31 – Cobertura do VI sobre a EFVM. Os pontos em azuis representam uma captura de dados. O traçado em vermelho representa a EFVM.

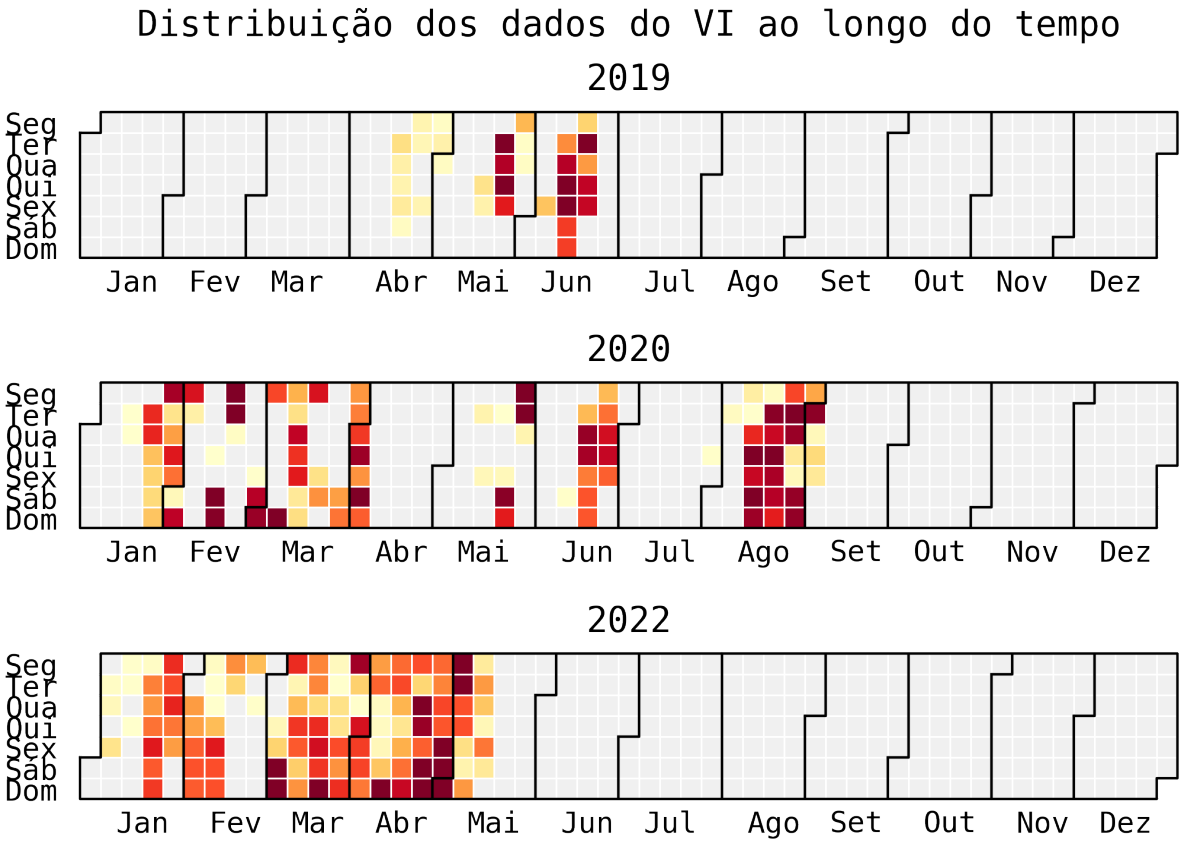


Figura 32 – Cobertura temporal dos dados coletados pelo VI. Quanto mais escuro um ponto, maior a quantidade de registros feitos naquele dia.

Uma vez bem conhecida a natureza dos dados do VI, é crucial explorar o relacionamento que esse *dataset* tem com os dados de falhas e ocorrências na via. Isso é outra necessidade do modelo multidimensional proposto, já que a distância de um registro de veículo para uma falha no tempo é o fator chave para os modelos de aprendizado de máquina. Assim, analisando essa dimensão em conjunto com os defeitos registrados, podem-se observar alguns comportamentos interessantes. Cruzando cada registro do VI com uma falha futura na mesma região, observa-se que há considerável correlação entre a proximidade do registro à ocorrência com algumas variáveis. Essas correlações são apresentadas pelas Figuras 33 e 34.

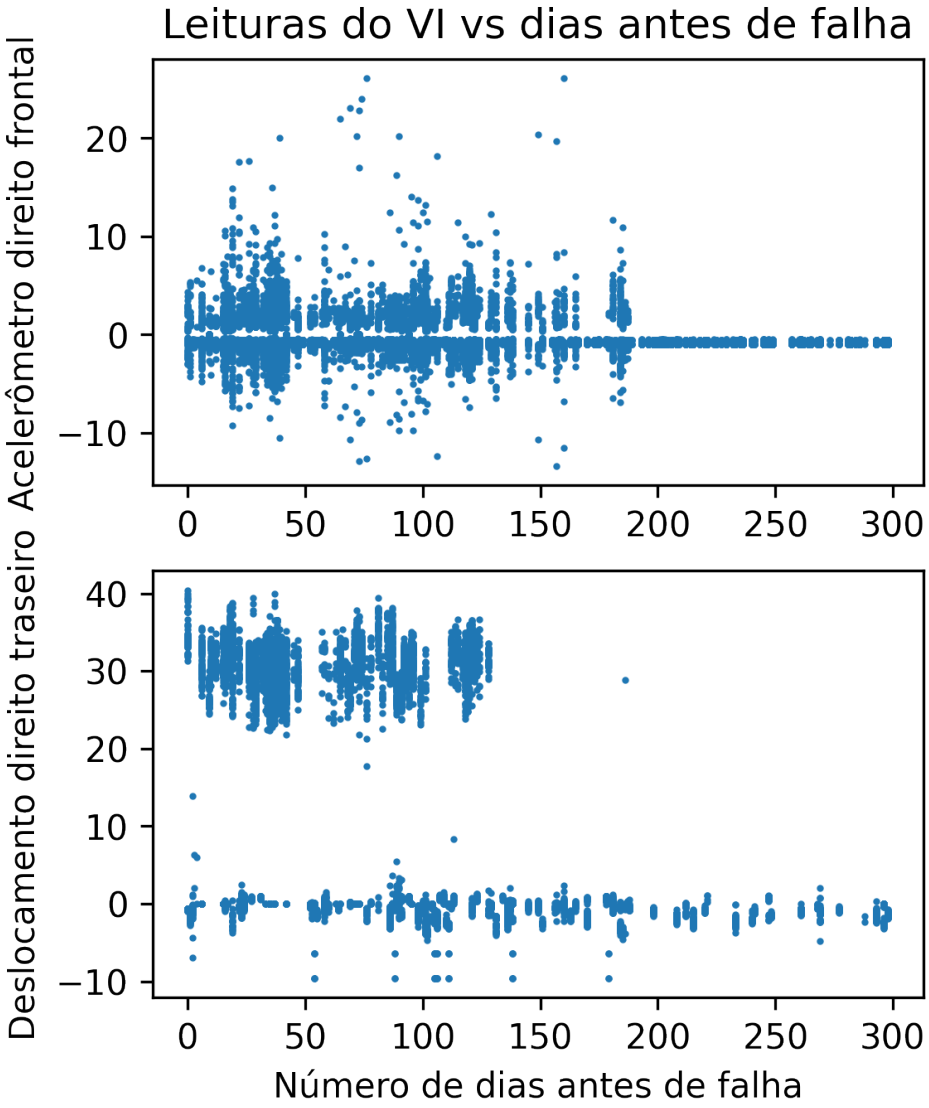


Figura 33 – Variação do deslocamento traseiro direito e acelerômetro dianteiro direito respectivamente com o número de dias anteriores à falha registrada no mesmo local.

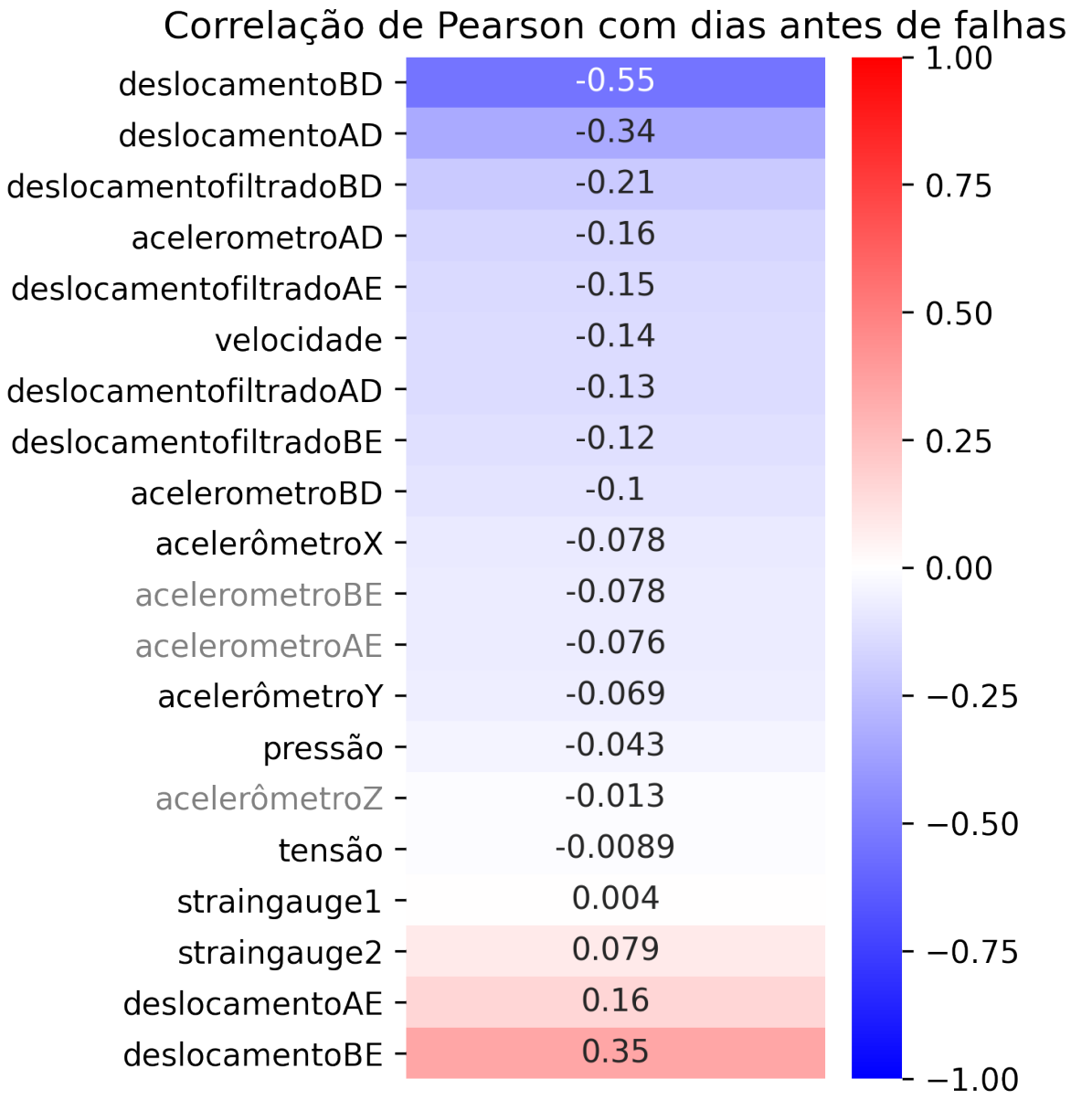


Figura 34 – Mapa de calor de correlação de Pearson entre as medidas do VI e os a quantidade de dias antes de falha. As variáveis em cinza não apresentaram significância estatística, calculada através do coeficiente de correlação de Pearson, e avaliada em relação ao valor 0,05 (valores menores são considerados estatisticamente significantes). As demais em preto passaram nesse teste.

### 4.2.2 Carro Controle

Depois de feita a análise exploratória dos dados do VI, repetiram-se os mesmos estudos em cima dos dados disponibilizados do CC. Assim, de início, foram identificadas as variáveis de interesse do *dataset*, apresentadas na Tabela 3, bem como a distribuição das variáveis de sensores, ilustrada em forma de histogramas na Figura 35. Com isso, tem-se bem definida a dimensão de sensores do modelo multidimensional do CC.

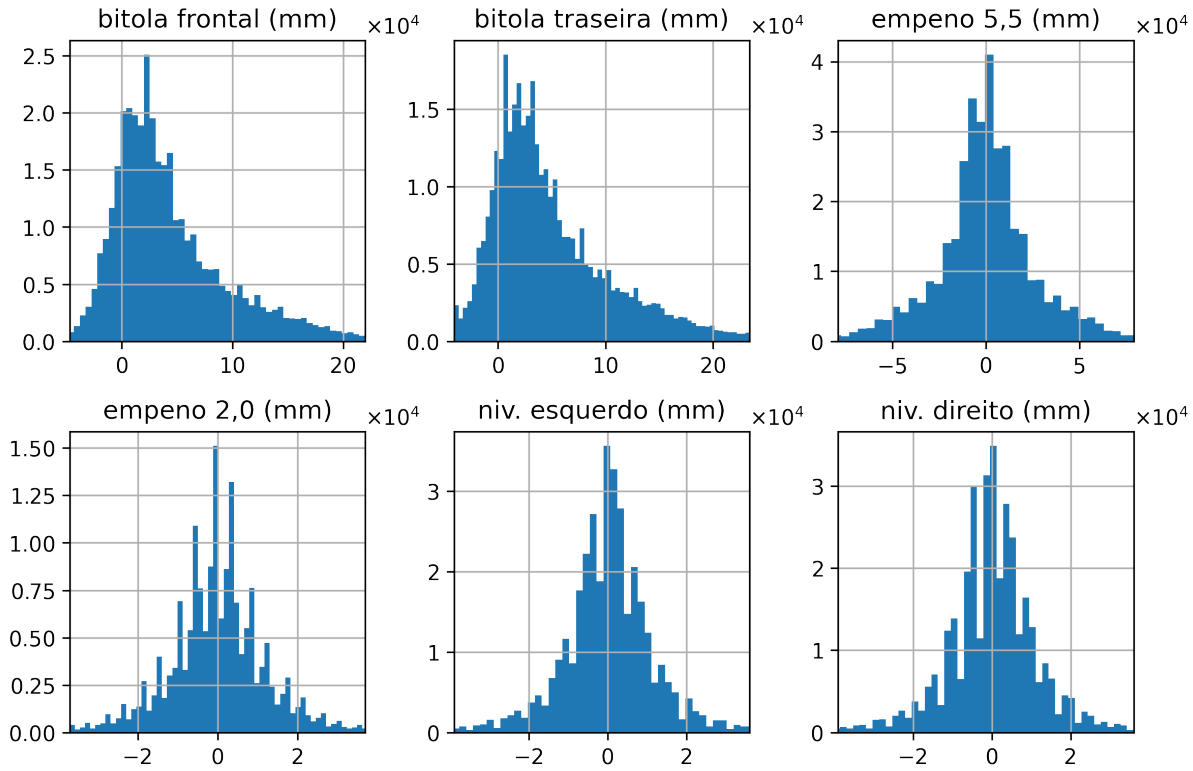


Figura 35 – Gráficos de distribuição de algumas variáveis de interesse do *dataset* do CC.

Um comportamento importante a ser analisado com especial atenção no caso do CC é a presença das variáveis de registros de sensores do CC, e como isso varia ao longo do tempo. Foi observado que, ao contrário do VI, onde praticamente todos os dados de sensores são constantes ao longo das viagens do veículo, o CC possui algumas leituras que só estão presentes nas últimas viagens do conjunto de dados. É possível e, conforme a Figura 36 mostra, provável que as métricas coletadas pelo veículo foram evoluindo com o passar dos meses, de forma que algumas medidas não estão disponíveis nas primeiras viagens.


Figura 36 – Proporção de valores ausentes para cada variável de leitura de sensor do CC. Observa-se que algumas estão sempre ausentes nas viagens anteriores a Julho, e outras tem um comportamento estável ao longo do ano.

Depois, seguindo o já executado com o VI, as dimensões seguintes do modelo, espaço e tempo, são mapeadas. A Figura 37 mostra que a cobertura espacial do Carro Controle é completa, cobrindo toda a EFVM. É interessante ressaltar que essa cobertura se mostrou inclusive superior à do Vagão Instrumentado, que não cobre todas as ramificações da via. Já a Figura 38 ilustra a distribuição dos registros do CC ao longo do tempo. Aqui, também é relevante comparar com a distribuição temporal do Vagão Instrumentado, o que leva à percepção de que há pouca coincidência entre eles. Isso se dá pela complementariedade a qual ambos os veículos são empregados no campo. A comparação pode ser observada na Figura 19.


Figura 37 – Cobertura do CC sobre a EFVM. Os pontos em azuis representam uma captura de dados. O traçado em vermelho representa a EFVM.

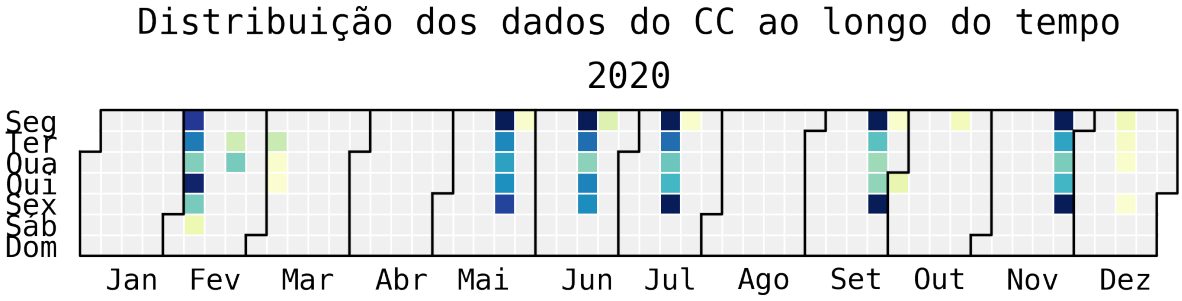


Figura 38 – Cobertura temporal dos dados coletados pelo CC. Quanto mais escuro um ponto, maior a quantidade de registros feitos naquele dia.

Diferente do VI, o CC efetua medidas diretas de componentes da via. Portanto, suas medições podem ser diretamente relacionadas com variadas métricas de avaliação da qualidade da via utilizadas tanto em normas quanto em campo. No caso da EFVM, a norma NBR 16.387/2016 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2016) foi utilizada como base para a elaboração de critérios de interdição e limitação de velocidade da via. Assim, é relevante conferir com esses critérios a qualidade aferida da via através do CC. A Figura 39 mostra os pontos de todo o *dataset* do CC em que algum desses limites foi atingido, e em qual data e região dessas ocorrências.


Figura 39 – Ocorrências de leituras do CC por data que ultrapassaram algum dos limites desenvolvidos pela Vale a partir da norma NBR 16.387/2016.

## 4.3 Aprendizado de máquina

### 4.3.1 Clustering

Nessa seção, discutem-se os resultados obtidos com a aplicação do algoritmo de aprendizado não supervisionado *K-means* nos dados de leituras do VI. Salvo onde especificado de forma diferente, todas as variáveis de leitura dos sensores modelados foram utilizadas para os experimentos, permitindo que o algoritmo fique o mais próximo possível da realidade captada.

#### 4.3.1.1 Resultados


Figura 40 – Visualização das leituras do VI em grupos encontrados pelo experimento de *clustering* em uma data específica, onde cada cor representa um grupo distinto. Os pontos do VI são ligeiramente deslocados de forma aleatória para visualização completa, devido à escala da figura.

Tendo o *dataset* sido devidamente filtrado para apenas os principais elementos da via (curvas e tangentes), o algoritmo foi executado repetidas vezes. Para se averiguar o resultado, foram empregados mapas com cada ponto representado por uma cor correspondente ao seu grupo atribuído. Como foram realizadas várias viagens do VI ao longo da via, cada data distinta foi exibida de cada vez. Nesse primeiro momento, procura-se por alguma região com clara tendência a um grupo sobre os demais, e como isso poderia indicar alguma falha. Um exemplo dessa visualização é apresentado pela Figura 40. Outro aspecto analisado inicialmente foi o comportamento das quatro principais variáveis calculadas em cada grupo, conforme exemplo da Figura 41.

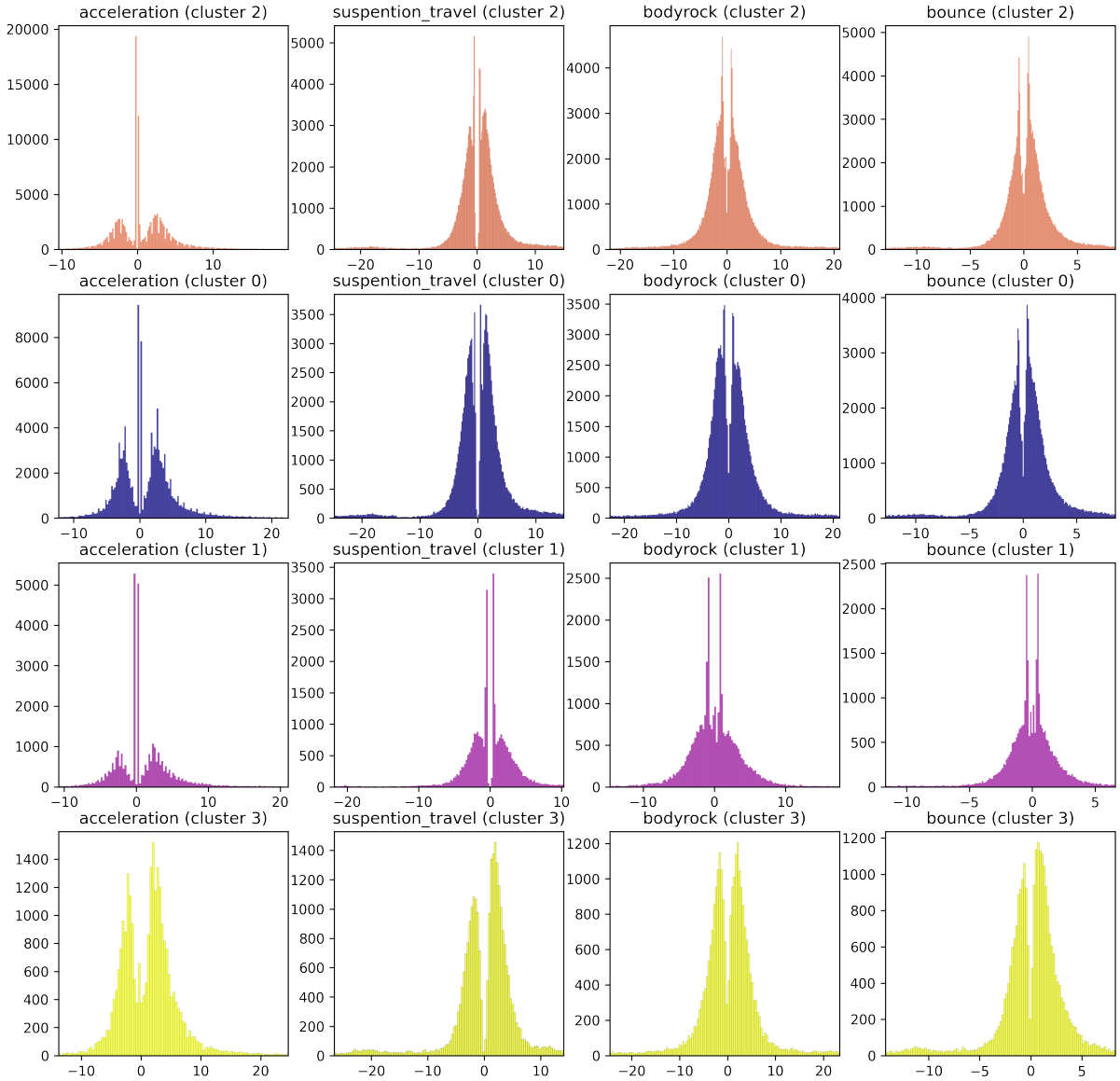


Figura 41 – Histogramas das quatro principais variáveis calculadas do VI para cada grupo encontrado no experimento com $K = 4$. Cada linha está ordenada em ordem crescente de nível de severidade, isto é, o cluster 2, nesse resultado, é a menor severidade e o cluster 3, a maior.

Uma dos desafios mais relevantes de experimentos de *clustering* é encontrar um número adequado para $K$, ou seja, a quantidade de *clusters* a serem encontrados. O *K-means* (e, de forma geral, os algoritmos mais comuns de aprendizado não supervisionado) sempre encontra os melhores *clusters* para um conjunto de dados fornecido, mas não faz parte de seu algoritmo procurar definir qual o melhor número para aquela amostra. Portanto, é necessário fazer um estudo à parte para se procurar definir esse número $K$.

Um dos métodos mais tradicionais para essa tarefa é fazer o gráfico do joelho, isto é, um gráfico em que o nível de similaridade dos *clusters* encontrados é medido em função

do $K$ utilizado. Naturalmente, a curva é sempre decrescente, uma vez que, quanto mais *clusters* são esperados, maior será a similaridade dentro de cada *cluster*. O que se espera observar é um ponto crítico em que há uma mudança nítida de melhoria de similaridade, representado, de forma gráfica, por uma mudança de variação na curva (daí o "joelho"no nome da técnica).

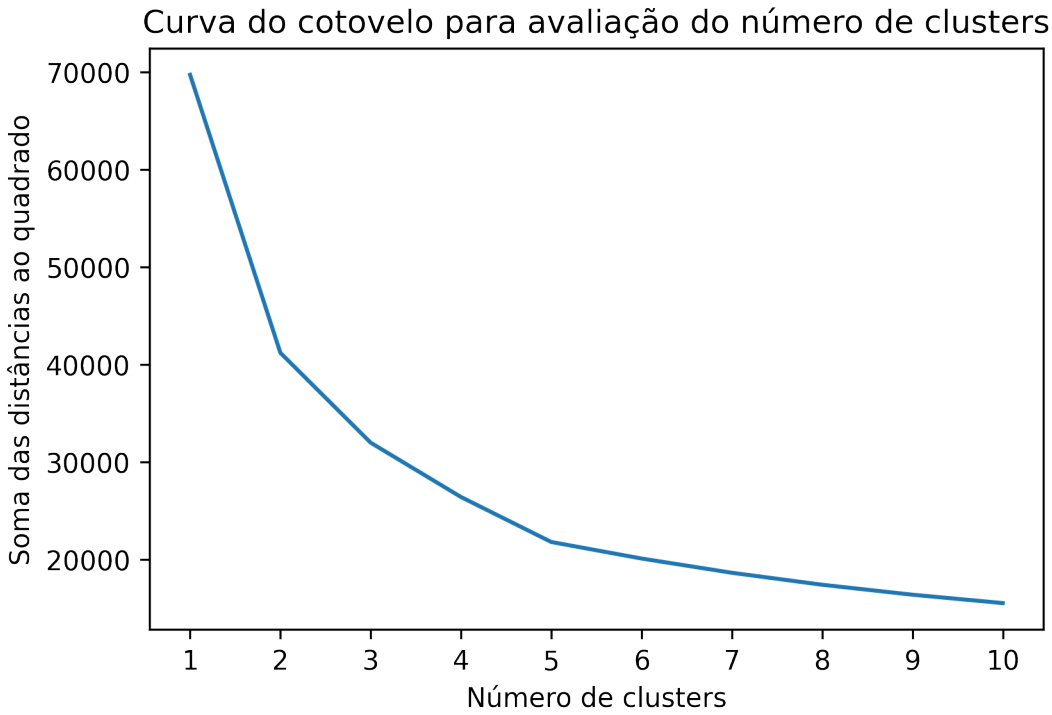


Figura 42 – Gráfico do joelho para avaliar o melhor $K$ para o conjunto de dados do VI.

Na Figura 42, nota-se que o $K$ ideal não é muito claro, estando em algum lugar entre 3 e 5. Portanto, apesar de se ter esse intervalo de confiança, ainda fica necessário o estabelecimento de técnicas de validação para se avaliar a qualidade dos *clusters* encontrados. Assim, de posse de uma estratégia de validação, pode-se repetir o mesmo procedimento: varia-se o $K$, e observa-se como a validação se comporta em função de $K$.

Uma hipótese importante a ser validada, mencionada na Seção 3.2, é se os eventuais *outliers* das variáveis estariam erroneamente influenciando o agrupamento feito pelo algoritmo. Assim, foram realizados experimentos exclusivos para avaliar o impacto dos *outliers* no *clustering*. Conforme discutido, foi utilizado o método IQR para essa tarefa. A Figura 43 ilustra esse procedimento para uma variável de exemplo.

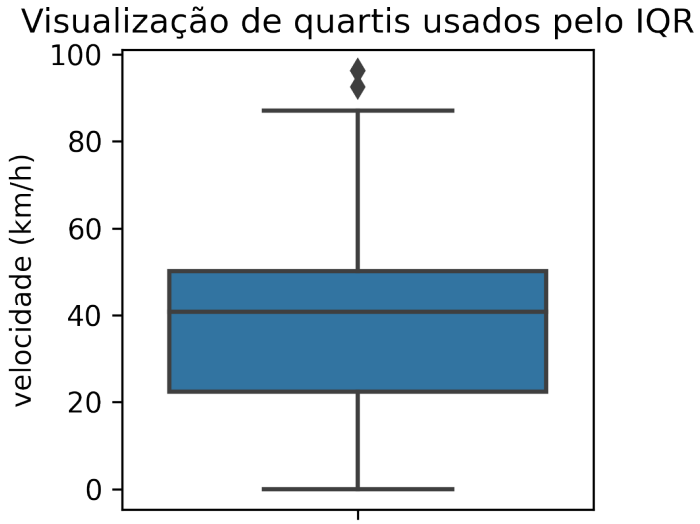


Figura 43 – Remoção de *outliers* através de IQR, onde a caixa azul representa os dados dentro do intervalo interquartil, e os pontos fora desta representam os *outliers*.

#### 4.3.1.2 Validação

De posse dos diferentes resultados de *clustering*, é necessário validar os *clusters* encontrados utilizando alguma estratégia de validação. Seguindo o que foi apresentado na Seção 3.2, é interessante avaliar como varia a frequência de cada *cluster* em função do tempo em relação a uma falha registrada, já que espera-se que os *clusters* representem níveis de severidade da deterioração da via.

Outro aspecto discutido anteriormente é a questão da validação do melhor $K$ para o experimento. Assim, as execuções do algoritmo de agrupamento e a subsequente estratégia de validação foram repetidas para o intervalo de $K$ previamente identificado pelo método do joelho como mais promissor.

Em relação à remoção de *outliers*, testes foram realizados variando apenas a execução ou não dessa etapa. Como resultado, observou-se que a não remoção proporcionou resultados significativamente melhores para todos os $K$. Os resultados apresentados a seguir, portanto, foram obtidos sem nenhum *outlier* filtrado.

Por fim, também é importante comparar com algum nível de severidade previamente utilizado, para se medir se houve alguma melhoria em relação a algum *benchmark*. Conforme apresentado na metodologia, um critério simples e puramente estatístico, denominados Níveis estatísticos, foi empregado para se definir os níveis de severidade no início dos trabalhos a partir do proposto pela Universidade de Monash. Assim, esse será a base de comparação utilizada nos resultados.

Os resultados da validação da maior e da segunda maior severidades, denominadas respectivamente severidade 1 e 2, estão presentes nas Tabelas 5 e 6. Para o nível inter-

mediário de severidade 3 foram aplicadas estratégias de validação das associações por simulações com a variação ao longo do tempo.

Algumas métricas utilizadas para aprendizado supervisionado, apresentadas na Seção 2.2.2, como *recall* e acurácia, foram empregadas para a avaliação de desempenho dos ensaios de *clustering*.

| | *Recall* | Acurácia |
|---|---|---|
| Níveis estatísticos | 0,47% | 96,24% |
| $K = 3$ | 6,54% | 88,29% |
| $K = 4$ | 21,03% | 91,44% |
| $K = 5$ | 21,50% | 91,87% |
| $K = 6$ | 2,80% | 96,11% |

Tabela 5 – Valores de *recall* e acurácia para as diferentes simulações realizadas com registros do VI até 3 dias antes de falha (severidade 1).

| | *Recall* | Acurácia |
|---|---|---|
| Níveis estatísticos | 5,83% | 95,34% |
| $K = 3$ | 49,43% | 56,03% |
| $K = 4$ | 5,19% | 89,16% |
| $K = 5$ | 4,54% | 89,53% |
| $K = 6$ | 4,38% | 89,66% |

Tabela 6 – Valores de *recall* e acurácia para as diferentes simulações realizadas com registros do VI até 7 dias antes de falha (severidade 2).

Para identificação de rótulos para severidade 3 e posteriores foi realizado um estudo variando o número de dias antes de ocorrência para conseguir identificar o comportamento dos agrupamentos à medida que registros mais antigos também são incorporados. A tendência esperada com esta análise é identificar um decréscimo da porcentagem de registros do *clusters* de maiores severidades (1 e 2) e o aumento da proporção de dados para os *clusters* das demais severidades.

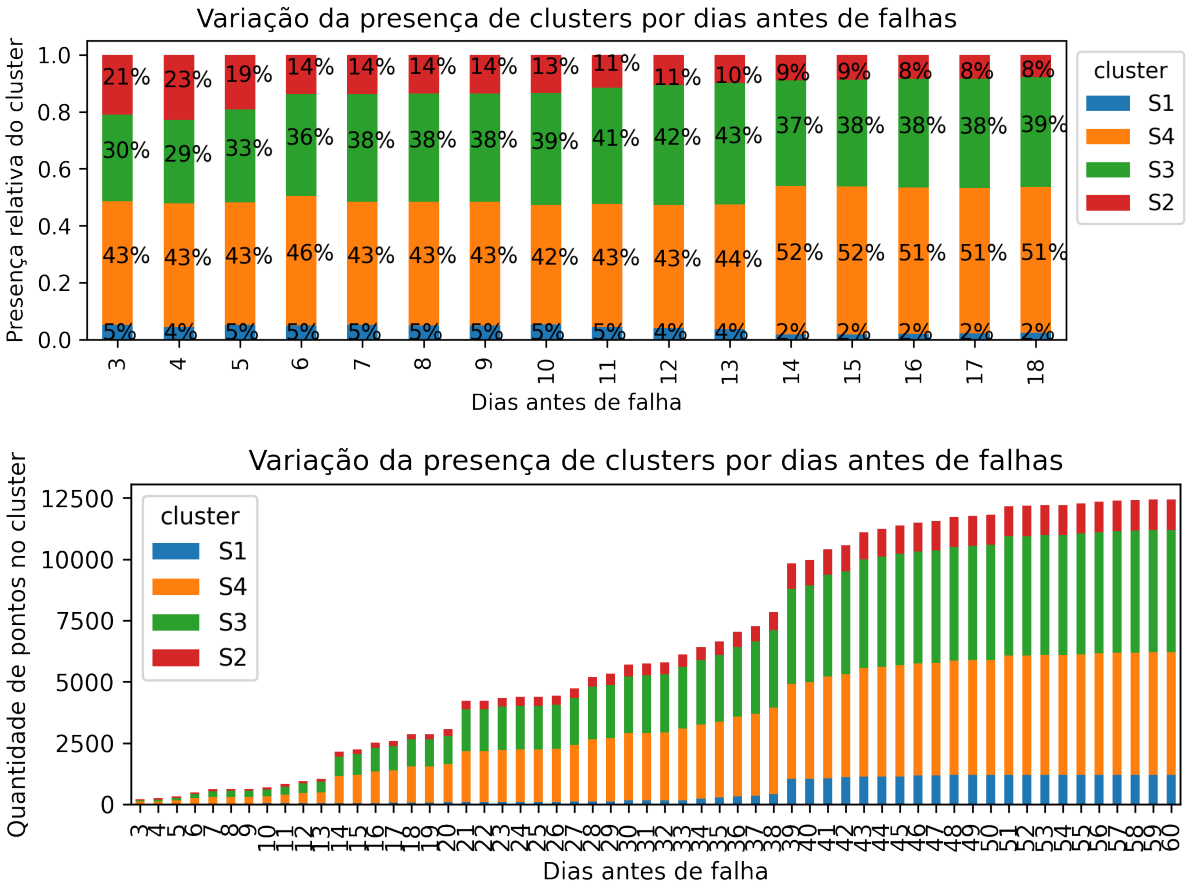


Figura 44 – Variação da porcentagem de *clusters* conforme aumenta-se o intervalo de dias entre coleta de dados do VI e ocorrências: a) em porcentagem e b) em números absolutos.

Conforme ilustrado pela Figura 44, é interessante observar que os *clusters* de maior severidade vão diminuindo à proporção que aparecem, conforme os dias antes da falha vão progredindo, o que é o esperado. Porém, foi observado que esse comportamento eventualmente se reverte quando aumenta-se o intervalo de dias anteriores ao registro de uma ocorrência. Tal fator impede a conclusão definitiva sobre a melhor janela em dias para as severidades 3 e seguintes. Isto acontece devido aos grandes saltos de quantidade de dados entre algumas janelas de tempo, sendo assim, observou-se que as alterações nas proporções dos *clusters* de severidade em função do tempo não apresentaram um padrão claro.

Um estudo foi conduzido na sequência para cada ocorrência identificada no período de dados disponíveis para o VI, mas nenhuma conclusão pôde ser obtida quanto ao padrão do comportamento dos níveis de severidade encontrados à medida que se aumenta a janela de tempo.

### 4.3.2 Classificação

Com base nos resultados de *clustering*, surge o questionamento sobre como um algoritmo de aprendizado supervisionado desempenharia sobre o conjunto de dados, utilizando a estratégia de validação empregada como algoritmo de rotulação dos dados para treinamento.

Um dos primeiros parâmetros do estudo é a definição dos níveis de severidade. Conforme exposto anteriormente, define-se um nível de severidade $s$ todos os registros de um veículo ocorridos numa região em que há falha observada em até $d_s$ dias. Dessa forma, a quantidade de níveis e de dias que definem esses níveis são variados ao longo dos experimentos.

Como algoritmo classificador, vários foram testados nesses experimentos: *random forests*, regressão logística, SVMs, perceptron multicamadas, AdaBoost, árvores de decisão, *naive Bayes* e *gradient tree boosting* (PEDREGOSA et al., 2011). Grande parte desses algoritmos foram amplamente empregados na literatura para tarefas de manutenção preditiva, conforme detalhado no Capítulo 2, e outros por apresentarem características interessantes para os experimentos.

Para a etapa de pré-processamento dos dados, algumas opções foram trabalhadas. Normalização e *scaling* foram técnicas mais comumente empregadas, além de serem requisitos para alguns dos algoritmos empregados. *Scaling* em particular foi analisado utilizando *scalers* para transformar as variáveis em Gaussianas centradas em 0 com variância unitária, valores entre 0 e 1 e máximos em valor unitário. Para redução de dimensionalidade, PCA e LDA foram as duas alternativas exploradas durante os experimentos.

Devido à natureza do *dataset* de ocorrências, existem duas escalas de geolocalização possíveis de serem utilizadas: localização via elemento da via e localização via quilometragem específica. Nem todos os registros de ocorrências possuem ambas as informações; na realidade, apenas 7% dos dados possuem tanto a localização via elemento quanto via quilometragem. Como consequência, para o cruzamento de dados desse *dataset* com os do VI e CC, é necessário definir qual escala pretende-se utilizar. A escala de quilometragem é preferível em relação ao do elemento porque ela é mais próxima do exato local em que foi observado a avaria na via, permitindo uma análise mais precisa, principalmente quando comparado à localização através do elemento, que é por vezes muito maior do que a região efetivamente afetada pela anomalia. Entretanto, por justamente ser uma área menor, existem menos dados dos veículos que se encontram na exata região de uma ocorrência dentro da janela de tempo em análise. Além disso, também é uma informação mais escassa do que a de elemento, portanto também se dispõe de menos dados para cruzar. Tem-se, portanto, um *trade-off* entre precisão e quantidade de dados para treinamento.

Assim, optou-se por trabalhar a questão da escala como mais um parâmetro de

estudo dos experimentos, que pode ser variado em conjunto com os demais. Além das opções de usar exclusivamente uma das escalas, uma terceira alternativa explorada foi um misto dentre elas, onde usou-se o nível de quilometragem como referência para a maior severidade, e o de elementos para as demais.

Outro fator relevante para se destacar ainda sobre geolocalização de falhas é a observação de que o VI, em particular, é um vagão utilizado como parte de uma composição. Isso significa que a influência do defeito geométrico na via pode afetar o VI antes ou depois de sua posição cruzar com o registro da falha. Consequentemente, foi estabelecido outro parâmetro de estudo para a definição de uma margem LRS de tolerância; isto é, um valor de LRS que expande os valores limites do defeito de forma a aumentar a região considerada da ocorrência. Ou seja, se $LRS_{min}$ e $LRS_{max}$ forem os limites LRS mínimos e máximos da localização de uma falha, então uma margem de tolerância $x$ é utilizada de tal forma que $LRS_{min} - x \leq LRS \leq LRS_{max} + x$ define se um registro do VI de posição $LRS$ está dentro da região de falha. Apesar do CC ser um carro que não acompanha uma composição, essa estratégia foi preservada para ele.

Um ponto de atenção que foi observado, principalmente quando se utilizavam maiores períodos em dias para determinar os níveis de severidade (e.g., 200 dias), foi a questão das regiões da via em que não se havia ocorrências. A princípio, assume-se que tais regiões sempre devem ser consideras sem severidade, dado a ausência de defeitos. Porém, essa presunção leva em conta a ausência de dados ao invés de uma confirmação de boas condições. Além disso, o *dataset* de falhas termina em 2020, o que abre a possibilidade de haver ocorrências nessas áreas em 2021, o que acabaria por entrar em alguns níveis de severidade, dependendo dos valores de dias definidos. Portanto, um parâmetro extra adicionado aos ensaios foi a exclusão de dados dos veículos dessas regiões sem nenhum registro de falha no *dataset* de ocorrências.

Um dos principais parâmetros de controle dos experimentos de classificação foram os filtros aplicados ao *dataset* do VI e CC. Devido à natureza temporal dos dados, cada ano específico foi trabalhado individualmente, gerando resultados próprios. Devido à diferença tanto de quantidade de registros, falhas disponíveis no mesmo período e distribuição ao longo dos meses, cada recorte por ano também teve ainda filtros para alguns meses de interesse, bem como em regiões da via específicos. Como consequência, a combinação desses vários parâmetros de filtro junto com as opções de algoritmos classificadores e algoritmos de pré-processamento gerou uma grande quantidade de ensaios realizados.

Devido à esperada diferença de quantidade de dados correspondentes às diferentes severidades, as estratégias de balanceamento de *dataset* para treinamento, *undersampling* e *oversampling*, foram experimentadas. A primeira consiste em descartar amostras das classes mais dominates, de forma a equilibrar a quantidade dentre todos os rótulos. No caso, os dados descartados da base de treinamento foram movidos para a base de testes. Já a

segunda técnica consiste em reutilizar os registros das classes menos populosas, resultando em quantidades novamente próximas entre os rótulos. A diferença entre a classe mais dominante e a menos é determinada através de um parâmetro experimental.

Em relação à divisão do *dataset* dos veículos entre treinamento e testes, um parâmetro de porcentagem para dividir aleatoriamente os registros em cada um dos subconjuntos foi empregado. Também foi implementado a técnica de validação cruzada, onde o conjunto de dados é dividido $N$ vezes entre recortes para treinamento e um para testes, e o modelo é treinado várias vezes (tradicionalmente, $N$ vezes), variando os pedaços utilizados para cada uma dessas etapas.

Após o modelo ser treinado e testado na primeira seleção do *dataset* do VI e CC, um novo filtro independente sobre a totalidade dos dados é executado para o estabelecimento de um novo conjunto para validação. Da mesma forma que os filtros iniciais para o *dataset* de treinamento e testes consistem de vários parâmetros do experimento, esses recortes para validação também são compostos por filtros por ano, mês, dias, LRS e RHs, possibilitando, portanto, amplas escolhas nos eixos temporal e espacial. Na prática, os dois conjuntos de filtros foram empregados de forma a evitar que um mesmo dado estivesse presente nos dois conjuntos de dados.

#### 4.3.2.1 Validação com falhas selecionadas

Para a definição dos filtros para as bases de treinamento e validação, um fato relevante acerca dos *datasets* dos veículos e de falhas é a questão temporal; executando-se os dados do VI de 2022, em que não se tem os registros de ocorrências desse período, ambos os conjuntos terminam ao final de 2020. Consequentemente, o que observa acerca da varição temporal dos dados é que, conforme a Tabela 7 e Figura 45 mostram, a diversidade de proximidade em dias de registros do VI com falhas diminui conforme se aproxima do final do ano. Dessa forma, recortes tradicionais muito empregados em séries temporais não seriam adequados à essa realidade, necessitando, então, recorrer-se a seleções espaciais.

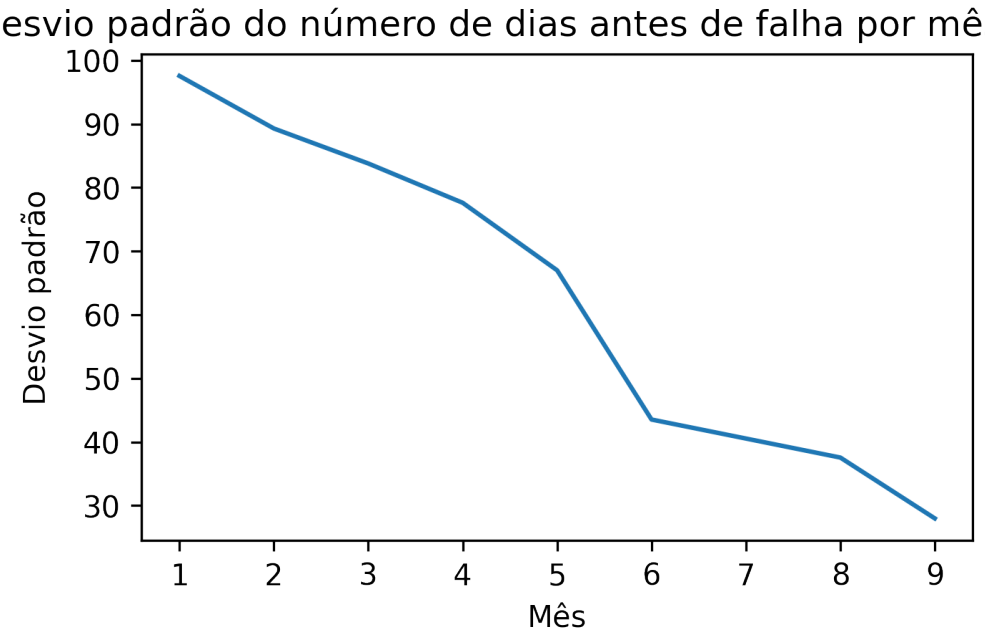


Figura 45 – Desvio padrão dos dias antes de falha para os registros do VI a cada mês. Observa-se claramente que, conforme esperado devido à disponibilidade dos dados, a variância da distância em dias diminui à medida que se caminha em direção ao final do ano de 2020.

| mês | quantidade | média | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 1 | 8195 | 143,762 | 97,341 | 0 | 51 | 117 | 246 | 336 |
| 2 | 17262 | 85,316 | 88,566 | 0 | 29 | 31 | 107 | 324 |
| 3 | 18375 | 70,702 | 82,708 | 0 | 6 | 23 | 162 | 298 |
| 4 | 6246 | 95,072 | 81,750 | 1 | 14 | 66 | 162 | 266 |
| 5 | 3242 | 95,999 | 69,261 | 0 | 24 | 107 | 134 | 214 |
| 6 | 3736 | 96,910 | 50,312 | 0 | 76 | 96 | 136 | 189 |
| 8 | 10765 | 51,338 | 36,005 | 0 | 22 | 40 | 81 | 133 |
| 9 | 2037 | 53,591 | 31,044 | 4 | 31 | 31 | 85 | 117 |

Tabela 7 – Valores estatísticos relevantes sobre a distância em dias de registros do VI a falhas da EFVM.

Conforme discutido na conceituação do problema, é desejável que os níveis de severidade encontrados representem bem a deterioração do estado da via conforme os dias passam, aumentando a severidade à medida que se aproxima de um registro de falha. Portanto, procuraram-se na base de dados algumas ocorrências que não apenas tenham dados de veículos associados a elas na janela de tempo proposta, como também tenham registros do VI ou CC em múltiplos dias diversos, permitindo o acompanhamento da variação do nível de severidade predominante ao longo do tempo. É importante ressaltar

que, para essa validação, todos os dados dos veículos na região da ocorrência em questão foram excluídos da base de treinamento, evitando um resultado erroneamente positivo.

#### 4.3.2.2 Clustering para aprimoramento de rotulação

Dado que a geolocalização de uma leitura dos veículos com uma falha depende consideravelmente da precisão do registro feito da ocorrência, bem como da escala escolhida e margem de tolerância adotada, é esperado que o resultado final desse processo tenha alguma imprecisão. Pensando em melhorar esse resultado, *clustering* foi teorizado como uma potencial solução para distinguir dessa seleção dados afetados pela falha e dados não afetados.

Inspecionando-se o resultado do algoritmo *K-means* (outros algoritmos, como *linkage* simples, *linkage* completo e espectral também foram experimentados) para $K = 2$, comprovou-se que de fato não há grupos distintos relacionados à proximidade do tempo com uma falha. O que se observa são distinções dentro de cada visita do VI, muitas vezes tomando forma de uma seção contida na seleção original, tal como ilustra a Figura 46.

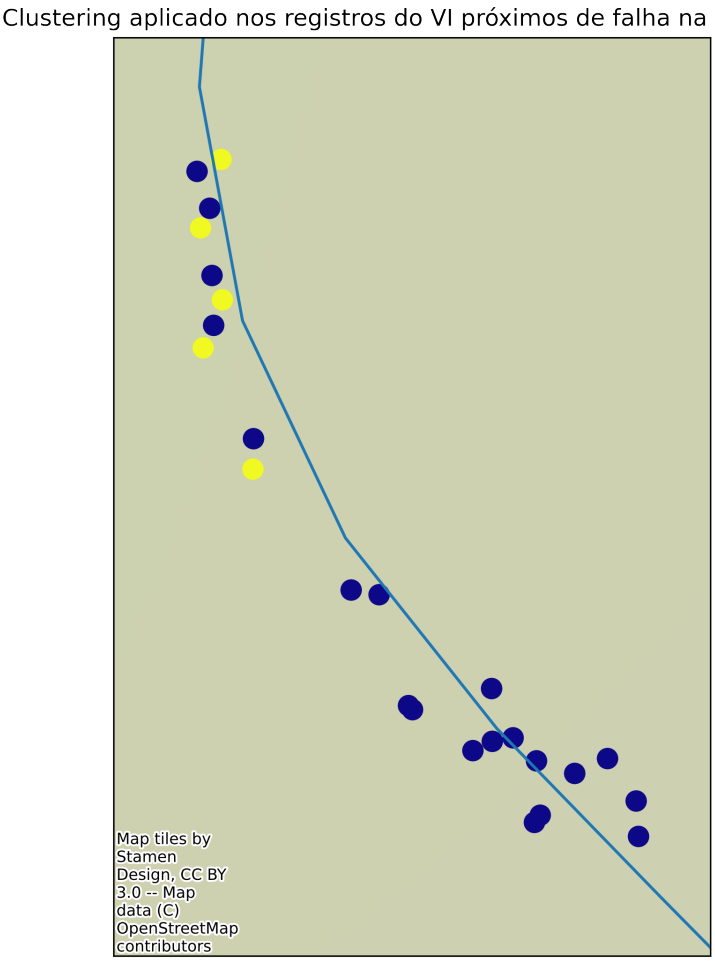


Figura 46 – Dados do VI pré-rotulados com a maior severidade agrupados após execução de algoritmo de *clustering* para uma falha na região EH 82/83, onde cada cor representa um *cluster* diferente. Nota-se que há um agrupamento regional em um subsetor da seleção, o que corrobora a hipótese dessa técnica separar dados afetados pela falha e dados não afetados.

#### 4.3.2.3 Resultados do Vagão Instrumentado

De forma geral, os resultados obtidos pelas diversas execuções dos classificadores apresentaram boas métricas de acurácia e precisão para o *dataset* de testes, mas desempenho consideravelmente inferior para a seleção de validação, especialmente quando suficientemente distinta ao *dataset* de treinamento. Um dos exemplos desses resultados está detalhado a seguir, onde se definiu um único nível de severidade com a janela de 90 dias (sempre existe a classe padrão "sem severidade"complementando os estudos, então trata-se de um classificador binário), utilizando os dados do VI de Março a Maio de 2020, com o classificador de *random forests* utilizando normalização e 5 dimensões de PCA. Os resultados do treinamento estão detalhados na Tabela 8. Para a validação, utilizou-se os dados de Janeiro e Fevereiro de 2020, e os resultados constam na Tabela 9. Uma falha selecionada para inspecionar o modelo em ação está ilustrada na Figura 47.

| | precisão | *recall* | *f1-score* | quantidade |
|---|---|---|---|---|
| Severidade máxima | 81,70% | 53,28% | 64,50% | 595 |
| Sem severidade | 85,85% | 95,96% | 90,63% | 1758 |
| acurácia | | | 85,17% | 2353 |
| média | 83,78% | 74,62% | 77,56% | 2353 |
| média ponderada | 84,80% | 85,17% | 84,02% | 2353 |

Tabela 8 – Métricas de precisão e acurácia do modelo aplicado no conjunto de dados de teste.

| | precisão | *recall* | *f1-score* | quantidade |
|---|---|---|---|---|
| Severidade máxima | 19,28% | 13,78% | 16,07% | 1901 |
| Sem severidade | 78,59% | 84,58% | 81,48% | 7114 |
| acurácia | | | 69,65% | 9015 |
| média | 48,94% | 49,18% | 48,77% | 9015 |
| média ponderada | 66,08% | 69,65% | 67,68% | 9015 |

Tabela 9 – Métricas de precisão e acurácia do modelo aplicado no conjunto de dados de validação.

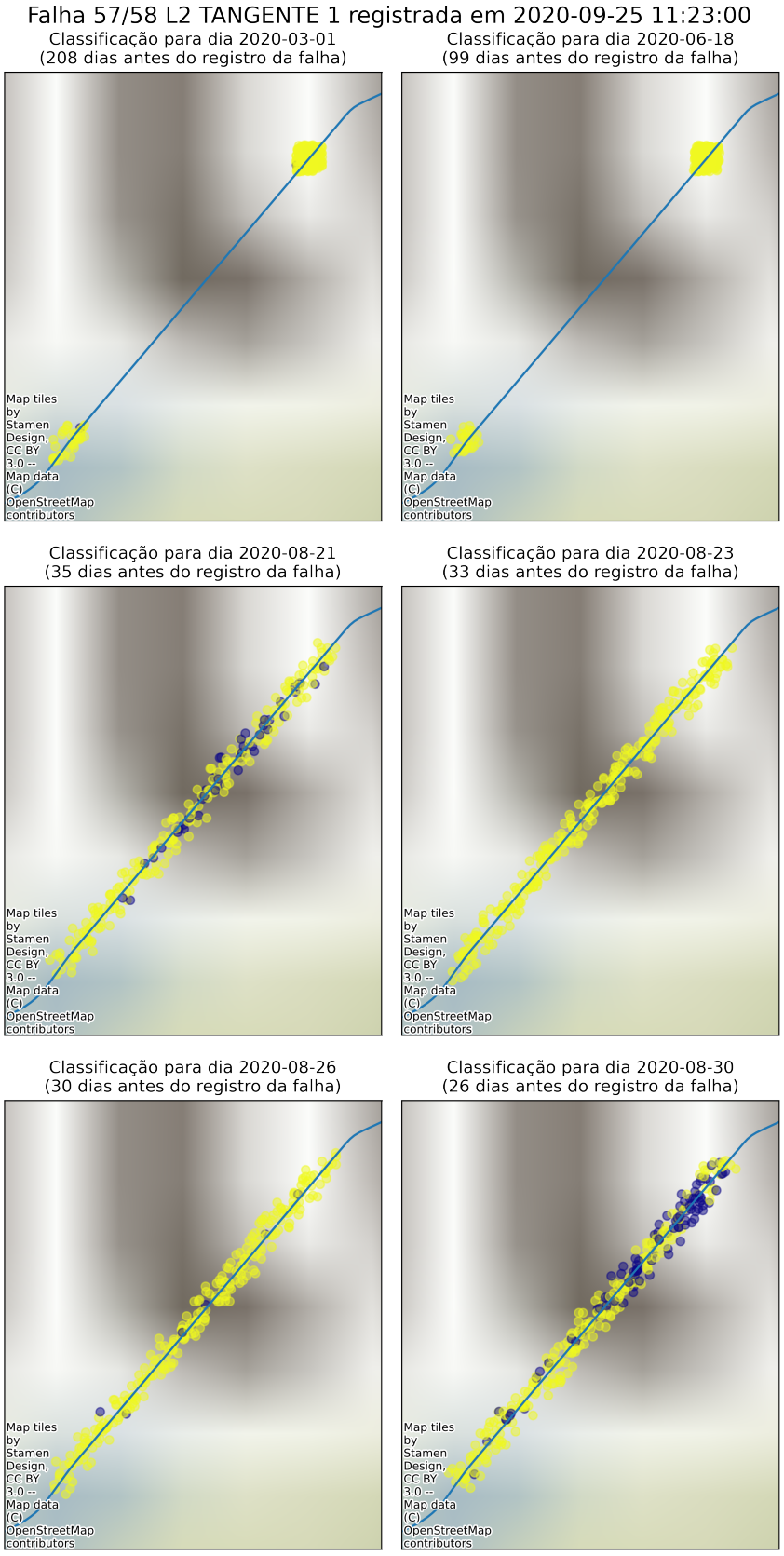


Figura 47 – Dados do VI numa ocorrência, classificados segundo o nível de severidade determinado pelo modelo treinado. Quanto mais escura a cor de um ponto, maior a sua severidade determinada pelo modelo.

#### 4.3.2.4 Resultados do Carro Controle

Dados os resultados do VI aquém do esperado, particularmente os de validação, convém avaliar o desempenho que a mesma estratégia apresenta com os dados do CC. A diversidade de parâmetros do experimento é a mesma do VI, mudando apenas o *dataset* ao qual eles são aplicados. Um dos exemplos dos resultados para o CC está detalhado a seguir, onde se definiu um único nível de severidade com a janela de 60 dias (novamente tratando-se de classificador binário), utilizando os dados do CC do segundo semestre 2020, com o classificador de *random forests* utilizando normalização sem redução de dimensionalidade. Também foi empregado *K-means* acompanhando de PCA de 8 dimensões para o aprimoramento da rotulação. Os resultados do treinamento estão detalhados na Tabela 10. Para a validação, utilizou-se os dados do primeiro semestre de 2020, e os resultados constam na Tabela 11. Uma falha selecionada para inspecionar o modelo em ação está ilustrada na Figura 48.

| | precisão | *recall* | *f1-score* | quantidade |
|---|---|---|---|---|
| Severidade máxima | 83,69% | 66,58% | 74,17% | 1218 |
| Sem severidade | 89,98% | 95,86% | 92,82% | 3812 |
| acurácia | | | 88,77% | 5030 |
| média | 86,84% | 81,22% | 83,49% | 5030 |
| média ponderada | 88,46% | 88,77% | 88,31% | 5030 |

Tabela 10 – Métricas de precisão e acurácia do modelo aplicado no conjunto de dados do CC de teste.

| | precisão | *recall* | *f1-score* | quantidade |
|---|---|---|---|---|
| Severidade máxima | 20,12% | 11,87% | 14,93% | 7953 |
| Sem severidade | 77,66% | 86,66% | 81,91% | 28109 |
| acurácia | | | 70,17% | 36062 |
| média | 48,89% | 49,27% | 48,42% | 36062 |
| média ponderada | 64,97% | 70,17% | 67,14% | 36062 |

Tabela 11 – Métricas de precisão e acurácia do modelo aplicado no conjunto de dados do CC de validação.

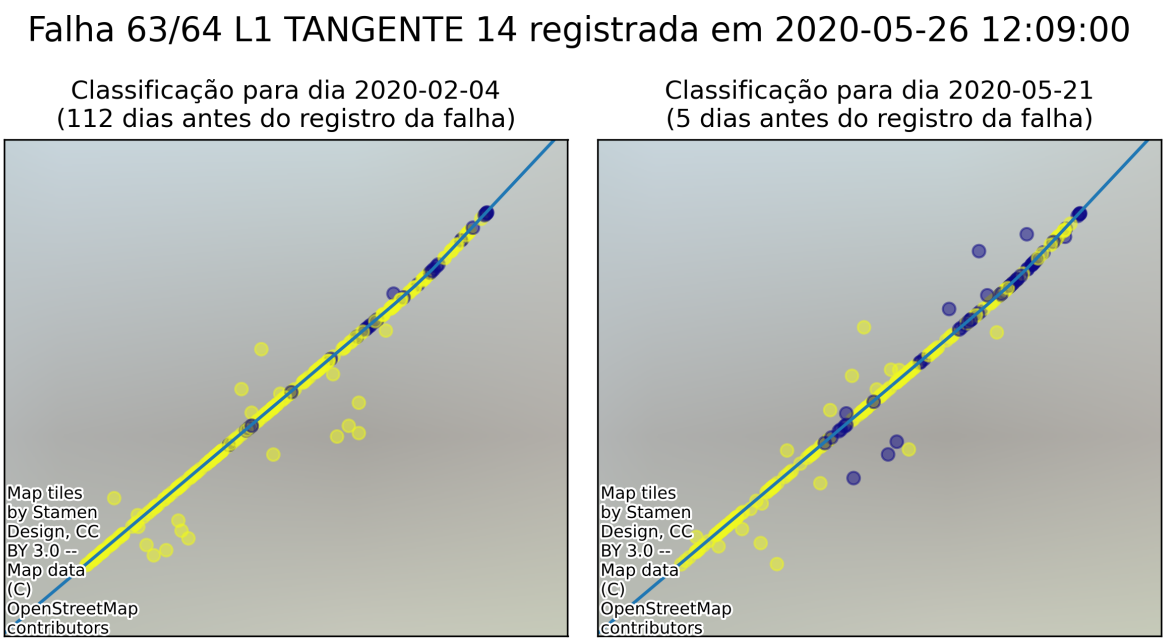


Figura 48 – Dados do CC numa ocorrência, classificados segundo o nível de severidade determinado pelo modelo treinado. Quanto mais escura a cor de um ponto, maior a sua severidade determinada pelo modelo.

### 4.3.3 Regressão

De posse dos resultados de aprendizado de máquina supervisionado, e a consequente aparente falta de generalização dos modelos, convém avaliar diretamente o relacionamento entre distância em dias de falhas aos registros de leitura dos veículos. Dado que o CC é o que faz a medida direta da via, ele é o mais adequado a mostrar alguma correlação.

#### 4.3.3.1 Dataset completo

Para início dos experimentos, foi utilizada uma amostragem aleatória da totalidade do *dataset*, filtrando apenas por seções da via com registros de falha, uma vez que a ausência de defeitos impossibilita a quantificação da distância no tempo para um evento. Nesse primeiro momento, foram utilizadas *random forests* como algoritmo de regressão.

Algumas métricas relevantes para esse tipo de aprendizado de máquina dessa primeira iteração são apresentadas na Tabela 12, com o peso de cada variável para essa regressão descritas na Tabela 13.

| $R^2$ *score* | Erro médio absoluto | Variância explicada *score* |
|---|---|---|
| 0,851 | 18,216 | 0,851 |

Tabela 12 – Métricas relevantes para primeira iteração do modelo de regressão, utilizando amostragem do conjunto de dados completo do CC.

| Variável | Importância relativa |
|---|---|
| Bitola frontal | 12,583% |
| Bitola traseira | 10,089% |
| Suspensão | 09,367% |
| Variação da suspensão | 15,595% |
| Empeno (corda 5,5m) | 04,464% |
| Empeno (corda 1,7m) | 02,001% |
| Raio inverso | 17,385% |
| Nivelamento esquerdo | 01,946% |
| Nivelamento direito | 01,849% |
| Inclinação trilho esquerdo | 10,533% |
| Inclinação trilho direito | 09,966% |
| Acelerômetro X | 01,357% |
| Acelerômetro Y | 01,521% |
| Acelerômetro Z | 01,336% |

Tabela 13 – Importância relativa de cada variável do CC para a construção do modelo de regressão.

Para se concretizar o que esses números representam num experimento prático, foi selecionada uma falha aleatória para inspeção com o modelo. Todos os dados do CC ocorridos na região da falha em viagens anteriores ao evento foram obtidos, e os pontos presentes no *dataset* utilizado para treinamento do modelo foram removidos. Os resultados para os registros restantes é apresentado nas Figuras 49 e 50.

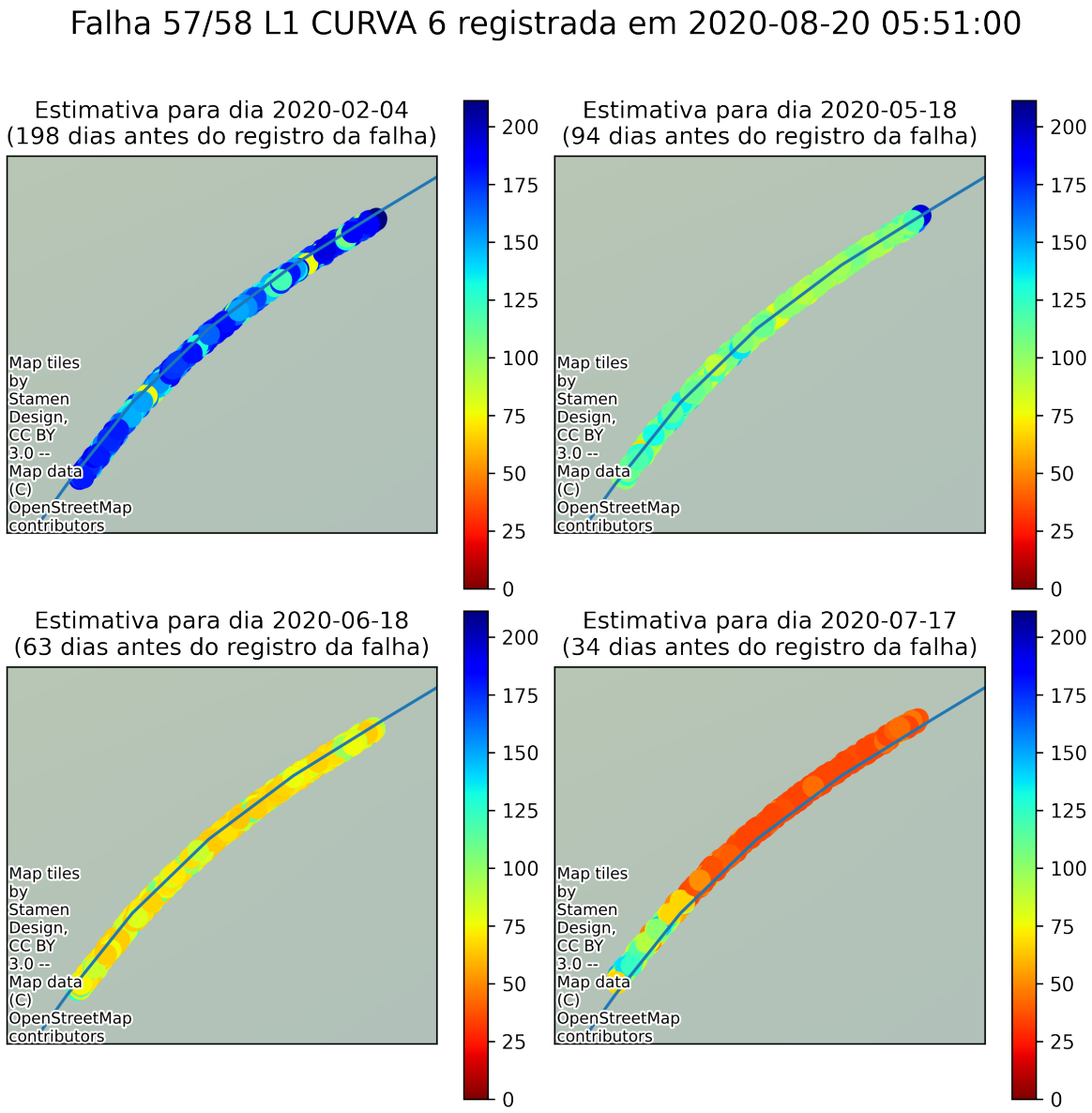


Figura 49 – Dados do CC numa ocorrência, com a distância em dias para a falha estimada pelo modelo de regressão representada em cores.

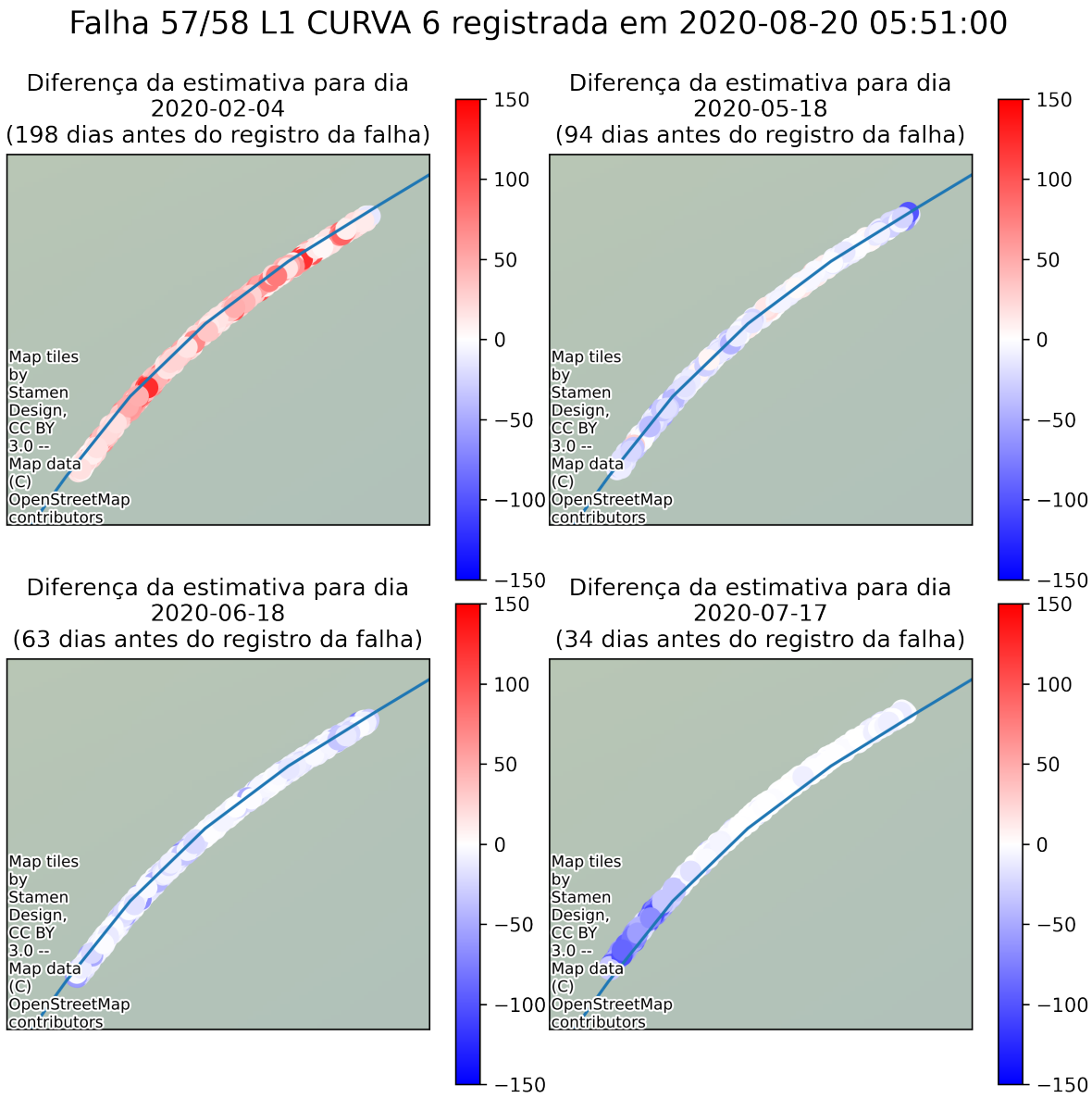


Figura 50 – Dados do CC numa ocorrência, com a diferença entre a distância em dias para a falha estimada pelo modelo de regressão e o valor esperado representada em cores.

#### 4.3.3.2 Eliminando dados de uma falha

Agora, para avaliar a capacidade do modelo de generalizar o que foi aprendido, um novo experimento foi realizado, desta vez selecionando-se de antemão um defeito aleatório, e selecionando uma amostragem aleatória dos dados CC para as demais regiões da via, excetuando-se a localização da falha selecionada. Para esse experimento, alguns algoritmos de regressão foram testados, e os resultados apresentados a seguir foram obtidos utilizando-se redes neurais *feed-forward* com duas camadas de neurônios. Vale destacar que os hiperparâmetros foram calibrados utilizando *tuner* automáticos, com o objetivo de

encontrar aqueles que minimizem os erros.

As mesmas métricas de desempenho do modelo foram coletadas, cujo resultado é apresentado na Tabela 14, bem como o resultado da execução do algoritmo na falha de validação pré-selecionada, ilustradas pelas Figuras 51 e 52.

| $R^2$ *score* | Erro médio absoluto | Variância explicada *score* |
|---|---|---|
| 0,798 | 22,042 | 0,798 |

Tabela 14 – Métricas relevantes para a segunda iteração do modelo de regressão, utilizando amostragem do conjunto de dados do CC excluindo completamente a região da falha de validação.

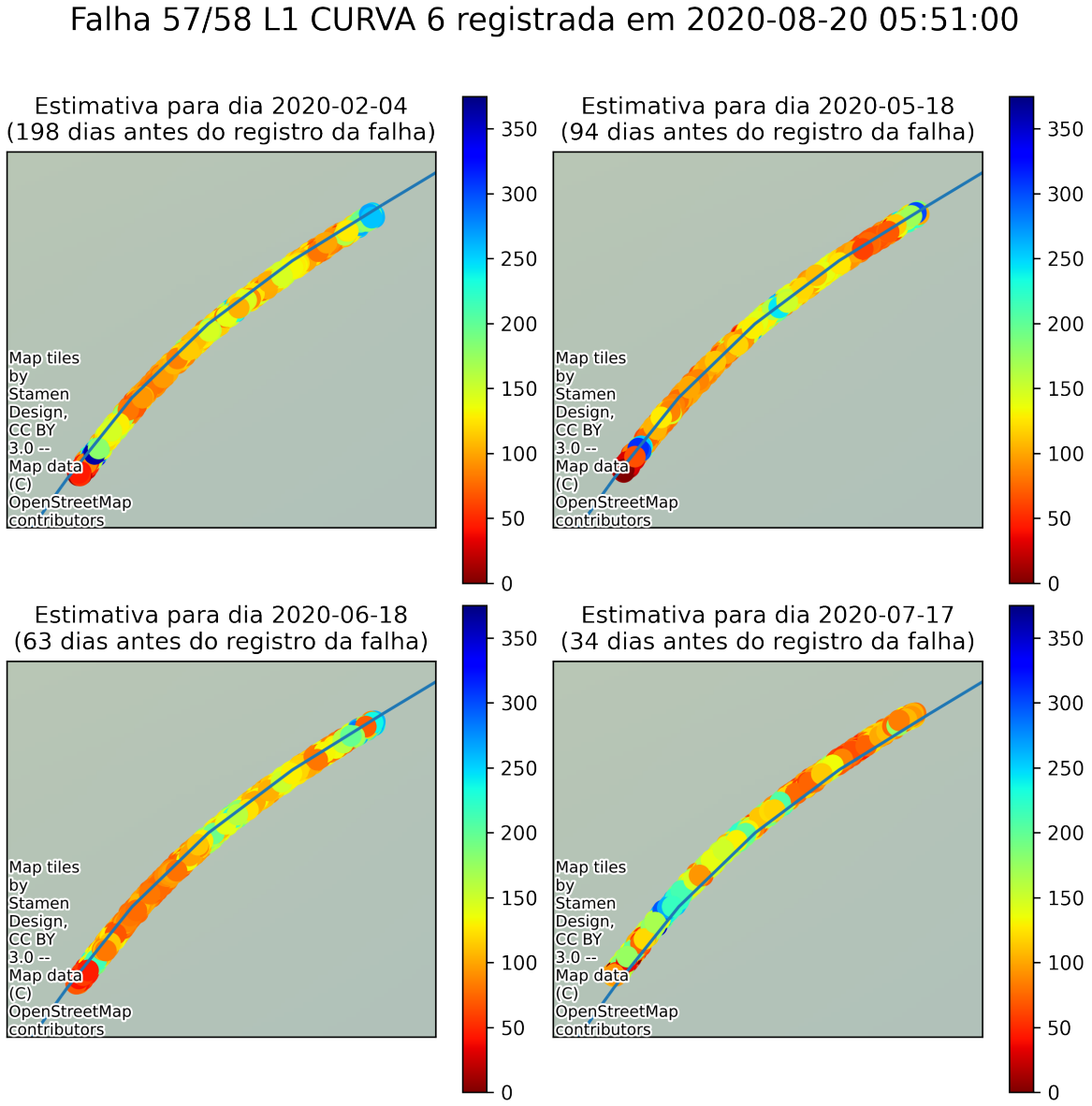


Figura 51 – Dados do CC numa ocorrência, com a distância em dias para a falha estimada pelo modelo de regressão representada em cores.

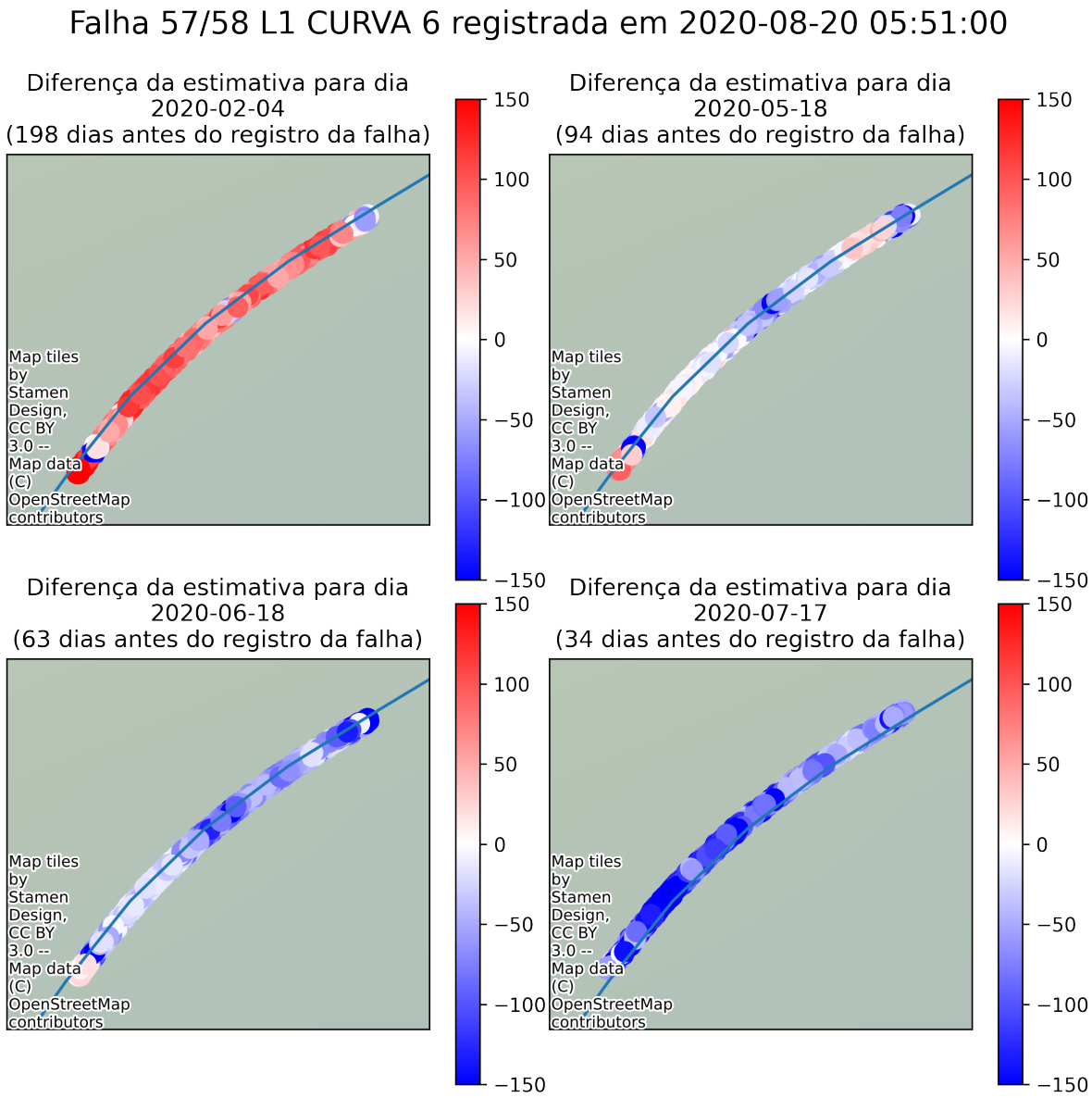


Figura 52 – Dados do CC numa ocorrência, com a diferença entre a distância em dias para a falha estimada pelo modelo de regressão e o valor esperado representada em cores.

#### 4.3.3.3 Cruzamento de todas as falhas

Por fim, a fim de entender se há alguma falha que é comportamentalmente similar a alguma outra, um experimento de validação cruzada foi realizado. Desta vez, para cada falha do *dataset*, uma amostragem dos dados CC que cobrem sua região foi utilizada para treinamento do modelo de regressão, e o desempenho contra amostragens nas regiões de todas os demais defeitos foi avaliado. Aqui, o mesmo modelo de rede neural, com os mesmos hiperparâmetros definidos pelo *tuner* foi empregado para treinamento.

Os cinco melhores resultados são detalhados na Tabela 15 a seguir.

| Falha de treinamento | Falha de validação | $R^2$ *score* |
|---|---|---|
| Desnivelamento transversal EH 53/54 L1 tangente 5 | Desnivelamento longitudinal EH 58/59 L1 curva 3 | 0,61219 |
| Desnivelamento transversal EH 53/54 L1 tangente 5 | Quebra de dormente de aço EH 40/41 L2 tangente 9 | 0,55526 |
| Desnivelamento longitudinal EH 40/41 L2 tangente 9 | Quebra de dormente de aço EH 40/41 L2 tangente 9 | 0,49471 |
| Recalque de Plataforma | Desalinhamento EH 70/71 L2 curva 10 | 0,48839 |
| Desnivelamento longitudinal EH 40/41 L2 tangente 9 | Desnivelamento transversal EH 53/54 L1 tangente 5 | 0,44410 |

Tabela 15 – As falhas cujos modelos de regressão treinados com dados do CC registrados em suas regiões obtiveram as melhores métricas em registros do CC em outros elementos de defeitos.

Para fins de comparação, o par de falhas com melhor $R^2$ *score* foi avaliado individualmente, com o modelo retreinado e executado nos pontos do CC em questão. Os resultados são apresentados pelas Figuras 53 e 54.

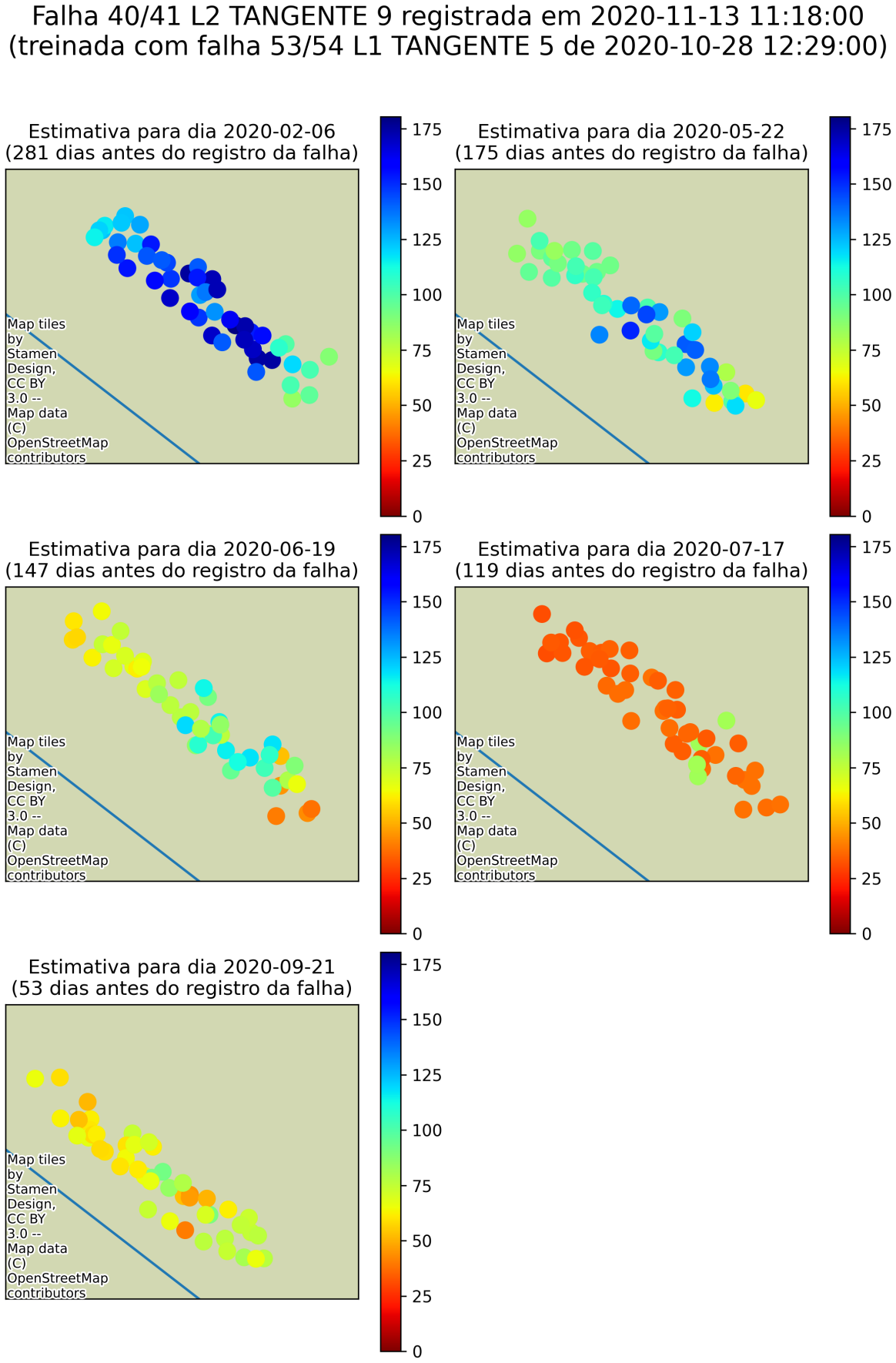


Figura 53 – Dados do CC numa ocorrência, com a distância em dias para a falha estimada pelo modelo de regressão representada em cores.

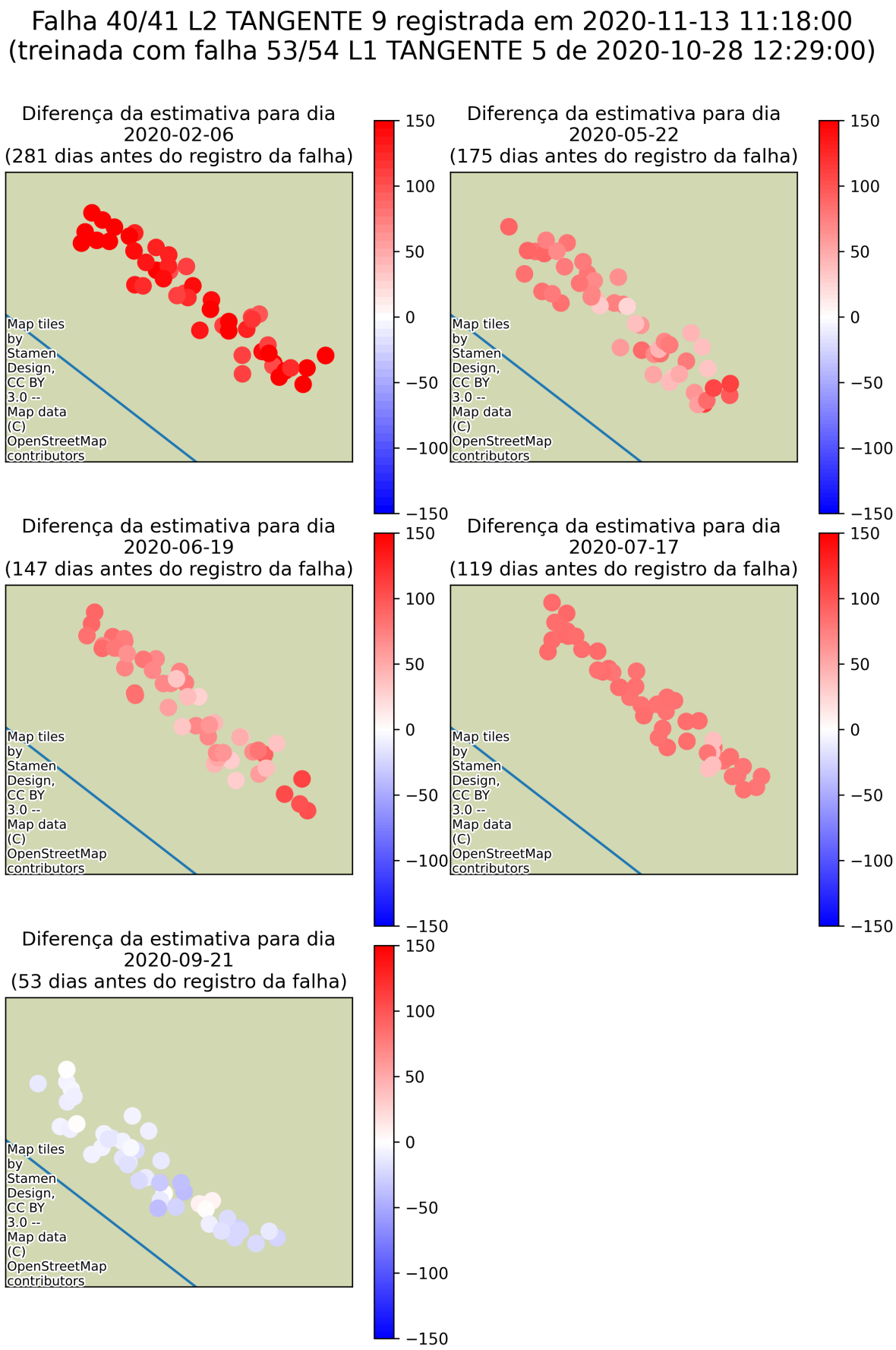


Figura 54 – Dados do CC numa ocorrência, com a diferença entre a distância em dias para a falha estimada pelo modelo de regressão e o valor esperado representada em cores.

## 4.4 Integração com ferramenta Datamap

Conforme discutido na Seção 3.3, a base da plataforma de visualização do Datamap já havia sido construída anteriormente, trazendo uma base confortável para a expansão de funcionalidades adequadas à visualização dos resultados obtidos. Entretanto, o que se tinha até o momento era apenas o esqueleto de uma aplicação que tem o potencial de se tornar a ferramenta de visualização do projeto, mas ainda não era presente qualquer funcionalidade com dados reais.

Como parte da contínua construção dessa plataforma, foi implementado um novo botão à interface, capaz de subir arquivos GeoJSON do computador do usuário. Esses arquivos são gerados pelo experimentos de aprendizado de máquina, e eles trazem os dados de falhas e as severidades encontradas. O ferramental construído por trás como parte da gestão de dados (detalhados na Seção 4.1) também é capaz de facilmente exportar dados do VI e CC de acordo com uma gama de filtros e seleções, permitindo de forma prática a geração de GeoJSONs através de *Jupyter notebooks*, por exemplo. Os resultados dessa implementação são ilustrados pela Figuras 55 e 56.

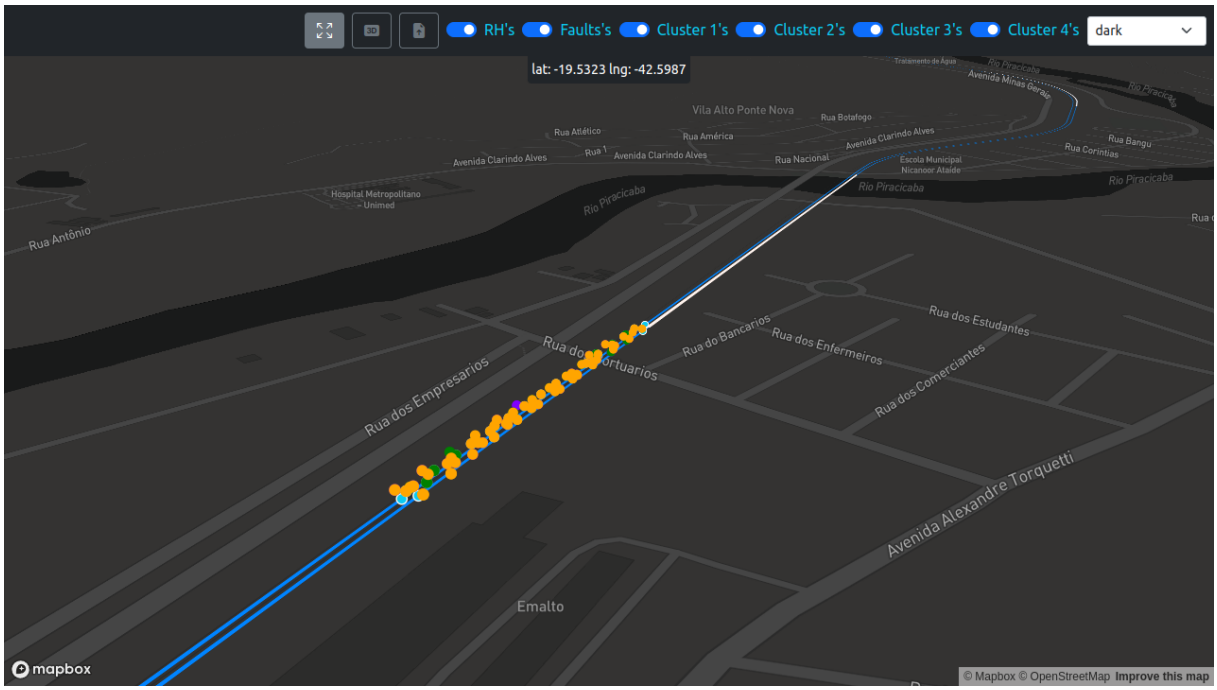


Figura 55 – Interface do Datamap após atualização com botão de *upload* dos resultados do *clustering*. No topo, as opções de falhas e *clusters* encontrados são exibidas após a importação dos resultados. No centro, os pontos coloridos ilustram leituras do VI, em que cada cor representa um *cluster* distinto. Há também um segmento de reta em rosa, ilustrando uma falha ocorrida naquele trecho.

Além disso, alguns detalhes também foram acertados, como a construção de gráficos com as variáveis relevantes a partir dos pontos do VI e CC selecionados (anteriormente, os

gráficos eram apenas *mock-ups*, utilizando números aleatórios), e eventuais correções de responsividade.

Foi também implementado no *backend* da aplicação o controle de algumas das funcionalidades do Datamap. Assim, para fins de demonstração, alguns arquivos GeoJSON de dados de falhas, VI e CC foram construídos e pré-carregados no servidor, e o *backend* faz o controle entre carregar na visualização esses arquivos ou pedir ao usuário o upload. O *backend* também passa a controlar a dimensão em função da qual os dados são agrupados na tela (data, *cluster*, viagem, linha, etc). Com isso, tem-se o controle dinâmico do comportamento da plataforma, sem a necessidade de mudanças no código-fonte ou de implantações de novas versões da aplicação nos ambientes produtivos.

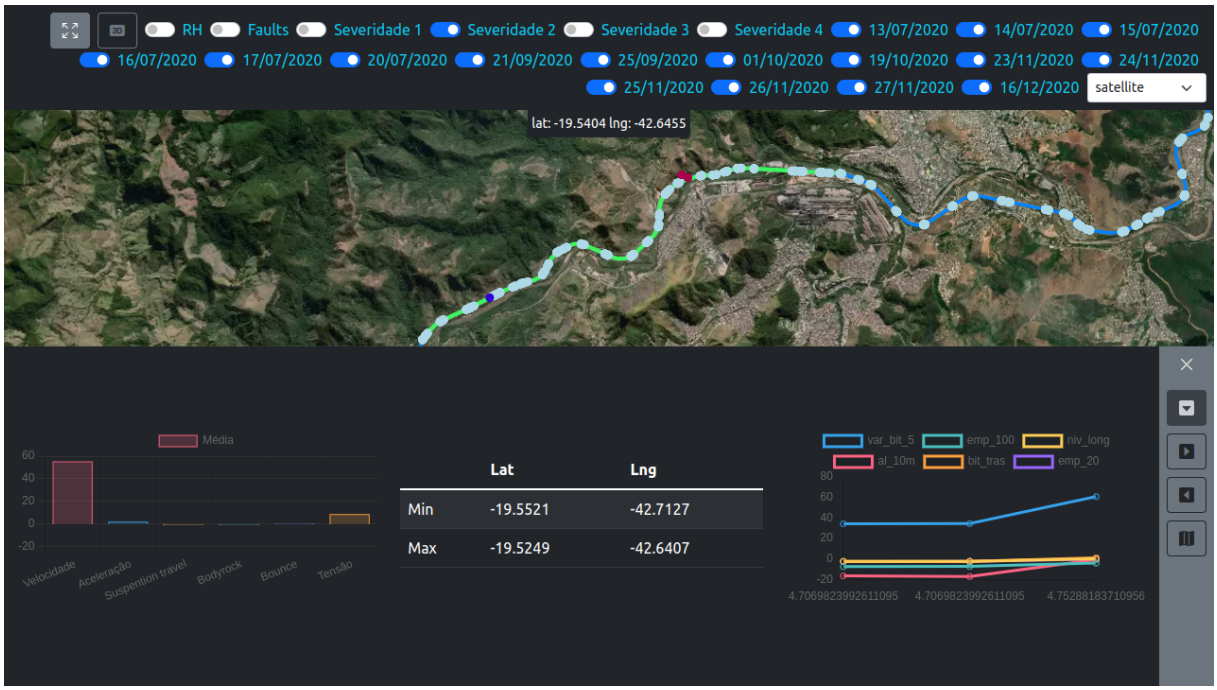


Figura 56 – Interface do Datamap com dados do Vagão Instrumentado e Carro Controle. Os dados do VI foram agrupados via *clustering*, e cada cluster atribuído a uma severidade, que pode ser filtrada na tela. Os dados do Carro Controle estão divididos por data. Na parte inferior da tela, há gráficos de variáveis importantes do VI e CC no trecho selecionado (em verde).

Dessa forma, agora está disponibilizada uma ferramenta capaz de ilustrar as mais diversas características dos dados do projeto. Também fica possível o entendimento do comportamento dos *clusters* encontrados e sua variação no espaço e no tempo, através da visualização dos pontos na ferrovia e filtros por data.

# 5 Conclusão

As técnicas de *Big Data* e Ciência dos Dados utilizadas nesse projeto contribuíram positivamente para a ingestão, processamento e análise dos dados, além da construção de um modelo de níveis de severidade melhorado sobre o sistema atual.

Os modelos unificados dos dados possibilitaram a flexível e precisa busca seletiva por dados, permitindo que estudos fossem sempre conduzidos em cima de uma base centralizada, além de ser crucial para a agregação de novos registros. Os experimentos de aprendizado de máquina e a aplicação de visualização dos dados se beneficiam de um contrato bem-especificado para os dados, consequência de uma separação de responsabilidades do sistema trazida pela implementação de um processo de ETL para lidar com a incerteza e variedade de formatos de dados.

As abstrações criadas através da construção de um novo e tratado sistema linear de referência e o cálculo das variáveis importantes da área também foram pontos chave para o cumprimento dos objetivos. O LRS permitiu filtros precisos e análise espacial cruzada dos registros, condição necessária para a estratégia de validação dos experimentos de aprendizado de máquina. Da mesma forma, o modelo multidimensional foi central para a criação do contrato dos dados e abriu caminho para as contribuições já mencionadas na unificação do modelo de dados.

Os estudos de aprendizado de máquina não-supervisionados revelaram que existe uma relação mais relevante das leituras do VI e as falhas da ferrovia. Ao agrupar dados similares sem qualquer rótulo, fica nítido que essas relações multidimensionais são essenciais para melhorar a predição de manutenção. Em particular, os experimentos com 4 e 5 grupos apresentaram significativa aumento de precisão quando comparados aos existentes métodos estatísticos baseados na estratégia Monash.

Já os experimentos de aprendizado supervisionado demonstram a necessidade de uma maior variedade dos registros para o desenvolvimento de modelos mais direcionados. Devido às limitações dos conjuntos de dados disponíveis, resultando em um *dataset* disponível para treinamento constituído de basicamente alguns meses de dados, os modelos desenvolvidos, tanto de classificação quanto de regressão, não foram capazes de generalizar devidamente a relação de distância de falhas no tempo às leituras dos sensores, apesar de precisarem de apenas alguns poucos registros para absorverem uma ocorrência.

Finalmente, a aplicação de visualização de dados melhorada trouxe como contribuição uma plataforma capaz de ser utilizada tanto para estudos dos dados como apresentação dos resultados dos modelos preditores. Devido à flexibilidade de camadas parametrizáveis, a variedade de características dos dados que podem agora ser analisados através de mapas

e gráficos interativos é considerável. Também é importante ressaltar a importância de se permitir a prática visualização dos resultados dos modelos de aprendizado de máquina, requisito essencial para o emprego dos modelos no campo.

## 5.1 Perspectivas de continuidade

Como perspectivas futuras de continuidade, os processos de ingestão de dados encontram-se maduros para serem encarregados de processar novas levas de dados. Com mais dados tanto de registros do VI e CC quanto de ocorrências na via, os modelos de aprendizado de máquina supervisionados devem ganhar a capacidade de generalização, o que levaria a uma melhor acurácia com dados inéditos.

Outra possibilidade de continuação é a condução de inspeções *in loco* dos resultados de iminência de falha apontados pelos modelos preditores. Devido ao contexto global em que esse trabalho foi realizado, tais visitas ao local não foram possíveis de serem conduzidas nos momentos mais oportunos, privando o projeto de uma possível valiosa alternativa de validação dos níveis de severidade mais elevados obtidos. Com uma eventual disponibilidade de dados mais recentes de leituras dos veículos, essa oportunidade torna-se viável.

Devido a aspectos operacionais, os dados de ocorrências disponibilizados contém apenas intervenções corretivas, isto é, após a observação de um defeito ou outra avaria à via permanente. Os experimentos de aprendizado de máquina demonstram, pela questão da eliminação dos trechos em que não se havia registros de falha alguma durante todo o período do *dataset*, que um conjunto de dados contendo manutenções preventivas e visitas de inspeção com confirmação de operacionalidade normal da via seria de grande valia aos esforços de generalização dos modelos. Tem-se, então, outro fator de possível desenvolvimento futuro do projeto.

# Referências

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 16387*: Via férrea - classificação de vias. Rio de Janeiro, 2016. 25 p. Citado 2 vezes nas páginas 11 e 62.

BREIMAN, L. Random forests. *Machine Learning*, n. 45, p. 5–32, 2001. Citado na página 18.

CARVALHO, T. P. et al. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, v. 137, 11 2019. ISSN 03608352. Citado 7 vezes nas páginas 10, 17, 18, 19, 20, 21 e 24.

CASSARO, L.; SILVA, C. Estudo para parametrização dos registros do Vagão Instrumentado EFVM - relatório 10. Engenharia Ferroviária Via Permanente da VALE. 2021. Citado na página 12.

COMITÉ EUROPÉEN DE NORMALISATION. *Railway applications - Track - Track geometry quality - Part 5: Geometric quality levels - Plain line, switches and crossings*. [S.l.], 2017. Citado na página 24.

COSTA, J. et al. Metodologia de priorização da manutenção de via usando dados do carro controle para aumento da segurança na EFC. *Semanda de Tecnologia Metroferroviária, 22ª*, p. 22, 2016. Citado na página 10.

CURTIN, K.; TURNER, D. The geographic information science & technology body of knowledge. In: UNIVERSITY CONSORTIUM FOR GEOGRAPHIC INFORMATION SCIENCE. 4th quarter 2019 edition. ed. [S.l.]: John P. Wilson (ed.), 2019. cap. Linear Referencing. DOI: 10.22224/gistbok/2019.4. Citado na página 31.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: SPRINGER. *Ensemble Machine Learning: Methods and Applications*. New York, NY: Springer New York, 2012. p. 157–175. ISBN 978-1-4419-9326-7. Disponível em: <https://doi.org/10.1007/978-1-4419-9326-7_5>. Citado na página 18.

CÁTEDRA UNDERRAIL. *Big Data Analytics aplicado aos parâmetros de vagão instrumentado e carro controle no apoio da gestão de manutenção de via permanente - Relatório 1*. São Paulo, 2021. Citado na página 41.

CÁTEDRA UNDERRAIL. *Big Data Analytics aplicado aos parâmetros de vagão instrumentado e carro controle no apoio da gestão de manutenção de via permanente - Relatório 2*. São Paulo, 2021. Citado na página 43.

DAHLQVIST, F. et al. Growing opportunities in the internet of things. *McKinsey & Company*, p. 1–6, 2019. Citado na página 8.

DAMA INTERNATIONAL. *Data Management Body of Knowledge*. [S.l.]: Technics Publications, 2017. Citado na página 16.

DARBY, M. et al. The development of an instrumented wagon for continuously monitoring track condition. In: MONASH UNIVERSITY. *AusRAIL PLUS 2003, Investing in*

*Australian Rail - Strategies and Solutions*. Sydney, NSW, Australia, 2003. p. 7. Citado 2 vezes nas páginas 24 e 25.

DATAONE. *DataONE Education Module: Data Management*. 2016. <https://github.com/DataONEorg/Education/blob/1f8ab80060ef24c583d84219eb067a03d54070d7/_lessons/lessons/>. Citado 2 vezes nas páginas 16 e 17.

ERTEL, W. *Introduction to Artificial Intelligence*. 2. ed. [S.l.]: Springer Cham, 2018. (Undergraduate Topics in Computer Science). Citado na página 21.

FALCÃO, V. A. A importância do transporte ferroviário de carga para a economia brasileira e suas reais perspectivas de crescimento. *Revista de Engenharia Civil. Ed*, v. 45, p. 51–63, 2013. Citado na página 9.

FUMEO, E.; ONETO, L.; ANGUITA, D. Condition based maintenance in railway transportation systems based on big data streaming analysis. *Procedia Computer Science*, v. 53, 2015. ISSN 18770509. Citado na página 9.

IGUAL, L.; SEGUÍ, S. *Introduction to Data Science*. 1. ed. [S.l.]: Spring Cham, 2017. (Undergraduate Topics in Computer Science). Citado 2 vezes nas páginas 20 e 23.

JAMSHIDI, A. et al. A decision support approach for condition-based maintenance of rails based on big data analysis. *Transportation Research Part C: Emerging Technologies*, v. 95, 10 2018. ISSN 0968090X. Citado 2 vezes nas páginas 10 e 27.

JU-LONG, D. Control problems of grey systems. *Systems & Control Letters*, v. 1, n. 5, p. 288–294, 1982. ISSN 0167-6911. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016769118280025X>. Citado na página 20.

JÚNIOR, A. A. dos S. et al. Aprofundamento na avaliação dos parâmetros medidos com o vagão instrumentado. 2020. Citado 4 vezes nas páginas 24, 25, 26 e 36.

KAYACAN, E.; ULUTAS, B.; KAYNAK, O. Grey system theory-based models in time series prediction. *Expert Systems with Applications*, v. 37, n. 2, p. 1784–1789, 2010. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417409007258>. Citado na página 20.

KONDAKA, L. et al. Artfds–advanced railway track fault detection system using machine learning. In: SUMA, V. et al. (Ed.). *Inventive Systems and Control*. Singapore: Springer Nature Singapore, 2022. p. 609–624. ISBN 978-981-19-1012-8. Citado 2 vezes nas páginas 17 e 26.

LASISI, A.; ATTOH-OKINE, N. Principal components analysis and track quality index: A machine learning approach. *Transportation Research Part C: Emerging Technologies*, v. 91, 6 2018. ISSN 0968090X. Citado 2 vezes nas páginas 20 e 22.

LIAO, Y. et al. Prediction models for railway track geometry degradation using machine learning methods: A review. *Sensors*, v. 22, n. 19, 2022. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/22/19/7275>. Citado 4 vezes nas páginas 17, 18, 19 e 20.

LINGAMANAIK, S. et al. Using instrumented revenue vehicles to inspect track integrity and rolling stock performance in a passenger network during peak times. *Procedia Engineering*, v. 188, p. 424–431, 2017. ISSN 1877-7058. Structural Health Monitoring - From Sensing to Diagnosis and Prognosis. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877705817320568>. Citado 2 vezes nas páginas 8 e 24.

MATSUMOTO, C. Y. A importância do banco de dados em uma organização. *Revista de Ciências Empresariais*, v. 3, n. 1, p. 45–55, Jan 2006. Citado na página 17.

NUNEZ, A. et al. Facilitating maintenance decisions on the dutch railways using big data: The ABA case study. In: *2014 IEEE International Conference on Big Data (Big Data)*. Washington, DC, USA: IEEE, 2014. p. 48–53. ISBN 978-1-4799-5666-1. Citado 2 vezes nas páginas 10 e 27.

OpenStreetMap. *Planet dump retrieved from https://planet.osm.org* . 2017. <https://www.openstreetmap.org>. Citado 3 vezes nas páginas 32, 33 e 34.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 70.

RIBEIRO, F. S. *Contribuição para análise do custo do ciclo de vida de um sistema de gerência de pavimento ferroviário*. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado 3 vezes nas páginas 8, 9 e 24.

SAITO, L. S.; RIBEIRO, M. F. N. L. *Visualização de dados processados por machine learning para manutenção da EFVM*. Dissertação (Trabalho de Formatura) — Universidade de São Paulo (USP), 2021. Citado na página 43.

SALIERNO, G. et al. An architecture for predictive maintenance of railway points based on big data analytics. In: DUPUY-CHESSA, S.; PROPER, H. A. (Ed.). *Advanced Information Systems Engineering Workshops*. [S.l.]: Springer International Publishing, 2020. p. 29–40. ISBN 978-3-030-49165-9. Citado 2 vezes nas páginas 14 e 28.

SCHUMACHER, A.; EROL, S.; SIHN, W. A maturity model for assessing industry 4.0 readiness and maturity of manufacturing enterprises. *Procedia CIRP*, The Author(s), v. 52, p. 161–166, 2016. ISSN 22128271. Disponível em: <http://dx.doi.org/10.1016/j.procir.2016.07.040>. Citado na página 9.

SHARMA, S. et al. Data-driven optimization of railway maintenance for track geometry. *Transportation Research Part C: Emerging Technologies*, v. 90, 5 2018. ISSN 0968090X. Citado na página 28.

SILVA, P. H. O. *Protótipo para Monitoramento da Dinâmica de Vagões Ferroviários*. Dissertação (Mestrado) — Universidade Federal de São João del-Rei - Programa de Pós-Graduação em Engenharia Elétrica, 2019. Citado na página 24.

SKANSI, S. *Introduction to Deep Learning*. 1. ed. [S.l.]: Springer Cham, 2018. (Undergraduate Topics in Computer Science). Citado 4 vezes nas páginas 18, 19, 22 e 23.

SOBRINHO, O. G.; BERNUCCI, L. L. M.; CORRÊA, P. L. P.; MOTTA, R. dos S.; MACHICAO, J.; JUNQUEIRA, A. S.; COSTA, R. C. da; QUEIROZ, W. Dias de; SILVA, T. C. G. d. M.; FERRAZ, P. L.; CASSARO, L.; OLIVEIRA, L. Big data analytics in

support of the under-rail maintenance management at Vitória – Minas Railway. In: *2021 IEEE International Conference on Big Data (Big Data)*. Orlando, FL, USA: IEEE, 2021. p. 6026–6028. Nenhuma citação no texto.

SOBRINHO, O. G.; BERNUCCI, L. L. M.; CORRÊA, P. L. P.; MOTTA, R. d. S.; MACHICAO, J.; JUNQUEIRA, A. S.; COSTA, R. C. da; SILVA, T. C. G. d. M.; QUEIROZ, W. D. de; FERRAZ, P. L. Inteligência artificial e processamento de linguagem natural no monitoramento da estrada de ferro Vitória – Minas. In: *10º Congresso Rodoferroviário Português*. Lisboa, Portugal: Centro Rodoferroviário Português, 2022. Nenhuma citação no texto.

TUDEIA, B. et al. *Vagão Instrumentado - Projeto do Sistema de Aquisição de Dados*. [S.l.], 2019. 1-50 p. Disponível em: <www.isqbrasil.com.br>. Citado 4 vezes nas páginas 10, 25, 26 e 36.

WESTON, P. et al. Perspectives on railway track geometry condition monitoring from in-service railway vehicles. *Vehicle System Dynamics*, v. 53, 7 2015. ISSN 0042-3114. Citado 4 vezes nas páginas 9, 24, 25 e 26.

XU, G. et al. Data-driven fault diagnostics and prognostics for predictive maintenance: A brief overview*. In: *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. Vancouver, BC, Canada: IEEE, 2019. p. 103–108. ISBN 978-1-7281-0356-3. Citado 7 vezes nas páginas 10, 17, 18, 19, 20, 21 e 23.