NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA

CATARINA MENA RIBEIRO DE BRITO AFONSO

Licenciatura em Matemática Aplicada à Empresa e Tecnologia

ANÁLISE DE SÉRIES TEMPORAIS DE ABSENTISMO NO SETOR FERROVIÁRIO

DISSERTAÇÃO, 2022/2023

MESTRADO EM ANÁLISE E ENGENHARIA DE BIG DATA

Universidade NOVA de Lisboa
Setembro, 2023

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA

ANÁLISE DE SÉRIES TEMPORAIS DE ABSENTISMO NO SETOR FERROVIÁRIO

DISSERTAÇÃO, 2022/2023

CATARINA MENA RIBEIRO DE BRITO AFONSO

Licenciatura em Matemática Aplicada à Empresa e Tecnologia

Orientadora: Regina Bispo
Professora Auxiliar, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa

Coorientador: Ricardo Saldanha
Diretor do Departemento de Inovação, SISCOG - Sistemas Cognitivos, SA

Júri

Presidente: Paula Amaral
Professora Associada, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Arguente: Dulce Gomes
Professora Associada, Universidade de Évora

Vogal: Regina Bispo
Professora Auxiliar, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa

MESTRADO EM ANÁLISE E ENGENHARIA DE BIG DATA

Universidade NOVA de Lisboa
Setembro, 2023

Análise de séries temporais de absentismo no setor ferroviário

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo NOVAthesis (v6.10.5) [33].

Para os meus pais que sempre puxaram por mim para dar o melhor de mim.
Ao meu tio João Pedro, sem ele não estaria onde estou.
Ao André pelo apoio.

Agradecimentos

Quero agradecer à Faculdade de Ciências e Tecnologia da Universidade NOVA (FCT-NOVA) de Lisboa e à empresa SISCOG, por me deixarem fazer parte deste projeto e por acreditarem nas minhas capacidades.

Quero também agradecer à minha orientadora Regina Bispo, da FCT-NOVA, pelo apoio que me tem dado e por me desafiar todos os dias. De igual modo, quero agradecer ao meu co-orientador Ricardo Saldanha, da SISCOG, por me ter orientado e expandido os meus conhecimentos sobre o setor ferroviário. Ao Gonçalo Matos e ao Luís Albino, da SISCOG, quero agradecer a preciosa ajuda no desenvolvimento do projeto.

Por fim, quero enfatizar a importância da contribuição dos meus pais, do meu namorado e dos meus amigos, por me terem apoiado e motivado neste processo.

«Perfection is the enemy of progress.» (Winston Churchill)

Resumo

O absentismo constitui um problema para os operadores ferroviários, pois pode ter repercussões, como, por exemplo, o cancelamento de viagens. Pela natureza do serviço que prestam, os operadores ferroviários efetuam planeamentos operacionais. Porém, como alguns tipos de ausências dos tripulantes, nomeadamente aquelas motivadas por razões de doença, não se conseguem prever, os referidos operadores procuram outras ferramentas de apoio à decisão que possam ajudar a prevenir os impactos provocados pelo absentismo.

Esta Dissertação tem como finalidade desenvolver um modelo de previsão de ausências não programáveis que tenha capacidade de estimar o número de tripulantes ausentes por motivo de doença e/ou apoio à família, com base no histórico de atividade planeada e realizada dos tripulantes, para que possa ser englobado no planeamento.

Os dados foram disponibilizados por parte de um operador ferroviário da Europa do Norte, um cliente da empresa SISCOG. Neste caso, analisaram-se dados diários sobre a percentagem de maquinistas e de revisores ausentes de uma base operacional grande. Realizou-se um pré-processamento e uma análise exploratória, para perceber as características de cada série.

No final, foram comparados os resultados entre o modelo regressão dinâmica harmónica (RDH) e os algoritmos Prophet e NeuralProphet, com o método da janela fixa e método da janela móvel. Quanto à janela fixa, o NeuralProphet destacou-se em termos de precisão para a série dos revisores, com um MAPE de 12.86%. Para a série dos maquinistas, não se conseguiu concluir qual o método mais adequado, uma vez que as métricas eram muito semelhantes. No entanto, sugere-se que o Prophet, com um MAPE de 20.70%, seria a escolha mais indicada para a série devido ao seu custo computacional. Porém, os resultados da janela móvel revelaram que o modelo de RDH obteve um bom desempenho ao longo dos diferentes horizontes de previsão, em comparação com o NeuralProphet. Quanto aos dados dos maquinistas, os modelos obtiveram um MAPE pequeno no último horizonte de previsão, em comparação com os restantes.

Palavras-chave: Absentismo, Operador ferroviário, Planeamento operacional, Tripulantes, Métodos de Previsão, Séries temporais

Abstract

Absenteeism is a problem for railway operators because it can have repercussions such as cancellation of trips. Due to the nature of the service delivered, railway operators carry out operational planning. However, as some types of absences of crew members, particularly those caused by illness, cannot be predicted, they seek other decision support tools that can help prevent the impacts caused by absenteeism.

This dissertation aims to develop a prediction model for unscheduled absences that can estimate the number or percentage of crew members unexpectedly absent due to illness and/or family support, for a given time period, based on the planned and performed activity history of the crew members, so that it can be incorporated into the planning.

The data was made available from a railway operator in Northern Europe, a customer of the SISCOG company. In this case, we analyzed data on the percentage of absent train drivers and guards from selected operational bases. Data was pre-processed, and then an exploratory analysis was carried out, to understand the characteristics of each series.

Finally, the results produced by the dynamic harmonic regression (DHR) model and Prophet and NeuralProphet algorithms were compared, by using the fixed window and sliding window methods. Regarding the fixed window method, the NeuralProphet stood out for its precision for the series related to the guards, with a MAPE of 12.86%. For the series related with train drivers we could not conclude which method was the most appropriate, as the metrics were very similar. However, we decided that Prophet, with a MAPE of 20.70%, would be the most appropriate choice for this series, because of its computational cost. Additionally, the sliding window results revealed the DHR model as having a good performance along the diferent prediction horizons, when compared with NeuralProphet. regarding the train drivers data, the models obtained a small MAPE value in the last prediction horizon, compared to the less distant prediction horizons.

Keywords: Absenteeism, Railway operators, Planning, Crew, Forecasting Methods, Time series

Índice

Anexos

Índice de Figuras

Índice de Tabelas

Glossário

Artificial Neural Network	Modelo que se baseia na estrutura do cérebro humano, imitando a forma como os neurónios biológicos sinalizam entre eles. As *artificial neural networks* são constituídas por uma camada de nós, contendo uma camada de entrada, uma ou mais camadas ocultas, e uma camada de saída [12]. *(pp. xviii, 12)*
DataFrame	Estrutura de dados que organiza os dados numa tabela bidimensional em linhas e colunas [13]. *(p. 46)*
Support Vector Machine	Modelo de classificação (também usado em regressão), baseado na definição de um hiperplano num espaço N-dimensional (N representa o número de variáveis) [18]. *(pp. xix, 12)*
Synthetic Minority Oversampling Technique	Algoritmo aplicado em conjuntos de dados que tenham uma representação de classes desigual, para controlar o desequilíbrio entre classes [53]. *(pp. xix, 9)*
feature selection	Processo de seleção de variáveis no desenvolvimento de um modelo preditivo[6]. *(pp. 9, 12, 13)*
features	Propriedade mensurável ou característica de um objeto [9]. Pode também ser denominada, também, por variável ou atributo. *(pp. 10–13, 49)*
fuzzy neural network	Modelo que combina *Artificial Neural Network* e *lógica fuzzy [10] (p. 12)*

imbalanced data Conjunto de dados que tem uma representação de classes desigual [21]. *(pp. 9, 11)*

multilayer perceptron É uma *artificial neural network* totalmente conetada. Cada neurónio de uma camada recebe como *input*, o *output* de todos os neurónios da camada anterior[29]. *(p. 12)*

naive Bayes Modelo probabilístico, utilizado em tarefas de classificação, baseado no teorema de *Bayes* [43]. *(pp. 10, 12)*

outlier Observação(ões) que apresenta um grande afastamento das restantes observações, ou que é inconsistente com estas. *(pp. xii, 12, 23, 30–32, 51, 57–59, 78)*

pickle Processo de conversão de um objeto em memória num fluxo de *bytes*, que pode ser armazenado como um arquivo binário no disco. Quando se carrega num programa *Python*, esse arquivo binário é de novo um objeto *Python*. [28]. *(p. 46)*

validação cruzada *k-fold* Consiste em repartir em k sub-conjuntos ou *folds*. Cada *fold* contém o mesmo número de exemplos. Um dos *folds* é usado para a validação e os restantes para treino. O modelo é treinado com os *folds* de treino e avaliado no *fold* de validação [35]. *(p. 33)*

Siglas

ACF *Autocorrelation Function (pp. xi, 24–26, 56, 67)*
ADF *Augmented Dickey-Fuller (pp. 26–28, 56)*
AIC *Akaike's Information Criterion (pp. 38–40)*
AICc *Akaike's Information Corrected Criterion (pp. 38, 39)*
ANN *Artificial Neural Network (p. 12)*

BIC *Bayesian Information Criterion (p. 38)*

CFS *Correlation Feature Set (pp. 9, 10)*

DF *Dickley-Fuller (p. 27)*
DL *Deep Learning (pp. 10, 12, 14, 44)*

IQR *Interquartile range (pp. xii, 58)*

KNN *K-nearest neighbors (p. 10)*
KPSS *Kwiatkowski–Phillips–Schmidt–Shin (pp. 26, 28, 39, 56)*

LSTM *Long Short-Term Memory (pp. 13, 14)*

MAE *Mean Absolute Error (pp. xiii, xiv, 31, 32, 63, 66–68, 70, 73–76, 87)*
MAPE *Mean Absolute Percentage Error (pp. vi, vii, xiii, xiv, 32, 63, 66, 68, 70, 73–77, 87)*
ML *Machine Learning (pp. 10, 11, 13, 14, 33)*
MSTL *Multiple Seasonal-Trend decomposition using LOESS (pp. xi, xii, 23, 52–55)*

NaN *Not a Number (p. 48)*

PACF *Partial Autocorrelation Function (p. 26)*

RDH regressão dinâmica harmónica *(pp. vi, xii–xv, 41, 42, 62–64, 66–68, 70, 71, 73–75, 77, 88, 89)*

RMSE *Root Mean Squared Error (pp. xiii, xiv, 12, 31, 63, 66, 68, 70, 73–76, 87)*

SMAPE *Symmetric Mean Absolute Percentage Error (p. 32)*

SMOTE *Synthetic Minority Oversampling Technique (pp. 9–11)*

STL *Seasonal and Trend decomposition using Loess (p. 23)*

SVM *Support Vector Machine (pp. 12, 13)*

1

Introdução

O presente trabalho foi desenvolvido no âmbito do estágio na empresa SISCOG[1], subordinado ao tema "**Análise de séries temporais de absentismo no setor ferroviário**". Este estágio surge da necessidade da SISCOG melhorar a capacidade de prever ausências em relação a um estudo embrionário que se realizou internamente, resultando na publicação [39]. O presente documento relata o trabalho desenvolvido no seguimento desse estudo, conduzindo à realização de uma Dissertação de Mestrado em Análise e Engenharia de *Big Data* da FCT-NOVA.

Uma empresa de transporte ferroviário tem como principal atividade prestar serviços de transporte de passageiros e/ou de mercadorias por caminho de ferro [44], através da circulação de comboios ao longo de uma rede que serve várias localidades de uma região ou país. Esta tem um papel fundamental nas ligações urbanas, suburbanas e interurbanas, e como tal, necessita de criar condições para cumprir as suas funções [57].

Todos os comboios precisam de maquinistas e de alguns revisores para poderem circular e assim realizar os serviços de transporte pretendidos. Também, para garantir a operação dos comboios, é essencial cumprir um planeamento operacional. No entanto, se ocorrerem imprevistos, como por exemplo, a **ausência de trabalhadores** para um determinado serviço, esse plano operacional deve ser reajustado tão rapidamente quanto possível. Para estas situações, devem ser consideradas ferramentas de despacho para apoiar o processo de decisão.

Para além dos imprevistos, que são frequentes, os operadores ferroviários, um pouco por todo o mundo, têm sido forçados a cortar diversos serviços devido à escassez de recursos humanos. De uma forma geral, na Europa, a procura de viagens aumentou após o levantamento das restrições da pandemia, intensificando a pressão sobre a indústria ferroviária e as perturbações enfrentadas pelos passageiros. Por exemplo, as consequências da falta de trabalhadores nos operadores ferroviários Nacionais Holandeses, fizeram-se sentir no mês de setembro de 2022. Durante esse mês, em que também foram feitos grandes ajustes aos horários, os comboios encontravam-se demasiado ocupados, sem ser devido a

[1] A SISCOG - Sistemas Cognitivos, SA é uma empresa de *software* que fornece sistemas de apoio à decisão para o planeamento e gestão de recursos no setor dos transportes tendo por objetivo melhorar a produtividade e qualidade do negócio e a satisfação dos seus empregados e clientes [49].

circunstâncias inesperadas, tais como uma colisão ou uma perturbação [40]. O segundo maior operador ferroviário do Reino Unido, *Transpennine Express*, também sofreu no verão várias interrupções generalizadas, que quase colapsaram os serviços de *Avanti West Coast*, uma das linhas ferroviárias intermunicipais mais movimentadas do Reino Unido, devido à falta de maquinistas [19].

Uma forma de reduzir os riscos provocados pelo absentismo e pela falta de recursos humanos, e que constitui também o tema principal desta Dissertação, é a previsão de ausências, um instrumento cada vez mais importante para ajustar de maneira proativa a capacidade disponível às necessidades de força laboral, ajudando num melhor planeamento dos turnos, folgas e ausências.

1.1 A importância do planeamento no domínio ferroviário

O planeamento operacional é de vital importância para o domínio dos transportes. Pode-se dizer que sem planeamento não seria possível fornecer serviços de transporte regulares tanto no domínio ferroviário como noutros domínios. Sem planeamento não seria possível oferecer um serviço de transporte baseado em horários publicados (porque os horários são já o resultado de um planeamento), não seria possível saber se à partida há recursos suficientes para fornecer o serviço pretendido, não seria possível coordenar as operações necessárias à prestação do serviço. Além disso, uma operação que não seguisse um plano previamente efetuado correria o risco de ser extremamente ineficiente e sujeito a muitas falhas, como por exemplo viagens que deixariam de se realizar por não haver recursos disponíveis à hora e no local em que são precisos.

O planeamento é executado numa sequência de etapas. Essas etapas podem ser classificadas de diversas formas, mas o critério mais comum tem como base a extensão do horizonte temporal do planeamento. Assente este critério, as etapas podem ser divididas nas seguintes fases: estratégica, tática, operacional e de planeamento a curto prazo. Esta hierarquia no tempo também permite que os produtos resultantes de uma fase de planeamento possam servir como *inputs* para fases posteriores, e que as decisões anteriores possam ser revistas de acordo com a informação atualizada [34].

- **Planeamento estratégico**: tem um horizonte de planeamento de vários anos ou mesmo décadas. O foco principal é o planeamento da capacidade [34].

- **Planeamento tático**: aloca a capacidade disponível com um horizonte de planeamento de 2 meses até um ano [34].

- **Planeamento operacional**: estabelece os planos bastante detalhados. O horizonte de planeamento varia entre 3 dias a 2 meses. Trata-se principalmente de ajustar os planos táticos para as próximas semanas [34].

- **Planeamento de curto prazo**: inclui problemas que surgem quando as operações de comboio se realizam (despacho em tempo real) ou pouco antes disso. Abrange etapas de planeamento com um horizonte de, no máximo, 3 dias [34].

Para além desta categorização, as etapas podem ser agrupadas com base num determinado objetivo: planeamento de linhas de serviço (*line planning*), planeamento de horários (*timetabling*), planeamento do material circulante (*rolling stock scheduling*) ou planeamento da tripulação (*crew scheduling*) [34].

- ***Line planning***: é o problema de determinar, dado o número de passageiros interessados em viajar entre cada possível par origem-destino, as linhas de serviço mais apropriadas, as suas frequências horárias e os seus padrões de paragens, para satisfazer a procura. Estas linhas de serviço são compostas por uma rota de um comboio e um padrão de paragens que especifica em que estações o mesmo pára [38].
- ***Timetabling***: corresponde à atribuição de tempos aos eventos de serviço (partidas/chegadas) resultantes do processo de *Line planning* [38].
- ***Rolling stock scheduling***: determina, entre outros aspetos, quais locomotivas e carruagens de passageiros são atribuídas aos veículos para os serviços anteriormente planeados [38].
- ***Crew scheduling***: diz respeito à atribuição de membros da tripulação (maquinistas e revisores) aos serviços associados ao *Rolling stock scheduling* [38].

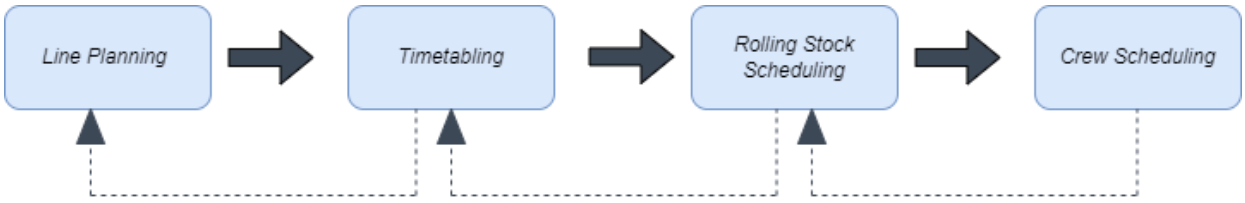

Figura 1.1: Sequência de etapas do planeamento operacional (adaptado de [38]).

1.1.1 Disrupções no transporte de passageiros

Numa viagem de comboio, os passageiros esperam chegar ao seu destino num determinado tempo estipulado pelo horário publicado. Naturalmente, eventos imprevisíveis podem ocorrer, causando atrasos ou até mesmo cancelamentos de comboios, e, consequentemente, os passageiros demoram mais do que o esperado. Exemplos de eventos imprevisíveis podem ser: avarias nas infra-estruturas, avarias no material circulante, acidentes e/ou condições meteorológicas severas [45].

Tais acontecimentos, bem como a perturbação que geram na operação, denominam-se por **disrupções**. Para as operações ferroviárias, define-se **disrupção** como um evento ou uma série de eventos (normalmente relacionados com a indisponibilidade de recursos)

que alteram os pressupostos com base nos quais os planos operacionais foram feitos ao ponto de os tornar inoperacionais. Por exemplo, um bloqueio da via inviabiliza o horário planeado, ou a **ausência de membros da tripulação** devido a doença pode levar a cancelamentos, dependendo do número de tripulantes faltosos [45].

Portanto, é muito importante limitar a consequências provocadas por estas perturbações. Um problema muito comum na gestão do tráfego ferroviário é que, devido às fortes interdependências na rede ferroviária e à eventual falta de robustez dos horários [2], é muito provável que as disrupções se propaguem pela rede no espaço e no tempo. Este fenómeno denomina-se por *knock-on effect* [45].

De modo a minimizar estes riscos e facilitar o planeamento, é necessário utilizar outros meios, como por exemplo, **Métodos de Previsão**, que contribuam para um replaneamento rápido e otimizado, por forma a realizar mudanças a tempo de publicar um novo plano de trabalho [48].

Neste trabalho, entre algumas das disrupções anteriormente mencionadas, será apenas abordado o **absentismo** dos maquinistas e revisores.

1.2 Absentismo

O absentismo refere-se à ausência de um funcionário no trabalho. As possíveis causas para estas ocorrências podem ser ausências programáveis, como férias, ou ausências não programáveis como nos casos de doença, emergências familiares, entre outras.

Embora seja expectável a falta de comparência de forma ocasional, o absentismo pode-se tornar um problema se acontecer de forma sistemática e com múltiplos trabalhadores. Pode ser impactante, principalmente, na produtividade da indústria e, a longo prazo, nas suas finanças. Os custos associados ao absentismo podem ser diretos ou indiretos. O custo financeiro direto advém, por exemplo, da mão de obra de reserva que é preciso ter para lidar com imprevistos, pagamento de horas extra e a falta de produtividade quando não há reposição. Os custos indiretos podem incluir a má qualidade de produtos, cancelamento dos serviços e insatisfação dos clientes [15].

Tendo em conta que as ausências não programáveis são mais difíceis, mas mais interessantes de prever, este estudo irá apenas focar-se nas ausências incentivadas por **doenças** ou **apoio à família**. Seguem-se algumas justificações possíveis que categorizam os dois casos:

- **Ausência por doença** incluindo, por exemplo, lesão em serviço e dias de tratamento.
- **Ausência por apoio à família** incluindo, por exemplo, falta para assistência a membro do agregado familiar, falecimento de cônjuge ou familiar e faltas do trabalhador, devido a gravidez ou a assistência a menor.

[2] Os horários encontram-se altamente otimizados de forma a reduzir custos. No entanto, esta eficiência é conseguida à custa de folgas reduzidas. Como as folgas são reduzidas, o horário tem pouca margem de manobra para absorver atrasos.

1.3 Replaneamento em caso de ausência dos tripulantes

Uma possível solução para reduzir os impactos provocados pelo absentismo no trabalho é estimar, tão antecipadamente quanto possível, o número de tripulantes inesperadamente ausentes.

Esta previsão seria de grande utilidade para apoiar a decisão dos planeadores no contexto do planeamento e gestão do trabalho de tripulantes, em particular na negociação de pedidos de dias de folga e no planeamento de turnos de reserva[3] [39]. Por exemplo, a **nível tático**, haverá uma melhoria no planeamento de férias e de reservas — reservas em excesso aumentam desnecessariamente os custos, enquanto que um número insuficiente de reservas pode originar situações em que não há recursos suficientes para fazer o que é preciso fazer, o que leva a mais viagens canceladas e despesas com compensações e horas extraordinárias. A **nível operacional** irá apoiar o processo de negociação, ajudando o planeador a não aceitar pedidos de folga em dias em que está previsto haver pouca capacidade disponível face às necessidades.

Para um operador ferroviário, poderia ser interessante, em termos de horizonte temporal, fazer previsões correspondentes às próximas 48 horas — permite dar tempo ao planeador para tentar negociar com os tripulantes o adiamento ou a compra de alguma folga, para o caso em que se preveja uma grande escassez de recursos disponíveis —, entre uma semana a um mês — para **planeamentos de curto prazo**, pode-se recomendar ao planeador não programar ausências ou folgas em dias em que se espera um nível de absentismo elevado ——, ou até um ano —— permite um melhor planeamento de longo prazo de reservas e formações, e um processo mais eficiente de planeamento anual de férias [39].

1.3.1 Produto CREWS

A solução referida para o replaneamento em caso de ausência dos tripulantes enquadra-se num dos produtos — o CREWS — que fornecem apoio à decisão para o planeamento e gestão do trabalho do pessoal circulante e não circulante, desenvolvido pela SISCOG.

A SISCOG Suite — ONTIME, FLEET, RAILNODE e CREWS — é composta por vários produtos independentes, mas integrados, que tratam do ciclo de planeamento e gestão de recursos em empresas de transporte. São soluções modulares que permitem desde o planeamento de longo prazo ao despacho em tempo real e análise pós-operação [50].

O **CREWS**, produto que se enquadra no contexto deste problema, ajuda a criar e a gerir planos de pessoal otimizados, incluindo tripulação e pessoal local, considerando horários e planos operacionais de veículos, competências e preferências do pessoal e restrições/regras laborais e operacionais. Para além disso, produz turnos e escalas cíclicas de longo prazo e planos de calendário de curto prazo. No planeamento de curto prazo, as

[3] Um turno de reserva é um intervalo de tempo em que o tripulante aguarda na sua base operacional, na eventualidade de ser necessário substituir um tripulante ausente ou para receber trabalho adicional resultante de alterações de última hora aos horários.

pessoas podem ser atribuídas a turnos seguindo padrões irregulares ou seguindo padrões de rotação regular baseados nas escalas cíclicas [50].

1.4 Objetivos

Como referido anteriormente, o planeamento deve assegurar, em cada local, em cada dia e em cada hora do dia, a presença de um número suficiente de tripulantes, para garantir os melhores padrões de qualidade de serviço, segurança ferroviária e segurança de pessoas e bens.

O número de tripulantes suficiente para assegurar a operação inclui os tripulantes de reserva, sendo estes determinados a partir da previsão de ausências.

Uma vez estabelecidos estes conceitos e a utilidade de se conhecer antecipadamente o número de tripulantes ausentes, **o objetivo central deste estudo é desenvolver um modelo preditivo de ausências não programáveis** capaz de estimar o número de trabalhadores ausentes por doença e/ou apoio à família tendo como base um conjunto de dados históricos de atividade planeada e de ausências, providenciado por um operador ferroviário da Europa do Norte, visando incorporá-lo futuramente no produto CREWS. Pretende-se, também, que as previsões sejam realizadas separadamente para cada tipo de função desempenhada pelo pessoal tripulante, isto é, maquinista ou revisor, e para cada base operacional.

Este trabalho desenvolveu-se em duas fases. Na primeira fase, que se realizou no período de 3 de outubro de 2022 a 28 de fevereiro de 2023 — Preparação da Dissertação — a signatária desenvolveu o trabalho introdutório do tema e apresentou um relatório incluindo um plano de trabalho a realizar para concretizar a Dissertação. O projeto foi avaliado e aprovado numa sessão pública, o que permitiu o seu seguimento. Na segunda fase, que se realizou no período de 28 de fevereiro de 2023 a 30 de setembro de 2023 — Dissertação — focou-se na parte prática do problema, onde se efetuou a análise exploratória dos dados e a construção de vários modelos preditivos, para a concretização dos objetivos da Dissertação.

1.5 Estrutura do relatório

No primeiro capítulo foi efetuada, de forma mais detalhada, uma breve introdução sobre a importância do planeamento no setor ferroviário, de maneira a contextualizar o tema. Seguidamente, foi feita uma apresentação sobre os problemas causados pelo absentismo, a solução e as vantagens do replaneamento rápido no caso de ausência dos tripulantes, de como esta solução se insere no produto CREWS, no contexto de planeamento e gestão de pessoal. Para concluir, foram expostos os objetivos da Dissertação.

Posteriormente, no segundo capítulo, é contemplado o Estado da Arte, onde se realiza a apresentação de trabalhos científicos relacionados com o tema proposto, abordando o

pré-processamento dos dados e os algoritmos utilizados, bem como as limitações com as quais os respetivos autores se depararam.

No terceiro capítulo, é relatada a metodologia, desenvolvendo o contexto e a base conceptual apresentando simultaneamente a metodologia aplicada.

O quarto capítulo aborda os dados disponíveis, bem como o procedimento para seleção e o tratamento dos mesmos, principalmente, a sua conversão em séries temporais, para analisar a evolução do absentismo ao longo do tempo. Também se realizou uma análise exploratória dos dados, com o propósito de recolher informações relevantes para tomada de decisão quanto à escolha do modelo consoante as características dos mesmos.

No quinto capítulo, são apresentados os resultados produzidos pelos métodos e, por fim, no sexto capítulo as conclusões do trabalho realizado e possíveis trabalhos futuros.

2

Estado da Arte

Neste capítulo são resumidos alguns trabalhos relacionados com o estudo do absentismo, descrevendo-se de forma sucinta para cada estudo os dados utilizados, as técnicas para o pré-processamento dos dados, os modelos usados para realizar as previsões, bem como a descrição de algumas limitações que os autores enfrentaram.

No trabalho apresentado por Skorikov et al [51], o elevado absentismo dos funcionários no trabalho pode ser prejudicial a nível económico e da produtividade. Para minimizar os riscos provocados pelo absentismo, os autores recorreram a técnicas de *data mining* para prever a ausência dos trabalhadores. Em específico, os autores analisaram um conjunto de dados proveniente de uma empresa de correio do Brasil. O *dataset* utilizado foi primeiramente publicado em Martiniano et al [36] e, neste momento, encontra-se disponível publicamente no repositório *UCI Machine Learning Database*[1]. O conjunto de dados inclui 740 instâncias e 21 atributos, referentes ao período compreendido entre julho de 2007 e julho de 2010. Contém informação relativa aos hábitos dos trabalhadores, viagens, trabalho, detalhes sobre as ausências, entro outros. A Tabela 2.1 descreve, em detalhe, os atributos contidos no *dataset*.

Tabela 2.1: Atributos dos dois *result sets* exportados.

Atributo	Tipo	Conteúdo/Observações
`ID`	*int*	Número de identificação trabalhador
`Reason for absence`	*int*	A razão do absentismo (código numérico entre 1 e 28)
`Month of absence`	*int*	Mês do absentismo do trabalhador (entre 1 e 12)
`Day of the week`	*int*	Entre 1 e 7 em que 1 representa domingo
`Seasons`	*int*	As estações do ano (entre 1 e 4)
`Travel expense`	*int*	Custo da viagem de casa para o trabalho (em real)
`Distance`	*int*	Distância entre casa e trabalho (em km)
`Service time`	*int*	Meses de serviço do trabalhador (antiguidade na empresa)

[1] `https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work`

`Age`	*int*	Idade do trabalhador (em anos)
`Workload / day`	*float*	Carga de trabalho média diária
`Hit target`	*int*	Objetivo para os trabalhadores
`Disciplinary failure`	*boolean*	1, se ocorreu alguma falha disciplinar no passado, ou 0 caso contrário
`Education`	*int*	Entre 1 e 4 dependendo do nível de educação do trabalhador
`Children`	*int*	Número de filhos do trabalhador
`Social drinker`	*boolean*	1 se bebe socialmente, ou 0 caso contrário
`Social smoker`	*boolean*	1 se fuma socialmente, ou 0 caso contrário
`Pet`	*int*	Número de animais de estimação do trabalhador
`Weight`	*int*	Peso do trabalhador (em kg)
`Height`	*int*	Altura do trabalhador (em cm)
`BMI`	*int*	Índice de massa corporal do trabalhador
`Absenteeism time`	*int*	Duração do absentismo (em horas)

Na fase de pré-processamento, Skorikov et al [51] efetuaram uma limpeza dos dados e verificaram a existência de valores omissos. Seguidamente, realizaram a discretização dos dados, isto é, converteram os dados contínuos em categorias discretas e, por fim, converteram os dados numéricos para dados categóricos, o que corresponde à fase de transformação dos dados. Posteriormente, os autores dividiram o conjunto de dados em 80% dos dados para o conjunto de treino e os restantes 20% para o conjunto de teste.

Relativamente à variável que se pretende prever, `'Absenteeism time'`, os dados são contínuos, porém, segundo Skorikov et al [51], seria mais apropriado se classificassem o tempo de absentismo em categorias, uma vez que permitiria ao modelo prever diferentes "graus" de absentismo no conjunto de teste. Para tal, os autores exploraram a distribuição de frequência recorrendo a um histograma e concluíram que o gráfico demonstrava visivelmente a existência de três classes (classe A - 0 horas, classe B - [1-15] horas, classe C - [16, 120] horas). Esta categorização resultou num desequilíbrio nos dados (*imbalanced data*) e isso pode ser prejudicial para a performance preditiva, tendo em conta que uma das classes pode receber um tratamento preferencial relativamente a uma classe que esteja em minoria. Para tratar deste desequilíbrio, Skorikov et al [51] aplicaram a técnica *Synthetic Minority Oversampling Technique* (SMOTE) no conjunto de treino, o que corresponde a gerar dados sintéticos para a classe que está em minoria, de modo a equilibrá-la com as restantes classes. Para além disso, os autores dividiram o processo de treino num conjunto com o SMOTE aplicado, e um conjunto sem o mesmo aplicado.

Skorikov et al [51] procederam à fase de *feature selection*, onde utilizaram a técnica *Correlation Feature Set* (CFS) e o *relief attribute evaluator*, baseada no algoritmo *relief*.

Relativamente à metodologia experimental, os autores utilizaram três abordagens; na Experiência A usam as 4 *features* mais influentes, `'Month of absence'`, `'age'`, `'Disciplinary Failure'` e `'Social Drinker'`, que se obtiveram a partir da aplicação da técnica CFS no conjunto de treino. Na Experiência B utilizaram os 19 atributos do *dataset*. E na Experiência C, como `'Disciplinary Failure'` obteve um maior *feature score* no algoritmo *relief* e um maior *information gain*, Skorikov et al [51] consideraram apenas utilizar esta variável para esta experiência.

Tendo em conta que o *dataset* é relativamente pequeno, a aplicação de técnicas de *Deep Learning* (DL) não seriam adequadas para este caso. Para cada experiência usaram diferentes classificadores de *Machine Learning* (ML), nomeadamente, *ZeroR*, *naive Bayes*, *K-nearest neighbors* (KNN) (com distância euclidiana, *Manhattan* e *Chebyshev*) e *tree-based (J48)*. Como referido anteriormente, cada classificador foi treinado com um conjunto de treino com a aplicação da técnica de SMOTE e com um conjunto sem a aplicação da mesma.

Para avaliar o desempenho dos modelos, Skorikov et al [51] utilizaram métricas, tais como *Precision*, *Recall*, *F-measure*, *ROC area* e *Accuracy*.

No que se refere aos resultados experimentais, o modelo que obteve a melhor performance foi o classificador KNN utilizando a distância *Chebyshev*, com uma *accuracy* aproximadamente de 92.3%, na experiência B com a aplicação da técnica SMOTE no conjunto de treino.

Outros trabalhos publicados utilizaram o mesmo *dataset* [36] e aplicaram uma metodologia semelhante para prever o absentismo no trabalho, por exemplo, Wahid et al [61], também utilizaram o atributo `'Absenteeism time'` como variável *target* do modelo. Tal como no trabalho anterior, Wahid et al [61] transformaram o problema numa tarefa de classificação, pois concluíram que os dados correspondentes a esse atributos são muito "dispersos", o que pode tornar mais complexa a obtenção de bons resultados na fase de previsão. Como tal, converteram o atributo contínuo numa coluna categórica, em que cada classe representa um determinado período de tempo. Esta categorização foi baseada nos critérios estabelecidos pela *International Labour Standards on Working Time*[2] da *International Labor Organization*.

Wahid et al [61] dividiram o conjunto dos dados em conjunto de treino (80%) e conjunto de teste (20%), e treinaram com os seguintes algoritmos de ML: *Decision Tree, Gradient Boosted Tree, Random Forest* e *Tree Ensemble*. Os autores concluíram que os classificadores baseados em árvores podem proporcionar uma estrutura bem definida que permite perceber o fluxo de decisões entre os valores e, por sua vez, investigar os possíveis resultados das escolhas dessas opções.

Para avaliar a performance dos modelos, Wahid et al [61] recorreram a 7 métricas de desempenho: *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), *False Negatives* (FN), *sensitivity*, *specificity* e *accuracy*.

[2] `https://www.ilo.org/global/standards/subjects-covered-by-international-labour-standards/working-time/lang--en/index.htm`

Quanto aos resultados obtidos, Wahid et al [61] concluíram que o algoritmo *Tree Ensemble* fornece uma *accuracy* mais baixa, de 79%. Com uma avaliação acima da média para todas as métricas preditivas de absentismo de trabalhadores, o classificador *Gradient Boosted Tree* produziu a *accuracy* mais alta de 82%.

Asiri e Abdullah [3] para além de pretenderem prever o absentismo no trabalho, também tiveram intenção de descobrir as causas e os fatores que levavam à ausência dos funcionários recorrendo, também, a algoritmos de ML, através do mesmo conjunto de dados [36].

Relativamente à metodologia utilizada, Asiri e Abdullah [3] verificaram que certos atributos, como `'Travel expense'` e `'Distance'`, contêm valores num intervalo muito grande. Para que os algoritmos de ML não tenham que lidar com dados dessa magnitude, os autores decidiram agrupar os dados, agregando instâncias individuais para categorizar os dados contínuos em grupos de intervalos. Neste caso, com recurso ao histograma, projetaram as frequências das ausências de acordo com os seguintes atributos: `'Travel expense'`, `'Distance'`, `'Age'`, `'BMI'` e a variável *target*, `'Absenteeism time'`.

Após o agrupamento dos dados, Asiri e Abdullah [3] investigaram quais os atributos que afetam mais o absentismo, estudando a relação entre os atributos e a variável *target*, através da correlação. Os autores concluíram que as variáveis com forte correlação são: `'Distance'`, `'Age'`, `'Children'`, `'Social Drinker'` e `'Height'`. A partir destas *features*, constitui-se um novo conjunto de dados.

Por fim, utilizaram os dois conjuntos de dados para a fase de treino dos modelos, neste caso, o *dataset* com os atributos com forte correlação, ao qual se chamou *Reduced Dataset* e ao *dataset* que contém todos os atributos (excepto o `'ID'`) anteriormente processados, *Base Database*. Asiri e Abdullah [3] optaram por utilizar os seguintes algoritmos: *Naïve Bayes, Decision Tree* e o *Random Forest*, por terem uma melhor performance com dados categóricos. Para avaliar os modelos, os autores utilizaram as métricas *accuracy*, *sensitivity* e *specificity*. Os resultados mostraram que o *Random Forest* teve uma melhor performance, produzindo uma *accuracy* de 92% com a utilização do *Base Database*.

Num estudo bastante similar, Shah et al [1] resolveram o problema de previsão do absentismo no trabalho como um problema de classificação binária recorrendo, novamente, ao mesmo conjunto de dados [36], onde a variável *target*, `'Absenteeism time'`, foi convertida para uma variável binária: absentismo moderado (0-5 horas por mês) e absentismo excessivo (>5 horas).

Outro procedimento que Shah et al [1] empregaram, e que os restantes artigos anteriormente mencionados não efetuaram, foi a estandardização dos dados. Para que um modelo de ML não tenha que lidar com atributos com diferentes escalas/ unidades (instabilidade numérica), deve-se reescalar, neste caso estandardizar, os atributos para que fiquem todos na mesma unidade. Assim, torna-se o processo de treino mais rápido.

Tal como no trabalho Skorikov et al [51], a categorização realizada na variável *target*, resultou num desequilíbrio dos dados. Para tal, aplicaram a mesma técnica, SMOTE, para lidar com o *imbalanced data*.

No projeto de Skorikov et al [51], afirmaram que a aplicação de técnicas de DL não seria adequada, uma vez que o conjunto de dados é pequeno. No entanto, Shah et al [1] treinaram um modelo com *Deep Neural Network* e obtiveram uma *accuracy* melhor que Skorikov et al [51] , com 97.5%. Relativamente aos restantes modelos, *Decision Tree* obteve uma *accuracy* de 82.83%, o *Support Vector Machine* (SVM) obteve 84.32%, e o *Random Forest* 82.43%.

Ao contrário dos quatro últimos artigos descritos, Araujo et al [2] recorreram ao mesmo conjunto de dados [36], mas realizaram a análise via regressão. O objetivo do estudo deles é prever o número de horas de absentismo no trabalho.

Araujo et al [2] executaram uma *feature selection* nos atributos, aplicando o algoritmo de *relief*, de modo a verificar quais destes têm um impacto mais significativo no conjunto de dados. Os autores selecionaram os dez atributos com o maior peso: `'Reason for absence'`, `'Seasons'`, `'Month of absence'`, `'Workload / day'`, `'Age'`, `'Hit target'`, `'Social smoker'`, `'Education'`, `'BMI'` e `'Weight'`, formando, assim, um novo conjunto com dimensões relevantes.

Também, contrariamente aos trabalhos mencionados, Araujo et al [2] utilizaram para o seu estudo um modelo híbrido, neste caso, aplicaram um modelo *fuzzy neural network*, com o intuito de obter previsões mais exatas, e outros modelos mais tradicionais como a regressão linear, *multilayer perceptron*, SVM e o *naive Bayes*. Como se trata de um problema de regressão, a métrica utilizada para avaliar a performance do modelo foi o *Root Mean Squared Error* (RMSE), que será descrita mais em detalhe no próximo capítulo.

Araujo et al [2] treinaram os modelos com dois *datasets*, um com todos os atributos (excepto o `'ID'`) e outro com as dimensões relevantes resultantes do algoritmo *relief*. Os autores concluíram que o modelo *fuzzy neural network* obteve melhores resultados, com um RMSE de 12.91 no conjunto de teste. Araujo et al [2] consideraram, como futuras melhorias, identificar *outliers* e considerar outros aspetos sociais para o modelo, como o assédio e atitudes negativas no trabalho, que afetam o absentismo e, indiretamente, a parte financeira da empresa.

Ferreira et al [17] utilizaram um *dataset* proveniente de vários documentos sobre o trabalho em geral, contendo 38 atributos e 2233 instâncias, e foram extraídos entre janeiro de 2008 e dezembro de 2016, para prever o absentismo quantificado em dias. Neste caso, os autores realizaram um problema de regressão.

Primeiramente, os autores reduziram o conjunto de dados de 38 atributos para 17 atributos constituindo, assim, um novo conjunto com as seguintes *features*: motivo de absentismo - através da classificação internacional de doenças, e razões para absentismo não determinado pela classificação internacional de doenças (subsídio de acompanhamento médico, consulta médica ou dentária, exame físico e fisioterapia) - frequência de absentismo individual, dia da semana, mês, tempo de absentismo em dias e horas, tempo de serviço em ano, idade, número de crianças, entre outros. Ferreira et al [17] utilizaram como modelo uma *Artificial Neural Network* (ANN) e utilizaram como métrica de desempenho o RMSE para avaliar a performance. Na fase de treino, obtiveram um RMSE de 0.0176, e de

teste, 0.0198.

Uma das limitações que Ferreira et al [17] enfrentaram foi a utilização de um grande número de *features*; isso dificultou resolver o problema de previsão e aumentou o custo computacional.

Segundo Lima et al [31], nas instituições públicas, geralmente, as ausências conduzem a inúmeras perdas pessoais, sociais e económicas. Identificar os fatores de absentismo preponderantes pode ser benéfico para este tipo de instituição pois permite tomar ações preventivas. Como tal, Lima et al [31] investigaram potenciais classificadores de ML para identificar agentes de segurança que sejam mais propensos a ausentar-se por um maior período de tempo. Neste estudo, os autores pretendem prever, com algoritmos de aprendizagem supervisionada, se é provável que o trabalhador falte ao trabalho por mais de 30 dias no ano seguinte, através dos dados extraídos de uma instituição pública de segurança do Brasil.

A instituição forneceu duas bases de dados: o primeiro conjunto de dados é composto por atributos comuns dos agentes e o segundo composto por todos os registos de todos os agentes que faltaram por doença, num período de 6 anos (2012-2017). Lima et al [31] realizaram uma *feature selection* manualmente, selecionando *features* biográficas como, por exemplo, a idade e o seu género, bem como *features* relacionadas com o trabalho, tais como o número de anos de serviço de um agente, o nível de hierarquia na instituição e a soma de dias ausentes durante o ano, entre outros. Cada instância descreve a situação de ausência para um determinado indivíduo num ano específico, ou seja, o conjunto de dados contém 6 entradas por trabalhador (uma para cada ano observado). Para a variável *target*, Lima et al [31] utilizaram a soma dos dias ausentes durante o ano para a converter numa tarefa de classificação binária. Os autores consideram que um agente pratica absentismo de longa duração se a soma de ausências durante o ano for superior ou igual a 30 dias.

Após a fase de pré-processamento, Lima et al [31] formaram cinco conjuntos de dados distintos. Cada instância de um conjunto é composta por um único agente a partir de k anos de serviço consecutivos, o k varia de 2 a 6. Uma vez que cada agente representado em cada conjunto pode não conter dados no período de 6 anos, os conjuntos para k superiores são compostos por menos instâncias.

Para a fase de treino, Lima et al [31] utilizaram algoritmos como *Multilayer Perceptron*, *Recurrent Neural Networks* e *Long Short-Term Memory* (LSTM) e o SVM, e, para avaliar os modelos, recorreram à métrica *accuracy*.

Lima et al [31] concluíram que em comparação com o SVM, as redes neuronais produziram melhores resultados atingindo uma *accuracy* de 78%.

No trabalho de Boot et al [5], estes autores utilizaram uma abordagem consideravelmente diferente dos artigos supra citados. Para identificar os possíveis fatores que originam ausências prolongadas e frequentes por motivo de doença nas companhias aéreas, os autores efetuaram uma análise de regressão logística nos dados.

O conjunto de dados utilizado é proveniente de uma companhia aérea, e contém uma combinação de variáveis sociodemográficas e relacionadas com o trabalho entre 2005 e

2008.

A análise da regressão logística permitiu desenvolver dois modelos preditivos, um para prever as ausências a longo prazo (>42 dias consecutivos) e outro para prever ausências de doença frequente (pelo menos 4 ocorrências de ausência por doença)

Boot et al [5] concluíram que a ausência por doença de longa duração foi prevista por uma combinação de diferentes variáveis, tais como idade mais avançada, gravidez recente, agravamento das condições no trabalho e ausência por doença antecedente. Enquanto que a ausência frequente por doença foi prevista por outras variáveis, como estado civil, não ter filhos de 16 anos ou mais, não ter uma licença de estacionamento da empresa, não ter trabalho por turnos, ter um emprego com requisitos operacionais especiais e ausência de antecedente por doença. Os modelos de ausência prolongada e frequente por doença tiveram uma capacidade de discriminação de 0.72 e 0.73 (valores mais próximos de 1 são modelos com maior capacidade discriminativa), e uma variância explicada de 10.9% e 14.2%, respetivamente.

Por fim, Wong [62] propôs um trabalho cujo objetivo consistia em desenvolver um modelo capaz de prever com precisão o volume de tripulantes indisponíveis para as tarefas de voo de uma companhia aérea para diferentes horizontes temporais, utilizando intervalos diários, semanais e mensais, para que o modelo pudesse ser uma ferramenta útil para os planeadores e incorporar no sistema de gestão da tripulação.

Relativamente aos conjuntos de dados, estes foram obtidos pelo repositório da companhia aérea *Sabre Airline Solutions*. Através do pré-processamento dos dados, Wong [62] obteve um conjunto de resultados com atributos, como o número de tripulantes ativos, o ano, o mês, a semana do ano, o dia do ano, e o dia da semana. Como referido anteriormente, a variável *target* consiste no número total diário, semanal e mensal de pessoas não disponíveis para as tarefas de voo.

Wong [62], para além de ter utilizado modelos de ML e DL, também utilizou um método de decomposição de Série Temporal para prever. O autor conclui que o modelo de decomposição da Série Temporal obteve uma precisão de 95.12% para a previsão diária, seguido de LSTM com uma precisão de 93.82%.

Todos os estudos apresentados anteriormente recorreram principalmente a algoritmos de ML e DL para desenvolver modelos capazes de prever o absentismo em diversos domínios. Apesar dos algoritmos baseados em árvores terem sido abordagens muito utilizadas, as redes neuronais exibiram um desempenho melhor, exceto no último estudo exposto. Nesse trabalho, Wong [62], para além de ter aplicado os algoritmos mencionados, também utilizou um método de previsão clássico, nomeadamente um método de decomposição, e focou-se em prever o número de tripulantes indisponíveis como um vetor de observações no tempo, isto é, uma série temporal. Em contraste, os primeiros trabalhos focaram-se em resolver o problema do absentismo como uma tarefa de classificação.

Como na literatura não existem muitos trabalhos que tratam os dados como séries

temporais, a presente Dissertação apresenta uma oportunidade única de explorar o absentismo sob forma de série temporal, e utilizar métodos de previsão de séries temporais clássicos e avançados, dependendo da complexidade dos dados, para prever o número de tripulantes ausentes no setor ferroviário.

3

Metodologia

Neste capítulo são introduzidos os fundamentos teóricos sobre a análise de séries temporais, entre eles os conceitos estruturantes e o contexto, que serão aplicados no desenvolvimento deste estudo.

Posteriormente, são apresentados alguns dos modelos de previsão presentes na literatura para o cumprimento dos objetivos do trabalho.

3.1 Conceitos fundamentais

3.1.1 Métodos de previsão

A **previsão** é fundamental para efeitos de **planeamento** em diversas áreas tais como: gestão de empresas, gestão da produção, *marketing*, entre outros, uma vez que oferece a capacidade de tomar decisões informadas e desenvolver estratégias através dos dados disponíveis. No contexto de análise de séries temporais, **prever**, significa estimar como a sequência de observações do passado irá evoluir no futuro [37].

Num mercado progressivamente mais competitivo e incerto, as empresas têm necessidade cada vez mais de desenvolver e aplicar **métodos de previsão**, para que se possa reduzir essa incerteza e os riscos associados à tomada de decisão. É neste sentido que se centraliza o tema deste estudo, para que seja possível efetuar um melhor planeamento no que toca à capacidade disponível de tripulantes face às necessidades operacionais.

Um **método de previsão** é um procedimento que permite calcular previsões. Os métodos podem ser categorizados em dois grupos:

- **Qualitativos**: baseiam-se em juízos subjetivos e na opinião intuitiva e experiente dos especialistas, e no estabelecimento de cenários e/ou paralelismos com situações semelhantes [37].

- **Quantitativos**: baseiam-se na manipulação matemática de dados históricos (quantificados), e têm como objetivo projetar no futuro padrões de comportamento que se identificam nos dados do passado [37].

A utilização de métodos qualitativos pode ser justificada quando não se tem acesso a dados históricos ou quando existem alterações significativas no passado e as condições não se mantêm no futuro o que invalida a hipótese de estabilidade dos padrões de comportamentos passados. A hipótese de estabilidade pressupõe que prevalecerão no futuro os padrões de comportamento identificados na série de valores passados [37]. Em contraste, os métodos quantitativos podem ser utilizados quando se dispõe de dados históricos e devem assentar no pressuposto de estabilidade desses padrões de comportamento passado [32].

Dentro dos **métodos quantitativos** existem:

- **Modelos causais**: procuram compreender como as variáveis explicativas influenciam a variável dependente, ou seja, a variável sobre a qual se pretende efetuar previsões de uma série são explicados não somente pelos valores passados da mesma, mas também por outras variáveis que de alguma forma possuam relação com ela. Através da análise de regressão é possível modelar e analisar dados que são compostos por uma variável resposta e uma ou mais variáveis independentes [37].
- **Modelos não causais**: assentam apenas na **análise da série** de valores passados, procurando caracterizar a sua forma de evolução e projetar no futuro esse padrão de comportamento [37].

A **análise de séries temporais** é um procedimento sistemático com objetivos bem definidos, sendo os principais:

- **Descrição dos dados**: permite determinar as propriedades de uma série, como por exemplo, identificar padrões a longo prazo e a existência de alterações súbitas e não sistemáticas na série. É fundamental ter este conhecimento prévio, pois é importante no processo de modelação, no que toca à escolha do modelo mais adequado que satisfaz as propriedades da série [37].
- **Modelação**: ajustar um modelo (univariado ou multivariado) aos dados que consiga explicar a variação da série ao longo do tempo [37].
- **Previsão**: pretende estimar os valores futuros da série, tendo em conta os valores do passado [37].

3.1.2 Processo estocástico e série temporal

Um processo estocástico define-se como uma coleção ordenada de variáveis aleatórias indexadas no tempo, representado por

$$\{X_t : t \epsilon T\}$$

onde,

T representa um conjunto ordenado de índices.

Quando o conjunto dos parâmetros T é um conjunto numerável, tipicamente $\{X_t : t = 1, 2, 3, ..., n\}$, diz-se que o processo tem parâmetro ou tempo discreto [37], caso contrário, diz-se que tem tempo contínuo $\{X_t : t > 0\}$.

Uma série temporal estocástica (também designada por sucessão cronológica) define-se como sendo uma sucessão de observações ordenadas no tempo

$$x_1, x_2, x_3, x_4, ..., x_t, ... \tag{3.1}$$

sendo x_t o valor observado no instante temporal t [37].

Em termos gerais, os **processos estocásticos estacionários** caracterizam-se por um equilíbrio estatístico em torno de um nível médio fixo, isto é, têm propriedades probabilísticas que são estáveis ou invariantes ao longo do tempo [37]. Uma série temporal estocástica é estacionária quando o valor médio e a variância são constantes ao longo do tempo.

O processo estocástico diz-se **fortemente estacionário** (ou estacionário em **sentido estrito**) quando para todo o n e todo $t_1, ..., t_n$, a distribuição conjunta de variáveis aleatórias $X_{t_1}, X_{t_2}, ..., X_{t_n}$ é idêntica à distribuição conjunta de $X_{t_1+k}, X_{t_2+k}, ..., X_{t_n+k}$, qualquer que seja $k \epsilon \mathbb{N}$ [42]

$$\begin{aligned} F(X_{t_1}) &= F(X_{t_1+k}), \\ F(X_{t_1}, X_{t_2}) &= F(X_{t_1+k}, X_{t_2+k}), \\ &... \end{aligned} \tag{3.2}$$

onde,

k representa o desfasamento temporal; e

$F(\cdot)$ representa a função de distribuição de X.

O processo estocástico diz-se **fracamente estacionário** (ou estacionário em **sentido lato**) quando o valor médio e a variância são constantes ao longo do tempo e a covariância depende apenas do número de *lags* k[37]

$$E(X_t) = E(X_{t+k}) = \mu \tag{3.3}$$

$$Var(X_t) = Var(X_{t+k}) = \sigma^2 \tag{3.4}$$

$$Cov(X_{t_1},\ X_{t_1+k}) = Cov(X_{t_2},\ X_{t_2+k}) = \gamma_k. \tag{3.5}$$

Uma sucessão cronológica diz-se estacionária se o processo estocástico que gera a sucessão for estacionário. Caso contrário, diz-se **não estacionária** [37].

Em **processos estocásticos não estacionários** a dependência no tempo pode dificultar a realização da inferência dos dados fora da amostra e pode gerar piores previsões que as séries estacionárias [63]. Assim, tem-se:

- **Série temporal não estacionária em média** - quando o nível da série não é estável no tempo, podendo ter tendência crescente ou decrescente [63], isto é,

$$E(X_t) \neq \mu. \tag{3.6}$$

- **Série temporal não estacionária nas covariâncias** – quando a variância ou as covariâncias variam com o tempo [63], isto é,

$$Var(X_t) \neq \sigma^2 \tag{3.7}$$

$$Cov(X_{t_1},\ X_{t_1-k}) \neq Cov(X_{t_1},\ X_{t_1+k}). \tag{3.8}$$

Um processo estocástico $\{\epsilon_t\}$ é chamado de **ruído branco** quando

$$E(\varepsilon_t) = \mu_\varepsilon \tag{3.9}$$

$$Var(\varepsilon_t) = \sigma_\varepsilon^2 \tag{3.10}$$

$$\gamma_k = Cov(\varepsilon_t,\ \epsilon_{t+k}) = 0, \text{para todo } k \neq 0. \tag{3.11}$$

Séries com **tendência** e/ou **sazonalidade** não são estacionárias, pois, em geral, uma série estacionária não tem padrões previsíveis a longo prazo [37].

3.1.3 Caraterização da série temporal

A série temporal possui determinadas propriedades que a caraterizam. Estas propriedades podem ser identificadas através da representação gráfica dos valores da sucessão que se pretende estudar, em função do tempo. Este tipo de gráfico designa-se por **cronograma** e permite identificar alguns padrões que traduzem comportamentos ou **componentes** da sucessão.

Uma série pode decompor-se nas seguintes componentes básicas:

- **Tendência** (T_t): reflete uma evolução global no sentido crescente ou decrescente [37].

- **Sazonalidade** (S_t): representa a flutuação com periodicidade fixa. Os períodos de variação são ciclos sazonais por S períodos de tempo que podem estar associados ou não às estações do ano [37].

- **Ciclicidade** (C_t): reflete movimentos oscilatórios de médio prazo. São movimentos sem periodicidade fixa que só são detetáveis para séries longas. A dificuldade de separar os ciclos longos da tendência, é a razão porque alguns métodos tratam em conjunto ciclo e tendência [37].

- **Componente aleatória/ residual** (R_t): reflete a influência que o acaso exerce sobre o comportamento da variável ao longo do tempo [37].

As séries temporais podem exibir uma variedade de padrões[1] e, portanto, pode-se dividir a série nas componentes descritas acima. Quando se **decompõe** a série em componentes, geralmente, combina-se a tendência T_t e a ciclicidade C_t numa única componente de **ciclo-tendência** (para simplificar denomina-se apenas por **tendência**), pois a componente cíclica tem um comportamento de médio longo prazo e, consequentemente, a sua identificação apenas é possível quando se dispõe de séries longas [37]. Assim, pode-se representar matematicamente a série como combinação dessas componentes na seguinte forma:

$$X_t = f(T_t,\, S_t,\, R_t) \tag{3.12}$$

onde,

X_t representa a sucessão cronológica univariada;

T_t representa a componente de tendência no instante t;

S_t representa a componente de sazonalidade no instante t; e

R_t representa a componente residual no instante t.

Para algumas séries temporais, pode haver mais de uma componente sazonal, correspondendo aos diferentes períodos sazonais.

3.1.4 Transformações e ajustes

Geralmente, ajustar os dados históricos pode tornar a tarefa de previsão mais simples, por exemplo, ajustes no calendário e transformações matemáticas. Estes ajustes e transformações servem para simplificar os padrões dos dados históricos removendo fontes de variação conhecidas ou simplesmente tornando o padrão mais consistente ao longo do conjunto de dados [24]:

1. Ajustes da população:

Quaisquer dados que sejam afetados por alterações populacionais podem ser ajustados para fornecer dados *per capita*, isto é, considerar os dados por pessoa (ou por mil pessoas, ou por milhão de pessoas) em vez do total. Por exemplo, considerando um estudo sobre o número de camas de hospital numa determinada região ao longo do tempo, os resultados são mais fáceis de interpretar se eliminarmos os efeitos das alterações populacionais, considerando o número de camas por mil pessoas. Assim, é possível verificar se foram registados aumentos reais no número de camas ou se os aumentos se devem inteiramente ao aumento da população [24].

2. Transformações matemáticas:

[1] Apesar das séries temporais apresentarem sempre a componente aleatória, nem todas apresentam as três primeiras componentes mencionadas, pois um determinado padrão comportamental pode não ser explicado pelos restantes comportamentos [42].

Quando uma sucessão não tem variância constante ao longo do tempo, deve-se efetuar uma transformação por forma a tornar a variância constante. Estas transformações devem ser feitas antes da aplicação de um método de previsão [37].

Por exemplo, deve-se utilizar uma **transformação logarítmica** quando a variância da série aumenta com o nível da série.

Considerando as observações originais de uma série

$$x_1, ..., x_T \tag{3.13}$$

Então

$$w_t = log(x_t), \tag{3.14}$$

representa as observações transformadas.

Esta transformação é particularmente útil, pois as diferenças na série dos logaritmos representam as variações relativas da série original.

A transformação apresentada é apenas aplicada para séries de valores positivos. Esta limitação pode ser mitigada adicionando uma constante à série – o que não afeta a sua estrutura de correlação [37]. Portanto, outra particularidade da transformação logarítmica é restringir as previsões de forma a que estas permaneçam positivas na escala original [24].

Quando se transformam os dados e se aplica um método de previsão, obtêm-se previsões para os valores transformados em vez de previsões para os dados originais. Para se obter as previsões para os dados originais, é necessário reverter a transformação nas previsões, sendo o inverso da transformação logarítmica dado por

$$x_t = e^{w_t}. \tag{3.15}$$

No entanto, esta reversão pode introduzir algum enviesamento nas previsões [37].

Para além da transformação logarítmica, existem outras transformações que podem ser úteis, por exemplo, raízes quadradas e raízes cúbicas. Estas transformações denominam-se por **transformações de potência** e podem ser escritas da seguinte forma genérica

$$w_t = x_t^p, \ p \in \mathbb{N} \tag{3.16}$$

A **transformação *Box Cox*** pode representar ambas as transformações logarítmica e de potência, dependendo do parâmetro λ e pode ser expressa como

$$w_t = \begin{cases} log(x_t) & \text{se } \lambda = 0; \\ \frac{x_t^\lambda - 1}{\lambda} & \text{caso contrário.} \end{cases} \tag{3.17}$$

para valores de x_t positivos [24].

Para reverter esta transformação:

$$x_t = \begin{cases} exp(w_t) & \text{se } \lambda = 0; \\ (\lambda w_t + 1)^{1/\lambda} & \text{caso contrário.} \end{cases} \tag{3.18}$$

3.1.5 Decomposição das séries temporais

Como se referiu no subcapítulo 3.1.3, as séries temporais podem exibir uma variedade de padrões que se traduzem em componentes. Por vezes, é útil isolar cada uma dessas componentes, de modo a facilitar a identificação das características que compõem a série, e definir o método de previsão adequado com base nessas caraterísticas, mas também podem ser usadas para melhorar a precisão da previsão. As componentes das séries podem ser decompostas de forma **aditiva** ou **multiplicativa**:

1. Modelo Aditivo:

Descrito pela expressão:

$$X_t = T_t \ + \ S_t \ + \ R_t \tag{3.19}$$

Aplica-se em situações em que a magnitude da componente sazonal não varia com o nível da série[2][37].

2. Modelo Multiplicativo:

Descrito pela expressão:

$$X_t = T_t \ \times \ S_t \ \times \ R_t \tag{3.20}$$

É adequado quando as flutuações sazonais aumentam ou diminuem proporcionalmente ao nível da série [37]. Uma alternativa ao uso da decomposição multiplicativa é primeiro transformar os dados até que a variação da série se torne estável ao longo do tempo e, de seguida, usar uma decomposição aditiva. Quando uma transformação logarítmica é aplicada, isso é equivalente a usar uma decomposição aditiva nos dados originais, isto é, a expressão (3.20) equivale a

$$log(X_t) = log(T_t) \ + \ log(S_t) \ + \ log(R_t) \tag{3.21}$$

[2]O nível da série são os valores que a série tipicamente pode assumir, se não for observado comportamento crescente ou decrescente a longo prazo

3.1.5.1 Decomposição STL e decomposição MSTL

O *Seasonal and Trend decomposition using Loess* (STL), desenvolvido por Cleveland et al [7], é um método versátil e robusto para decomposição de séries temporais, sendo *Loess* um método não paramétrico para estimação de relações não lineares [24]. Tradicionalmente, este método decompõe na forma aditiva (Equação 3.19), porém é possível efetuar uma decomposição na sua forma multiplicativa (Equação 3.20), aplicando uma transformação logarítmica na série original e, de seguida, realizar a transformação inversa das componentes obtidas [25].

A utilização deste método tem múltiplas vantagens [25]:

- O método STL consegue lidar com qualquer tipo de sazonalidade, não apenas dados mensais e trimestrais;
- A componente sazonal pode variar ao longo do tempo, e a sua taxa de mudança pode ser controlada pelo utilizador;
- A suavidade da componente tendência também pode ser controlada pelo utilizador;
- Pode ser robusto para *outliers*, de modo que as observações ocasionais incomuns não afetem as estimativas da componente tendência e componentes sazonais. No entanto, os mesmos irão afetar a componente residual.

Geralmente, séries temporais de frequência[3] mais alta podem exibir padrões sazonais mais complexos, por exemplo, dados diários podem ter um padrão semanal, bem como um padrão anual [24]. Neste contexto, a decomposição STL não é capaz de extrair as duas componentes sazonais. Assim, uma nova abordagem para a decomposição de séries temporais com sazonalidade múltipla foi introduzida por [4].

O método *Multiple Seasonal-Trend decomposition using LOESS* (MSTL) estende a equação 3.19 para incluir várias componentes sazonais

$$X_t = T_t + S_t^1 + S_t^2 + \ldots + S_t^n + R_t \tag{3.22}$$

onde,

n representa o número de ciclos sazonais presentes em X_t.

Se a série temporal for sazonal, o MSTL aplica o algoritmo STL iterativamente a cada uma das frequências sazonais identificadas e, de seguida, a componente da tendência é calculada utilizando a última iteração do STL. Caso não seja sazonal, o MSTL determina apenas as componentes tendência e residual da série [4].

[3] A frequência de uma série temporal é o número de observações decorridas até que o padrão sazonal se repita. Isso é o inverso da definição de frequência na física (corresponde, na verdade, à definição de período na física), ou na análise de *Fourier* [24].

Os métodos de decomposição devem, em geral, ser utilizados como uma ferramenta de análise exploratória dos dados, permitindo identificar características da série que são fundamentais para a definição de um método de previsão, pois nem sempre funcionam bem em termos de previsão [37].

3.1.6 Função de autocorrelação

A função de autocorrelação (ou *Autocorrelation Function* (ACF), em inglês) é um indicador estatístico que mede a relação linear entre as observações de uma série temporal desfasadas por k períodos (X_t e X_{t-k}), por outras palavras, mede o quão fortemente o valor observado no instante t se encontra correlacionado com os valores observados no passado [42]. Para uma série temporal estacionária, a ACF é definida da seguinte forma

$$\rho(X_t,\ X_{t-k}) = \rho_k = \frac{\gamma_k}{\mathrm{Var}(X_t)},\ \text{para}\ k\ =\ 1,\ \ldots,\ n-1\ \text{com}\ -1 \leq \rho_k \leq 1 \tag{3.23}$$

onde,

γ_k representa a covariância entre X_t e X_{t+k}; e

$\mathrm{Var}(X_t)$ representa a variância de X_t.

Em geral, à medida que k aumenta é de esperar que a capacidade de memória do processo diminua (decrescimento de ρ_k). A forma como ρ_k decai, carateriza a memória do processo [37].

Pode-se estimar γ_k e ρ_k, a partir das observações $x_1,\ x_2,\ \ldots,\ x_t$, através de

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=k+1}^{T}(X_t - \overline{X})(X_{t-k} - \overline{X})}{\sum_{t=1}^{T}(X_t - \overline{X})^2} \tag{3.24}$$

$$\hat{\gamma}_k = \frac{1}{T-1}\sum_{t=k+1}^{T}(X_t - \overline{X})(X_{t-k} - \overline{X}) \tag{3.25}$$

O gráfico das estimativas da ACF para diferentes valores de k, é habitualmente designado por **correlograma** [37].

3.1.7 Análise do correlograma

Uma das aplicações da análise do correlograma é ajudar a averiguar se a série é estacionária ou não, identificando a presença ou não de tendência e/ou sazonalidade.

Observando as Figuras 3.1, 3.2, 3.3 e 3.4, em geral, numa série estacionária, a ACF tende abruptamente para zero após alguns (poucos) *lags*. O decaimento da ACF pode ser exponencial ou sob a forma de uma sinusoidal amortecida [37] - Figura 3.1. Numa série não estacionária, o decaimento para zero da ACF (quando k aumenta) é lento (linear), sendo as autocorrelações elevadas durante muitos *lags* [37] - Figura 3.2. Quando a série possui

sazonalidade, os coeficientes de autocorrelação são maiores a cada intervalo de tempo correspondente ao período sazonal [42] - Figura 3.3. Por fim, quando a série apresenta os dois comportamentos (sazonalidade e tendência), os valores dos *lags* só decrescem ao fim de muitas observações [42] e certos coeficientes de autocorrelação são maiores a cada intervalo de tempo correspondente ao período sazonal - Figura 3.4.

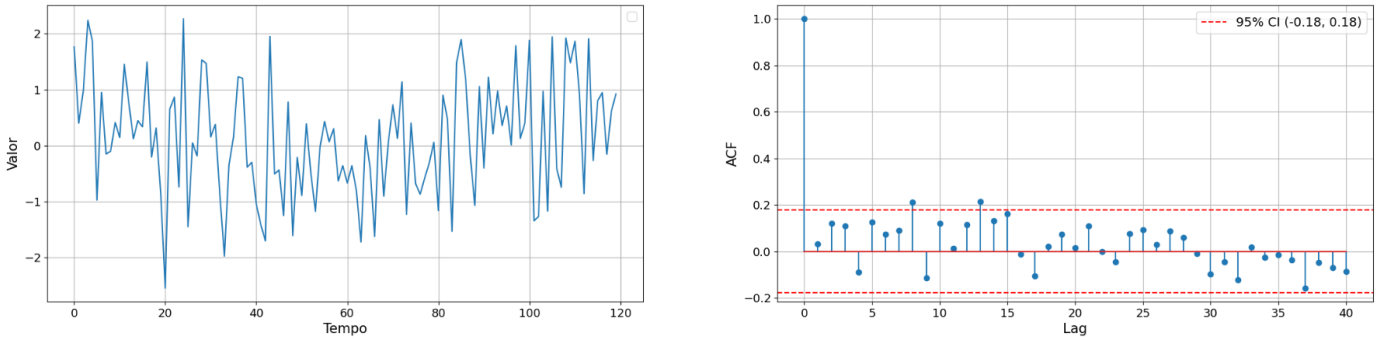

Figura 3.1: Série temporal estacionária, do lado esquerdo, e a sua ACF, do lado direito.

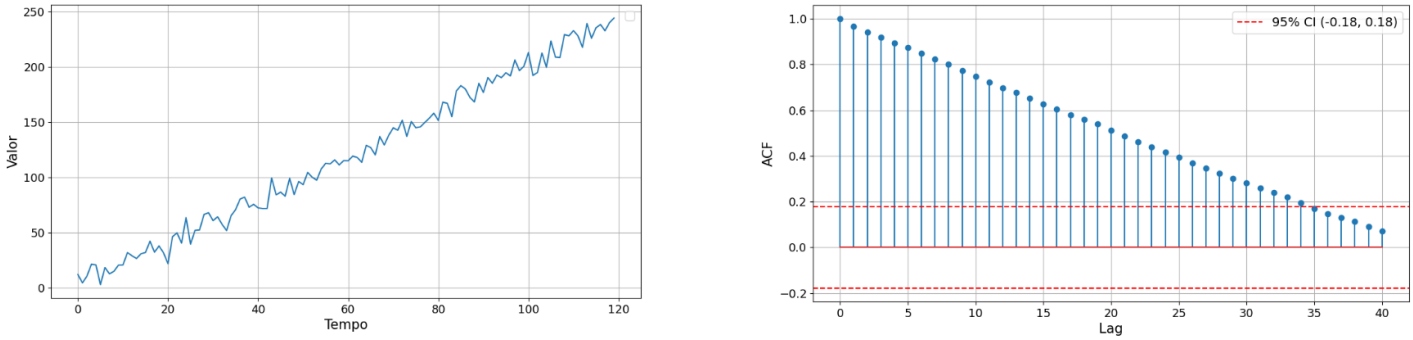

Figura 3.2: Série temporal não estacionária, do lado esquerdo, e a sua ACF, do lado direito.

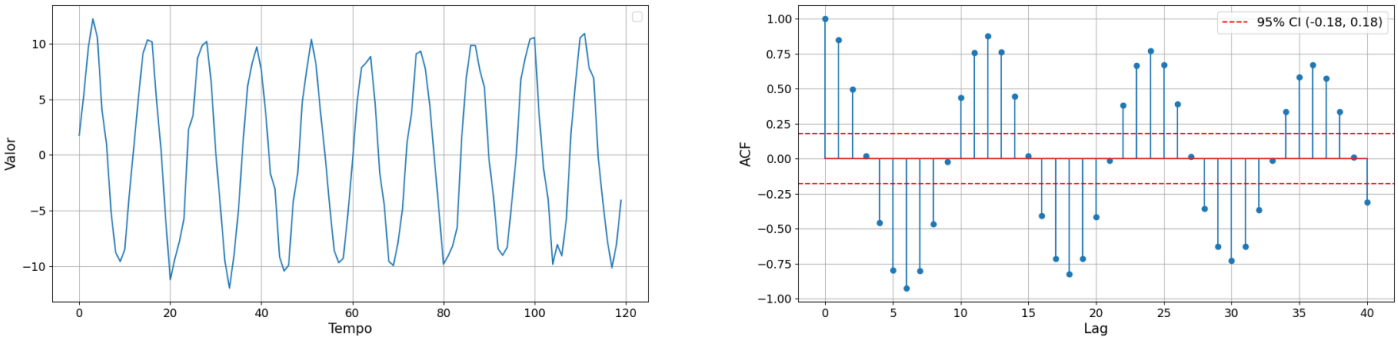

Figura 3.3: Série temporal com sazonalidade, do lado esquerdo, e a sua ACF, do lado direito.

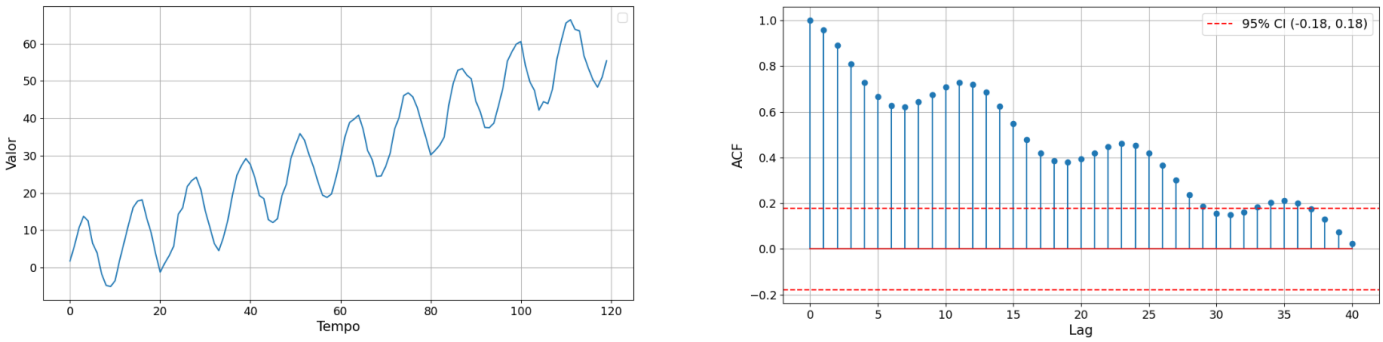

Figura 3.4: Série temporal com tendência e sazonalidade, do lado esquerdo, e a sua ACF, do lado direito.

3.1.8 Função autocorrelação parcial

A função de autocorrelação parcial (ou *Partial Autocorrelation Function* (PACF)) é semelhante à ACF) , diferindo deste último pelo facto de medir a correlação entre as observações de uma série temporal desfasados por k períodos, contabilizando os valores dentro desse intervalo. Por exemplo, se a primeira observação estiver relacionada com o segundo elemento da série, o segundo com o terceiro, então o primeiro poderá estar relacionado com o terceiro elemento da série [42]. O coeficiente de autocorrelação parcial é dado por

$$\rho_{k,k} = corr(X_t,\ X_{t-k}|X_{t-1},\ \ldots,\ X_{t-k+1}). \tag{3.26}$$

3.1.9 Teste de estacionaridade

Como referido anteriormente, a análise do cronograma e do correlograma poderão dar uma ideia dos padrões presentes nos dados, no entanto, existem testes que ajudam a averiguar a estacionaridade dos mesmos.

Para perceber se uma série temporal é estacionária ou não, pode-se recorrer a vários testes para fazer este estudo, nomeadamente:

- Teste *Augmented Dickey-Fuller* (ADF);
- Teste *Kwiatkowski–Phillips–Schmidt–Shin* (KPSS).

Estes pertencem a uma categoria de testes denominados por **teste da raiz unitária**. Um teste de raiz unitária testa se uma série temporal possui uma raiz unitária, o que implica que a série não é estacionária. A hipótese nula, geralmente, é definida como a presença de uma raiz unitária e a hipótese alternativa é estacionaridade, estacionariedade em tendência, dependendo do teste usado [11].

De uma forma geral, o teste de raiz unitária pode ser representado matematicamente da seguinte maneira

$$X_t = D_t + z_t + \varepsilon_t \tag{3.27}$$

onde,

D_t representa componente determinística (tendência, sazonalidade, entre outros);
z_t representa a componente estocástica; e
ε_t representa o processo de erro estacionário.

A tarefa do teste é determinar se a componente estocástica (z_t) contém uma raiz unitária ou é estacionária [11].

1. Teste ADF:

O teste *Dickley-Fuller* (DF) [14] é um teste de raiz unitária muito conhecido e amplamente utilizado. Considerando um modelo auto-regressivo de primeira ordem (AR(1))[4]

$$X_t = \varphi_1 X_{t-1} + \varepsilon_t \tag{3.28}$$

onde,

φ_1 é um parâmetro de auto-regressão e representa uma constante fixa desconhecida a estimar; e
$\varepsilon_t \overset{iid}{\sim} N(0, \sigma_t)$ representa o ruído branco.

As hipóteses a testar são:

$H_0 : \varphi_1 = 1$ − A série temporal tem raiz unitária (a série não é estacionária)
$H_1 : |\varphi_1| < 1$ − A série temporal não tem raiz unitária (a série é estacionária)

O modelo de regressão pode ser reescrito da seguinte maneira:

$$\Delta X_t = (\varphi_1 - 1)X_{t-1} + \varepsilon_t = \delta X_{t-1} + \varepsilon_t \tag{3.29}$$

com

$$\Delta X_t = X_t - X_{t-1} \tag{3.30}$$

$$\delta = \varphi_1 - 1 \tag{3.31}$$

Neste modelo, a série tem raiz unitária se $\delta = 0$, o que corresponde à hipótese nula, caso contrário, a série não tem raiz unitária se $\delta < 0$.

Nos modelos de DF, o termo do erro segue um processo de ruído branco. Caso o termo do erro seja correlacionado, a estatística de teste pode rejeitar a hipótese nula. Para eliminar a correlação no termo do erro, Said e Dickey [46] sugeriram adicionar um número maior de *lags* das diferenças da série X_t à regressão de teste [41]. Esta versão do teste

[4]Este modelo será definido mais em detalhe no capítulo 3.5.1

é conhecida na literatura como teste *Augmented Dickey-Fuller* (ADF), e faz o estudo da estacionaridade com base na seguinte regressão:

$$\Delta X_t = \beta_0 + \beta_1 + \delta X_{t-1} + \sum_{i=1}^{p-1} \alpha_i \Delta X_{t-i} + \varepsilon_t \tag{3.32}$$

onde,

β_0 representa a constante do modelo, também conhecida como *drift*;
β_1 representa o coeficiente de tendência;
δ representa o coeficiente de presença de raiz unitária; e
p representa o número de *lags* tomados na série.

Este teste também considera as hipóteses descritas anteriormente.

A estatística de teste é dada por:

$$T = \frac{\hat{\delta}}{sd(\hat{\delta})} \tag{3.33}$$

onde,

$\hat{\delta}$ representa um estimador para δ; e
$sd(\hat{\delta})$ representa um estimador para o desvio padrão do erro de δ.

Os valores críticos da estatística T são calculados por Dickley e Fuller [14] através de simulação de Monte Carlo [63].

2. Teste KPSS:

Ao contrário do teste anterior, o teste KPSS [30] considera as seguintes hipóteses:

$$H_0: \text{ A série estacionária em tendência.}$$
$$H_1: \text{ A série tem raiz unitária (a série não é estacionária).}$$

Este teste assenta na ideia de que a série temporal é estacionária em torno de uma tendência determinística, e pressupõe que a série pode ser decomposta da seguinte forma:

$$X_t = d_t + r_t + \varepsilon_t, \text{ sendo } r_t = r_{t-1} + \mu_t \tag{3.34}$$

onde,

d_t representa a componente de tendência determinista;
r_t representa a componente do passeio aleatório[5];
μ_t representa as flutuações do passeio aleatório; e

[5]O passeio aleatório é um caso especial dos modelos ARIMA.

ε_t representa o erro estacionário.

Considerando o modelo (3.34), os resíduos[6] são dados por $e_t = x_t - \hat{x}_t$. Define-se a soma parcial dos erros como

$$S_t = \sum_{i=1}^{t} e_t \tag{3.35}$$

Então, a estatística de teste é dada por:

$$LM = \sum_{t=1}^{N} \frac{S_t^2}{N^2 \hat{\sigma}_\varepsilon^2} \tag{3.36}$$

onde,

$\hat{\sigma}_\varepsilon^2$ representa o estimador para a variância dos erros da regressão.

3.2 Análise dos resíduos

A análise dos resíduos é útil para verificar se um modelo conseguiu explicar adequadamente a informação dos dados [37]. No entanto, nem todos os modelos de séries temporais requerem esta análise como parte essencial do processo de modelação. Portanto, a necessidade de analisar os resíduos depende do modelo e dos seus pressupostos.

Os resíduos são erros após o ajustamento de um modelo. Geralmente, em modelos de séries temporais, os resíduos são iguais à diferença entre as observações e os valores ajustados correspondentes

$$e_t = x_t - \hat{x}_t \tag{3.37}$$

Um bom método de previsão produzirá resíduos com as seguintes propriedades [24]:

- Os resíduos não estão correlacionados. Se existirem correlações entre os resíduos, então existe informação nos resíduos que deve ser utilizada para calcular as previsões.
- Os resíduos têm média zero. Se os resíduos tiverem uma média diferente de zero, então as previsões são enviesadas.

Qualquer método de previsão que não satisfaça estas propriedades pode ser melhorado. Porém, isto não significa que os métodos de previsão que satisfazem estas propriedades não possam ser melhorados. A verificação destas propriedades é importante para verificar se um método está a utilizar toda a informação disponível, mas não é a melhor forma de selecionar um método de previsão [24].

[6] Erros após o ajustamento.

Para além destas propriedades, é útil (mas não necessário) que os resíduos também tenham as seguintes propriedades:

- Variância constante (homoscedasticidade).
- Distribuição normal.

Estas propriedades facilitam o cálculo dos intervalos de previsão [24].

Em problemas reais, é importante reconhecer que os pressupostos muitas das vezes não são cumpridos.

Para além dos resíduos definidos em (3.37), é frequente calcular os **resíduos estandardizados** que são dados por

$$\frac{e_t}{SE(e_t)} \tag{3.38}$$

onde,
$SE(e_t)$ representa o erro padrão dos resíduos.

Os resíduos estandardizados permitem identificar *outliers*, e podem ser usados para comparar os resíduos entre diferentes modelos.

3.3 Avaliação das previsões

Para poder avaliar o desempenho de um modelo ou método isoladamente é necessário separar os dados disponíveis em duas subamostras: **amostra de treino** e **amostra de teste**. A amostra de treino, dada por $\{x_1, ..., x_T\}$, é utilizada para estimar os parâmetros do modelo e avaliar o seu ajuste, e a amostra de teste, dada por $\{x_{T+1}, x_{T+2}, ..., x_{T+h}\}$, é usada para avaliar a precisão das previsões, isto é, avaliar o modelo perante novas observações nunca antes vistas pelo modelo [25].

Uma vez que os dados de teste não são utilizados para determinar as previsões, os mesmos fornecem uma indicação fiável relativamente à capacidade de generalização do modelo [25].

A dimensão da amostra de teste depende de diversos fatores, como por exemplo, o horizonte temporal de previsão, método a utilizar, entre outros. Um valor típico é a dimensão da amostra corresponder a 20% da dimensão total da amostra [25].

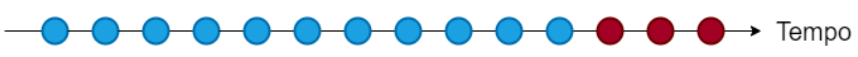

Figura 3.5: Cada ponto corresponde a uma observação. As observações a azul correspondem à amostra de treino e as observações a vermelho formam a amostra de teste (adaptado de [24]).

De ora em diante, este método que se encontra ilustrado na Figura 3.5 será apelidado de **janela fixa**.

O objetivo é minimizar os erros de previsão. O erro de previsão é a diferença entre um valor observado e sua previsão, e pode ser escrito como

$$e_{T+h} = x_{T+h} - \hat{x}_{T+h|T} \tag{3.39}$$

sendo que $\hat{x}_{T+h|T}$ corresponde às previsões de x_{T+h} tendo em conta $x_1, ..., x_T$ (isto é, h-passos que tem em conta todas as observações até ao momento T).

Pode-se medir a precisão da previsão resumindo os erros de previsão das seguintes maneiras [25].

3.3.1 Indicadores dependentes da escala dos dados

Os erros de previsão encontram-se na mesma escala que os dados. A principal vantagem das medidas dependentes da escala é que, geralmente, são fáceis de calcular e de interpretar, no entanto, não podem ser usadas para fazer comparações entre séries que envolvem unidades diferentes. Segue-se as medidas dependentes da escala mais utilizadas:

1. Média dos Erros Absolutos:

A Média dos Erros Absolutos (ou *Mean Absolute Error* (MAE), em inglês) representa a diferença média absoluta entre os valores reais e os valores previstos

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|e_t| \tag{3.40}$$

Quanto menor for o valor do MAE, maior é a precisão do modelo.

É fácil de calcular e útil para avaliar a precisão de uma série temporal, no entanto, não é adequado para comparar com outras séries que tenham unidades diferentes. Além disso, não deve ser utilizado se o objetivo não for penalizar os *outliers* [25].

2. Raiz Quadrada do Erro Quadrático Médio:

A Raiz Quadrada do Erro Quadrático Médio (ou *Root Mean Squared Error* (RMSE), em inglês) representa a raiz quadrada da diferença média quadrática entre os valores reais e os valores previstos

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}e_t^2} \tag{3.41}$$

Quanto menor for o valor do RMSE, maior é a precisão do modelo.

Tem as mesmas vantagens que o MAE, mas é mais sensível aos *outliers* e é mais difícil de interpretar que o primeiro indicador descrito [25].

3.3.2 Erro percentual

O erro percentual é dado por

$$p_t = \frac{e_t}{x_t} 100 \tag{3.42}$$

Uma das suas vantagens é o facto de não depender da escala dos dados, e pode ser usado para comparar o desempenho da previsão entre diferentes séries temporais.

No entanto, as medidas baseadas no erro percentual têm a desvantagem de ser indefinidas ou infinitas se a série temporal contiver zeros, isto é, $x_t = 0$, para qualquer t dentro do período de interesse, e gerar valores extremos, se qualquer x_t for próximo de zero. Outra desvantagem, estas medidas assumem que a unidade de medida tem um zero absoluto, ou seja, são válidas numa escala de rácios, mas não numa escala de intervalos. Por exemplo, não faz sentido medir a precisão das previsões da temperatura nas escalas *Fahrenheit* ou *Celsius* através do erro percentual, pois a temperatura tem um ponto zero arbitrário [25].

1. Erro Absoluto Percentual Médio :

O Erro Absoluto Percentual Médio (ou *Mean Absolute Percentage Error* (MAPE), em inglês) representa a média da diferença absoluta entre os valores reais e previstos dividido pelos valores reais

$$\text{MAPE} = \frac{1}{T} \sum_{t=1}^{T} |p_t| \tag{3.43}$$

Tal como as medidas anteriores, quanto menor for o MAPE, melhor a precisão do modelo.

Apesar de ser uma medida fácil de interpretar e ter as vantagens anteriormente mencionadas, tem a desvantagem de ser uma medida assimétrica, no sentido em que impõe uma penalidade mais pesada nos erros negativos (quando as previsões possuem valores mais altos que os valores reais) do que nos erros positivos [25]. Para evitar a assimetria do MAPE uma nova métrica de erro foi proposta.

O Erro Absoluto Percentual Médio Simétrico (ou *Symmetric Mean Absolute Percentage Error* (SMAPE), em inglês) é definida da seguinte maneira

$$\text{SMAPE} = \frac{200}{T} \sum_{t=1}^{T} \frac{|x_t - \hat{x}_t|}{x_t + \hat{x}_t} \tag{3.44}$$

No entanto, se x_t for próximo de zero, $\hat{x}_t$ é provável ser próximo de zero também. Assim, a medida ainda envolve a divisão por um número próximo de zero, tornando o cálculo instável. Além disso, o valor de SMAPE pode ser negativo.

Apesar desta medida ser muito utilizada, Hyndman e Koehler [27] não recomendam a sua utilização para avaliar a precisão da previsão [25].

3.3.3 *Backtesting* - Validação cruzada de séries temporais

Uma forma mais sofisticada de avaliar a performance do modelo de previsão e a capacidade de generalização, usando dados históricos, é a utilização de validação cruzada de séries temporais, procedimento mais conhecido por *backtesting*[7].

Em termos gerais, este procedimento envolve vários conjuntos de teste, cada um constituído por uma única observação. A amostra de treino consiste apenas numa sequência de instantes que ocorreram antes da 1ª observação incluída na amostra de teste. Deste modo, nenhuma observação futura pode ser utilizada na construção da previsão [24]. Existem dois métodos principais de *backtesting*, sendo que um deles denomina-se por **janela móvel**.

Na janela móvel, o tamanho da amostra de treino mantém-se o mesmo e a janela "desliza" pelos dados, com o intuito de criar vários pares de treino-teste. O seguinte diagrama ilustra o que foi descrito anteriormente:

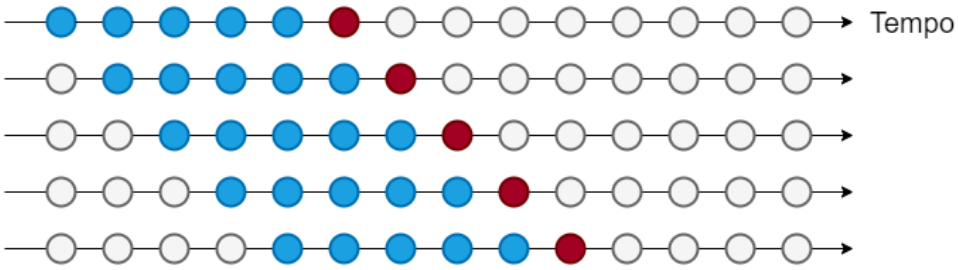

Figura 3.6: Janela móvel de tamanho 5 de previsão a 1-passo ($h = 1$), onde cada ponto corresponde a uma observação. As observações a azul correspondem às amostras de treino e as observações a vermelho formam as amostras de teste (adaptado de [24]).

A precisão da previsão é calculada através da média das amostras de teste.

Também se pode efetuar previsão a múltiplos-passos. Supondo que se pretende estudar um modelo que produza estimativas a 5-passos, o procedimento seria esquematicamente representado da seguinte maneira:

[7] Em ML, a validação cruzada *k-fold* é uma das formas mais conhecidas de validação cruzada. Quando se trata de séries temporais, não se deve aplicar esse tipo de validação, pois pretende-se prever o futuro com base no passado e com a validação cruzada *k-fold*, pode-se treinar com dados do futuro para prever o passado. Para não confundir os termos, decidiu-se denominar por *backtesting* nesta Dissertação.

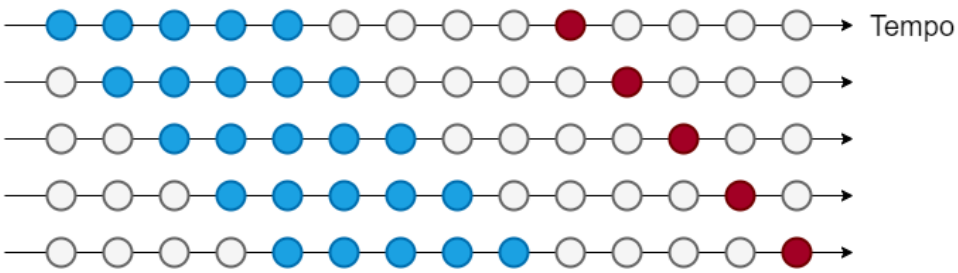

Figura 3.7: Janela móvel de tamanho 5 de previsão a 5-passos ($h = 5$), onde cada ponto corresponde a uma observação. As observações a azul correspondem às amostras de treino e as observações a vermelho formam as amostras de teste (adaptado de [24]).

3.4 Intervalos de previsão

Quando se obtém uma previsão, a estimativa obtida representa o meio do intervalo de valores possíveis que a variável aleatória pode assumir. No momento da previsão, é comum calcular, também, o seu **intervalo de previsão** [24].

O intervalo de previsão fornece um intervalo dentro do qual se espera que x_t se situe com uma determinada probabilidade. Numa forma geral, o intervalo de previsão pode ser escrito da seguinte forma

$$\hat{x}_{T+h|T} \pm c\hat{\sigma}_h \tag{3.45}$$

onde,
$\hat{\sigma}_h$ representa a estimativa do desvio padrão da distribuição de previsão no h-passo; e
c representa o valor crítico da distribuição *t-Student*[8];

Os intervalos de previsão são úteis, pois tornam mais claro o grau de incerteza associado a cada previsão. Por este motivo, as previsões pontuais não têm significância sem os intervalos de previsão que as acompanham [24].

Uma característica comum dos intervalos de previsão é o facto de aumentarem em amplitude à medida que o horizonte de previsão aumenta. Quanto maior for o horizonte de previsão, maior será a incerteza associada à previsão e, consequentemente, maiores são os intervalos de previsão. Ou seja, σ_h aumenta, normalmente, com h. No entanto, existem métodos de previsão não lineares que não possuem esta propriedade [24]. Para previsões a $h = 1$ passo, o desvio padrão residual fornece uma boa estimativa do desvio padrão previsto σ_1. Para previsão a h-passos, envolve métodos mais complicados para calcular. Estes cálculos assumem que os resíduos não são correlacionados [24].

[8]Geralmente, calculam-se intervalos a 95% cujo o valor do multiplicador c corresponde a 1.96.

Como existem inúmeros métodos de previsão disponíveis, cada um com as suas próprias abordagens para calcular os intervalos de previsão, e sendo o foco desta Dissertação mais direcionado para as análises e interpretações das previsões pontuais, apenas serão utilizadas as funções do *Python* dos modelos utilizados, que estimam de forma automática esses intervalos de previsão.

3.5 Modelos

A seguinte secção consiste na descrição dos modelos e algoritmos que vão ser usados para modelar as séries temporais, tendo em conta a posterior descrição e a análise dos dados deste trabalho.

Assim, são descritos os modelos ARIMA e regressão dinâmica, e os algoritmos Prophet e NeuralProphet.

3.5.1 Processo Auto-Regressivo Integrado de Média Móvel - ARIMA

O processo auto-regressivo de ordem p, AR(p), é um modelo clássico utilizado em séries temporais fracamente estacionárias e assenta no pressuposto de que a observação da variável no instante t se relaciona com observações da mesma variável em instantes anteriores [32], isto é,

$$X_t = \delta + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + ... + \varphi_p X_{t-p} + \varepsilon_t \tag{3.46}$$

onde,

δ representa uma constante;

$\varphi_1, ..., \varphi_p$ são os parâmetros de auto-regressão e representam constantes fixas desconhecidas a estimar; e

ε_t representa o erro gerado por um processo de ruído branco.

Admitindo que o processo é estacionário, então

$$\begin{aligned} \mu = E(X_t) &= \frac{\delta}{1 - \varphi_1 - \varphi_2 - ... - \varphi_p} \\ &\Leftrightarrow \delta = \mu(1 - \varphi_1 - \varphi_2 - ... - \varphi_p). \end{aligned} \tag{3.47}$$

A equação do modelo pode-se reescrever da seguinte forma

$$\begin{aligned} & X_t = \mu(1 - \varphi_1 - \varphi_2 - ... - \varphi_p) + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + ... + \varphi_p X_{t-p} + \varepsilon_t \\ \Leftrightarrow\ & X_t - \mu = \varphi_1(X_{t-1} - \mu) + \varphi_2(X_{t-2} - \mu) + ... + \varphi_1(X_{t-p} - \mu) \\ & \Leftrightarrow Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t, \text{ sendo } Y_t = X_t - \mu \\ & \Leftrightarrow \varepsilon_t = (1 - \varphi_1 L - \varphi_2 L^2 - ... - \varphi_p L^p) Y_t \\ & \Leftrightarrow \varepsilon_t = \varphi(L) Y_t \end{aligned} \tag{3.48}$$

onde,
L representa o operador *lag*; e
$\varphi(L)$ representa o polinómio auto-regressivo de ordem p.

Tal como no processo auto-regressivo, o processo de média móvel de ordem q, MA(q), é um modelo utilizado em sucessões cronológicas fracamente estacionárias. Em vez de usar os valores anteriores da variável preditiva, um modelo de média móvel utiliza os erros de previsão anteriores [20], isto é,

$$X_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \tag{3.49}$$

Sendo $Y_t = X_t - \mu$ a equação do modelo pode-se reescrever como

$$\begin{aligned} & Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \\ \Leftrightarrow\ & Y_t = (1 - \theta_1 L - \theta_2 L^2 - ... - \theta_q L^q)\varepsilon_t \\ \Leftrightarrow\ & Y_t = \theta(L)\varepsilon_t \end{aligned} \tag{3.50}$$

onde, $\theta(L)$ representa o polinómio de médias móveis de ordem q.

A combinação destes processos estocásticos estacionários resulta num modelo denominado ARMA(p, q). O objetivo é realizar inferências sobre um dado processo estocástico desconhecido [32], tal que,

$$X_t = \delta + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + ... + \varphi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \tag{3.51}$$

Admitindo que o processo é estacionário, tem-se

$$\varphi(L) Y_t = \theta(L)\varepsilon_t \tag{3.52}$$

Uma vez que a grande maioria das séries temporais não são estacionárias, é necessário recorrer a outros métodos que permitam a sua análise [32]. Algumas séries não estacionárias em média podem tornar-se estacionárias diferenciando uma ou mais vezes [37]

$$\Delta^d X_t = (1 - L)^d X_t, \text{ sendo } d \geq 1 \tag{3.53}$$

A ordem de diferenciação d é o menor número de vezes que a sucessão necessita de ser diferenciada para se tornar numa sucessão estacionária. Após a diferenciação pode-se utilizar o modelo ARMA [32].

Por outro lado, pode-se usar os modelos ARIMA($p,\ d,\ q$), que consistem na conversão de um processo não estacionário num processo estacionário através da diferenciação [32].

Os modelos ARIMA contemplam três tipos de processos distintos. Tal como no modelo ARMA, contém os processos AR de ordem p e MA de ordem q associados, e um processo de diferenciação I de ordem d [32]

$$\varphi(L)(1-L)^d Y_t = \theta(L)\varepsilon_t \tag{3.54}$$

onde, $(1-L)^d$ representa o processo de diferenciação de ordem d.

O modelo ARIMA($p,\ d,\ q$) não apresenta a componente da sazonalidade.

Os processos com sazonalidade assentam na hipótese que uma sucessão com sazonalidade S é caraterizada pela existência de correlação entre as observações espaçadas por um período múltiplo de S [37].

3.5.2 Processo Auto-Regressivo Integrado de Média Móvel com Sazonalidade - SARIMA

Os modelos SARIMA são uma extensão dos modelos ARIMA que suportam séries temporais univariadas com a componente sazonal, e podem ser representadas como SARIMA$(p,d,q)(P,D,Q)_S$.

Quanto à configuração dos seus hiperparâmetros, $p,\ d,\ q$ são iguais aos do modelo ARIMA, correspondem à parte não sazonal, enquanto que os hiperparâmtros $P,\ D,\ Q$ correspondem à parte sazonal, cujo D se refere à ordem de diferenciação sazonal [63]. Portanto, tem-se um processo auto-regressivo de ordem P com sazonalidade S - AR(P)$_S$ [37]

$$\begin{aligned} & Y_t = \varphi_1 Y_{t-s} + \varphi_2 Y_{t-2s} + \ldots + \varphi_P Y_{t-Ps} + \varepsilon_t \\ \Leftrightarrow\ & \varepsilon_t = (1 - \varphi_1 L^S - \varphi_2 L^{2S} - \ldots - \varphi_P L^{PS}) Y_t \\ \Leftrightarrow\ & \varepsilon_t = \varphi(L^S) Y_t \end{aligned} \tag{3.55}$$

onde, $\varphi(L^S)$ representa o polinómio auto-regressivo sazonal de ordem P.

E tem-se uma média móvel de ordem Q com sazonalidade S - MA(Q)$_S$ [37]

$$\begin{aligned} & Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-s} + \theta_2 \varepsilon_{t-2s} - \ldots - \theta_Q \varepsilon_{t-Qs} \\ \Leftrightarrow\ & Y_t = \theta(L^S)\varepsilon_t \end{aligned} \tag{3.56}$$

onde, $\theta(L^S)$ representa o polinómio de média sazonal de ordem Q.

A representação do SARIMA pode ser expressa da seguinte forma

$$\varphi(L)\varphi(L^S)(1-L)^d(1-L^S)^D Y_t = \theta(L)\theta(L^S)\varepsilon_t. \tag{3.57}$$

3.5.3 Função `auto_arima()`

Para ajustar os modelos ARIMA, é necessário seguir um procedimento específico, no entanto, existe uma função na biblioteca *Python* que permite modelar de forma automática esses modelos.

A função `auto_arima()` [52] do *Python* explora o modelo possível dentro das restrições de ordem fornecidas, e retorna o melhor modelo ARIMA de acordo com o valor do *Akaike's Information Criterion* (AIC) (parâmetro *default*).

O AIC dos modelos ARIMA define-se como

$$\text{AIC} = -2log(L) + 2(p + q + k + 1) \tag{3.58}$$

onde,

L representa a verossimilhança dos dados; e

k representa o número de parâmetros no modelo, se $k = 1$, então $c \neq 0$, e se se $k = 0$, então $c = 0$.

Quando o T é pequeno, neste caso, quando a amostra é relativamente pequena, é preferível recorrer ao *Akaike's Information Corrected Criterion* (AICc) [24]. O AICc dos modelos ARIMA define-se como

$$\text{AICc} = \text{AIC} + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - p - q - k - 2} \tag{3.59}$$

O *Bayesian Information Criterion* (BIC) dos modelos ARIMA, penaliza mais o número de parâmetros do que o AIC, e é definido como

$$\text{BIC} = \text{AIC} + [log(T) - 2](p + q + k + 1) \tag{3.60}$$

Bons modelos são obtidos minimizando o AIC, AICc ou o BIC [24].

Definidos o AIC, o AICc e o BIC, segue-se a descrição do comportamento *default* desta função [24]:

Algoritmo *Hyndman-Khandakar* para a modelação automática do ARIMA [26]

1. O número de diferenças $0 \leq d \leq 2$ é determinado fazendo repetições do teste KPSS.

2. Os valores p e q são posteriormente escolhidos minimizando o AIC após diferenciação dos dados d vezes. Ao invés de considerar todas as possíveis combinações de p e q, o algoritmo utiliza o método *stepwise* para percorrer o espaço do modelo.

   (a) Quatro modelos são ajustados:

   - ARIMA(0, d, 0),
   - ARIMA(2, d, 2),
   - ARIMA(1, d, 0),
   - ARIMA(0, d, 1),

   Uma constante é incluída, mas apenas no caso de $d \neq 2$. Se $d \leq 1$, um modelo adicional é também ajustado:

   - ARIMA(0, d, 0) sem uma constante.

   (b) O melhor modelo (com o menor valor AIC) é ajustado no passo (a) é considerado o "modelo atual"

   (c) Variações do "modelo atual" são consideradas:

   - Variar p e q em relação ao "modelo atual", com ±1;
   - Incluir/ excluir c do "modelo atual".

   O melhor modelo considerado (o "modelo atual" ou um das variações mencionadas) torna-se o novo "modelo atual".

   (d) Repetir o passo 2 (c) até que nenhum valor do AICc inferior possa ser encontrado.

3.5.4 Regressão dinâmica

Os modelos descritos no capítulo anterior, permitem a inclusão de informações provenientes de observações passadas de uma série, mas não a inclusão de outras informações que também possam ser relevantes. Por exemplo, os efeitos dos feriados/ eventos especiais ou de outras variáveis externas podem explicar algumas das variações históricas [24].

Considerando um modelo de regressão da forma

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \ldots + \beta_k X_{k,t} + \varepsilon_t \tag{3.61}$$

onde,

Y_t representa a variável dependente no tempo t que se pretende prever;

β_1, ..., β_k representam os parâmetros (desconhecidos) do modelo, refletindo o efeito esperado de cada preditor fixando os efeitos das restantes covariáveis; e

ε_t representa o termo de erro não correlacionado.

Os modelos de regressão possibilitam a inclusão de muitas informações relevantes a partir das variáveis preditoras, no entanto não permitem a dinâmica das séries temporais tal como pode ser tratada pelos modelos ARIMA [24].

Se permitir que os erros da Equação (3.61) sejam autocorrelacionados, os erros da série assumem seguir um modelo ARIMA. Para distinguir, vai-se substituir o erro ε_t por η_t. Por exemplo, se η_t seguir um modelo ARIMA(1,1,1), pode-se escrever da seguinte maneira

$$Y_t = \beta_0 + \beta_1 X_{1,t} + ... + \beta_k X_{k,t} + \eta_t, \tag{3.62}$$

Tal que

$$(1 - \varphi_1 L)(1 - L)\eta_t = (1 - \theta_1)\varepsilon_t \tag{3.63}$$

onde,

ε_t representa o ruído branco da série.

O modelo (3.63) possui dois erros - o erro do modelo de regressão, denotado por η_t, e o erro do modelo ARIMA[9], denotado por ε_t [24].

Assim, os modelos que combinam a regressão linear e o ARIMA num único modelo de regressão, denominam-se por **regressão dinâmica**.

A função `auto_arima()` [52], também, permite ajustar um modelo de regressão com erros ARIMA, se o argumento `xreg` for utilizado. O utilizador deve especificar as variáveis preditoras a serem incluídas, e a função irá selecionar o melhor modelo ARIMA para os erros. Se for necessário diferenciação, todas as variáveis são diferenciadas durante o processo de estimação, embora o modelo final seja expresso em relação às variáveis originais. O AIC é calculado para o modelo final, e este valor pode ser utilizado para determinar os melhores preditores. Ou seja, o procedimento deve ser repetido para todos os subconjuntos de preditores a serem considerados, e o modelo com o menor AIC é selecionado [24].

3.5.4.1 Regressão dinâmica harmónica

O modelo SARIMA é concebido para períodos mais curtos, como dados mensais ou dados trimestrais. O problema reside na estimação de $m - 1$ parâmetros para os estados sazonais iniciais, sendo que m corresponde ao período sazonal. Logo, para um elevado nº de parâmetros, estimar m torna-se quase impossível. Apesar do `auto_arima()` permitir

[9] Apenas os erros do modelo ARIMA são considerados ruído branco.

um período sazonal máximo de $m = 350$, na prática, o custo computacional é elevado a partir de $m = 200$, o que pode levar a cálculos lentos ou fazer com que o sistema fique sem memória. Adicionalmente, uma ordem de diferenciação sazonal grande não faz muito sentido para dados diários, já que envolve comparar o que aconteceu no presente com o que ocorreu exatamente há um ano [24].

Para séries temporais que exibem longos períodos sazonais, é preferível optar por uma regressão dinâmica harmónica (RDH) onde os padrões sazonais são modelados através de termos de *Fourier* utilizando erros ARMA [24].

Séries de *Fourier*:

O matemático Francês Jean-Baptiste Fourier mostrou que, uma série de senos e cossenos, para determinadas frequências, pode-se aproximar a qualquer função periódica [24]. Os **termos de *Fourier*** podem ser escritos da seguinte forma

$$\sin\left(\frac{2\pi kt}{m}\right) \text{ e } \cos\left(\frac{2\pi kt}{m}\right), \text{com } k = 1, 2, \ldots \tag{3.64}$$

onde,

k representa o número de pares seno e cosseno de *Fourier*, correspondendo à ordem da série.

Esta abordagem tem as seguintes vantagens:

- Permite qualquer comprimento de sazonalidade;
- Para séries que tenham mais do que um período sazonal, pode-se incluir termos de *Fourier* de diferentes frequências;
- A suavidade do padrão sazonal pode ser controlado por k, sendo que quanto menor for k, mais suave é o seu padrão;
- Pode-se recorrer simplesmente a erros ARMA quando a dinâmica é de curto prazo.

A desvantagem é o facto de se assumir que a sazonalidade é fixa, ou seja, o padrão sazonal não pode mudar ao longo do tempo. No entanto, na prática, a sazonalidade é normalmente constante, pelo que não é uma grande desvantagem, exceto para séries temporais longas [24].

3.5.5 Prophet

O Prophet é um algoritmo que foi introduzido pelo Facebook [54], e tem como objetivo prever séries temporais com base num modelo aditivo, onde as tendências não lineares são ajustadas à sazonalidade anual, semanal e diária, incluindo os efeitos de feriados. É mais eficaz em séries temporais que têm efeitos sazonais fortes. Para além disso, o Prophet é

robusto a dados ausentes e mudanças na tendência, e normalmente lida bem com valores discrepantes [16].

O Prophet inclui quatro componentes principais, podendo ser representado pela expressão

$$Y_t = g(t) + s(t) + h(t) + \varepsilon_t \tag{3.65}$$

onde,

Y_t representa a variável dependente no tempo t que se pretende prever;

$g(t)$ representa a função de crescimento que permite modelar as mudanças não periódicas do valor da série temporal;

$s(t)$ representa as mudanças periódicas devido à sazonalidade semanal ou anual;

$h(t)$ representa os efeitos dos feriados que ocorrem em horários potencialmente irregulares que podem incluir mais do que um dia; e

ε_t representa o termo do erro, que assume ser normalmente distribuído.

Na sua essência, funciona da seguinte maneira, de acordo com [55]:

- O Prophet deteta automaticamente alterações nas tendências selecionando pontos de alteração (*changepoints*) nos dados;
- Modela a componente sazonal anual através de séries de *Fourier*;
- Modela a componente sazonal semanal através de variáveis binárias;
- O utilizador fornece uma lista de feriados.

Este modelo tem a vantagem de ser computacionalmente mais rápido de estimar do que o modelo RDH, sendo completamente automatizado. Contudo, é pouco frequente que resulte numa precisão melhor em comparação com esse modelo [25].

3.5.6 NeuralProphet

A maioria dos problemas em análise de séries temporais requer que as previsões sejam interpretáveis. Simultaneamente pretende-se que as previsões sejam eficientes. Estes dois objetivos resultam num *trade-off*: capacidade de interpretação vs desempenho. Por exemplo, a eficiência, geralmente, implica com um modelo mais complexo. No entanto, essa complexidade raramente é sinónimo de capacidade de interpretação [59].

Existem duas classes de modelos:

- **Modelos estatísticos**: são matematicamente explicáveis e focam-se na relação entre variáveis e a sua significância (por exemplo, ARIMA).

- **Redes neuronais**: são mais complicadas, mas têm como objetivo obter a melhor performance possível.

Originalmente, o Prophet foi concebido para ser uma ferramenta facilmente funcional, ajustável e explicável para prever séries temporais, no entanto, tem uma performance baixa relativamente a outras abordagens [59]. Foi para colmatar essas limitações que o NeuralProphet foi criado.

O **NeuralProphet**, sucessor do Prophet, combina redes neuronais e algoritmos tradicionais de séries temporais, inspirados no Prophet e redes neuronais auto-regressivas, com o intuito de melhorar a eficiência e, ao mesmo tempo, limitar a perda de interpretabilidade [59].

O conceito central do modelo, é a sua capacidade de composição modular. O modelo é composto por módulos, cada um dos quais contribui com uma componente aditiva para a previsão, e todos os módulos devem produzir h *outputs*, isto é, previsões a h-passos. Estes valores são adicionados como valores previstos $\hat{y}_t, ..., \hat{y}_{t+h-1}$ para os valores futuros da série temporal $y_t, ..., y_{t+h-1}$. Caso o modelo dependa apenas do tempo, um número arbitrário de previsões pode ser gerado [58]. Segue-se a representação matemática equivalente a uma previsão a h-passos

$$Y_{t+h-1} = T(t+h-1) + S(t+h-1) + E(t+h-1) + F(t+h-1) + A(t+h-1) + L(t+h-1), \forall\, h \in [1, h] \tag{3.66}$$

onde,

$T(t+h-1)$ representa a tendência no tempo $t+h-1$;

$S(t+h-1)$ representa os efeitos da sazonalidade no tempo $t+h-1$;

$E(t+h-1)$ representa os efeitos dos eventos e feriados no tempo $t+h-1$;

$F(t+h-1)$ representa os efeitos da regressão no tempo $t+h-1$ para variáveis exógenas futuras conhecidas;

$A(t+h-1)$ representa os efeitos de auto-regressão no tempo $t+h-1$ com base em observações passadas; e

$L(t+h-1)$ representa os efeitos de regressão no tempo $t+h-1$ para observações passadas de variáveis exógenas.

Por exemplo, uma previsão a 1-passo à frente é equivalente a escrever a relação passada com $h = 1$

$$Y_t = T(t) + S(t) + E(t) + F(t) + A(t) + L(t). \tag{3.67}$$

Para além de melhorar a eficiência e limitar a perda da interpretabilidade, o NeuralProphet ultrapassa o Prophet abordando as seguintes questões:

- O NeuralProphet é altamente escalável, o que significa que consegue lidar com volumes grandes de dados e requisitos computacionais crescentes.
- Tem uma *framework* extensível, uma vez que é treinado, também, com métodos de DL.
- O NeuralProphet introduz contexto local com o suporte da auto-regressão e covariáveis do passado.

Relativamente à análise dos resíduos:

- O NeuralProphet não assume explicitamente nenhuma distribuição específica para os resíduos;
- Uma outra vantagem do NeuralProphet é a flexibilidade no tratamento de diferentes tipos de dados, incluindo aqueles com resíduos com distribuição não gaussiana;
- Estes modelos são concebidos para serem robustos e não se baseiam em pressupostos de normalidade.

4

Dados

Neste capítulo descreve-se a forma como foi efetuada a recolha e a seleção dos dados disponibilizados, bem como o seu pré-processamento e a realização de uma análise exploratória de dados temporais.

Como o estudo se foca na previsão de ausências não programáveis, foram disponibilizadas por parte de um operador ferroviário da Europa do Norte informações relevantes sobre os dados biográficos dos tripulantes e o histórico das atividades planeadas e das ausências realizadas entre 2009 e 2021, por cerca de 4000 trabalhadores, incluindo maquinistas e revisores, distribuídos geograficamente por mais de 30 bases operacionais.

Para aceder a essas informações, o cliente forneceu à SISCOG uma base de dados relacional em sistema *Oracle*, e os dados foram acedidos através de *queries SQL* juntamente com a linguagem *Python*. Para realizar as restantes etapas, recorreu-se apenas à linguagem de programação *Python*. Para viabilizar a análise do estudo, utilizaram-se as seguintes bibliotecas *Python*: *pandas* [56], *matplotlib*[23], *numpy* [22] e *statsmodels* [47].

Como os dados em questão são de elevada sensibilidade, os mesmos foram anonimizados pela SISCOG para possibilitar o estudo.

4.1 Descrição dos conjuntos de dados

4.1.1 Dados relacionados com atividades planeadas e ausências

A partir da base de dados, foram extraídos 66 conjuntos de dados. Cada conjunto é constituído por tuplos que contêm informação diária agregada relacionada com o número total de tripulantes que foram trabalhar e o número total de tripulantes ausentes, segmentados por tipo de função, isto é, maquinistas e revisores, e por base operacional (existindo no total registos sobre 33 bases operacionais).

É também importante referir que, para obter os resultados de cada coluna, foi necessário efetuar filtragens. Neste caso, para obter o número total de tripulantes que foram trabalhar, contabilizaram-se apenas os registos das pessoas que tiveram um dia de trabalho "normal" ou *part time*, descartando as folgas. Por último, como referido anteriormente, este projeto tem como objetivo estudar as ausências não programáveis e, portanto, os únicos registos

disponíveis na base de dados relacionados com este tipo de ausências são o absentismo em caso de **doença** ou **apoio à família**. Assim sendo, filtraram-se também os registos por tipo de ausência, considerando apenas estes dois tipos.

A Tabela 4.1 resume os atributos que compõe cada conjunto.

Tabela 4.1: Atributos de cada conjunto de dados.

Atributo	Tipo	Conteúdo
`day`	*int*	Dia do mês
`month`	*int*	Mês
`year`	*int*	Ano
`count_absence`	*int*	Número de tripulantes ausentes
`count_duties`	*int*	Número de tripulantes que foram trabalhar

Inicialmente, ao exportar os resultados, descobriu-se que determinados tripulantes poderiam exercer as duas funções, maquinista e revisor, por acumularem os dois tipos de contrato. Ou seja, para os trabalhadores que têm os dois tipos de contrato, o mesmo trabalhador vai ser contabilizado tanto no conjunto segmentado por maquinista como no conjunto segmentado por revisor, para a mesma base. No entanto, sabe-se que os tripulantes só podem realizar umas das funções em determinada data, dependendo de qual dos contratos se encontra ativo nesse momento. Logo, para corrigir o problema, quando se filtrou os registos de trabalho/ausência por tipo de função que cada tripulante realizaria, assumiu-se apenas, para cada trabalhador, a função correspondente ao contrato ativo nessa data.

Os resultados obtidos foram guardados em dois ficheiros *.xlsx*, um ficheiro relativo aos registos dos maquinistas e outro para os revisores. Cada ficheiro *.xlsx* inclui várias folhas de cálculo. Uma folha de cálculo contém o conjunto de dados respetivo a uma determinada base operacional.

Para facilitar o manuseamento desses dados, cada conjunto foi convertido num objeto *DataFrame* da biblioteca *Pandas* [56]. De seguida, com base no tipo de função desempenhada pelos tripulantes, criaram-se dois dicionários, de forma a guardar as *DataFrames* com uma chave associada, neste caso, à base operacional. No final, guardaram-se os dicionários num ficheiro *pickle* associado aos registos da função desempenhada, conforme se resume na Tabela 4.2:

Tabela 4.2: Resumo dos conjuntos de dados obtidos.

Nome do ficheiro	Função	Conteúdo
`drivers_daily_data.p`	Maquinistas	33 conjuntos de dados com os atributos da Tabela 4.1, cada um associado a uma base operacional.
`guards_daily_data.p`	Revisores	

4.2 Pré-processamento dos dados

Nesta secção, descrevem-se os problemas encontrados durante a análise dos conjuntos de dados e as abordagens utilizadas, nomeadamente a limpeza e a imputação de valores omissos. Estas abordagens têm como objetivo melhorar a qualidade de cada conjunto, tornando-os mais adequados para as tarefas seguintes. Dado o número elevado de bases operacionais, não vão ser estudados todos os conjuntos de dados disponíveis e, portanto, esta secção ajuda a realizar uma seleção prévia das bases operacionais a serem consideradas.

4.2.1 Limpeza e imputação dos dados

Como se viu anteriormente, os dados contêm o ano, mês e dia como atributos separados. Para que os dados sejam reconhecidos como séries temporais pela biblioteca *Pandas* [56] é necessário transformar esses atributos, neste caso, as colunas `'year', 'month', 'day'`, num único objeto *Timestamp*, e, de seguida, representá-lo como índice do conjunto de dados, para que o manuseamento dos dados possa ser realizado em relação ao elemento do tempo. Assim, eliminaram-se essas três colunas dos dados e o índice de cada *dataset* passou a conter datas dos registos, conforme se descreve na Tabela 4.3:

Tabela 4.3: Estrutura dos conjuntos de dados com a atualização do índice.

Componente	Nome	Conteúdo	Tipo
Índice	`date`	Data (formato yyyy-mm-dd hh:mm:ss)	*datetime64[ns]*
Atributos	`count_absence`	Número de tripulantes ausentes	*int*
	`count_duties`	Número de tripulantes que foram trabalhar	*int*

No estudo embrionário realizado pela SISCOG [39], os autores recorreram a um histórico de 7 anos. Como referido na introdução deste capítulo, os dados disponibilizados referem-se ao período entre o ano 2009 e o ano 2021, no entanto, a partir das informações daquele estudo, sabe-se que do início do ano 2009 até dezembro de 2012 existem registos com dados incorretos. Para exemplificar, o número de tripulantes que faltaram é desproporcionalmente grande face ao número de tripulantes que foram trabalhar. Para além disso, dados que incluem o ano 2020 devem ser removidos, pois foi um período de grande instabilidade em diversas áreas devido à pandemia do Covid-19, tendo uma enorme repercussão no trabalho. Logo, garantiu-se que os *datasets* não incluíssem, pelo menos, informação entre o ano 2009 e dezembro de 2012 e entre janeiro de 2020 e 2021.

Posteriormente, decidiu-se averiguar se cada *dataset* obtido tinha pelo menos 6 anos de dados dentro do período estipulado, o que corresponde aproximadamente a 2191 linhas. Para a mesma base, se um dos conjuntos de dados tiver um número total de instâncias inferior a 2191, todos os conjuntos de dados relativos a essa base são descartados. Assim, este procedimento resultou na remoção de 10 bases operacionais, permanecendo

informação para as restantes 23 bases operacionais, restando assim um total de 46 conjuntos de dados (quando segmentados por tipo de função desempenhada). Ademais, foi possível perceber que muitos dos conjuntos contêm menos de 7 anos de histórico.

Com base na análise precedente, verificou-se os conjuntos de dados que continham valores omissos - *Not a Number* (NaN) - que apresentavam dias em falta, isto é, dias que não têm registo do número de pessoas ausentes, do número de pessoas que foram trabalhar ou ambos, entre o período de janeiro de 2013 e dezembro de 2019. Durante a análise, descobriu-se que alguns *datasets* apresentavam os seguintes casos:

1. Existem registos do número de tripulantes que foram trabalhar, mas não existem registos do número de tripulantes ausentes num determinado dia;

2. Existem registos do número de tripulantes ausentes, mas não existem registos do número de tripulantes que foram trabalhar num determinado dia;

3. Não existem registos nem do número de tripulantes ausentes nem do número de tripulantes que foram trabalhar num determinado dia.

Relativamente ao caso 1, considerou-se que nesses dias não houve registos de ausências provocadas por doença nem por apoio à família, portanto, apenas para os casos cuja coluna do número de tripulantes ausentes apresentasse NaN, esses valores foram substituídos por 0. Quanto aos casos 2 e 3, assumiu-se que a falta de registos pode-se dever a erros informáticos ou falhas no sistema, pois é impossível não existirem registos de tripulantes que foram trabalhar, uma vez que os operadores ferroviários necessitam de manter o normal funcionamento das operações dos comboios durante 7 dias por semana e 24 horas por dia. Para esses casos, os valores em falta foram substituídos através de métodos de imputação univariados.

A utilização destes métodos é uma decisão muito importante, pois pode introduzir erros ou incertezas nos dados, principalmente, se o número de valores omissos for elevado. Como não se pretende estudar todas as bases operacionais, é razoável, nesta situação, descartar os conjuntos de dados que tenham muitos valores em falta. Portanto, neste estudo, removeram-se conjuntos de dados com base no seguinte critério:

- Para a mesma base, se um dos seus conjuntos de dados possuir uma percentagem de valores omissos superior a 5%, então pode-se descartar essa base operacional. Neste caso, a contabilização dos valores omissos tem em conta o que se descreveu nos casos 2. e 3..

Com este critério, o número de bases operacionais a serem estudadas reduziu de 23 bases operacionais para 20, obtendo-se um total de 40 conjuntos de dados.

Por fim, para imputar os valores em falta, recorreu-se ao método de interpolação linear. Este método constrói novos pontos a partir dos conjuntos discretos de pontos já conhecidos,

e tem a vantagem de ser computacionalmente eficiente. No entanto, a utilização deste não é muito precisa, o que pode afetar a qualidade dos dados. Devido ao elevado número de bases operacionais e tendo em conta a percentagem de valores omissos de cada conjunto, optou-se por aplicar este método.

Através da interpolação linear, cada conjunto de dados obtido passou a ter no total 2556 observações, correspondendo à informação entre janeiro de 2013 e dezembro de 2019 (7 anos consecutivos).

4.3 Análise dos dados

Após o pré-processamento dos dados, foram analisados os *datasets* por forma a explorar mais informações relevantes para tomar decisões informadas no que toca à seleção e criação de novas *features*, bem como na escolha dos algoritmos.

A relação entre a dimensão/tamanho de uma base operacional e o número de tripulantes que esta regista pode variar dependendo de vários fatores. Geralmente, bases operacionais maiores apresentam um número maior de trabalhadores. No contexto desta Dissertação, o número de tripulantes corresponde ao número de maquinistas e revisores planeados para trabalhar.

Para averiguar o tamanho de cada base, estudou-se o número médio diário de maquinistas e de revisores planeados por cada base operacional, ao longo do último ano dos dados. Para contabilizar o número total de tripulantes planeados, criou-se uma nova coluna para cada conjunto de dados denominada por `'count_planned'`. Esta variável resulta da soma entre o número de pessoas ausentes e o número de pessoas que foram trabalhar

$$Nº\, tripulantes\, planeados = n_1 + n_2 \tag{4.1}$$

onde,

n_1 representa o número de tripulantes ausentes (maquinistas ou revisores); e

n_2 representa o número de tripulantes que foram trabalhar (maquinistas ou revisores).

A seguinte Tabela apresenta os conjuntos atualizados com a introdução desta variável:

Tabela 4.4: Estrutura dos conjuntos de dados com a introdução da variável `count_planned`.

Componente	Nome	Conteúdo	Tipo
Índice	`date`	Data (formato yyyy-mm-dd hh:mm:ss)	*datetime64[ns]*
Atributos	`count_absence`	Número de tipulantes ausentes	*int*
	`count_duties`	Número de tripulantes que foram trabalhar	*int*
	`count_planned`	Número de tripulantes planeados para trabalhar	*int*

Posteriormente, somou-se o número de tripulantes planeados para trabalhar, para cada base e calculou-se a média diária do último ano dos dados (2019). Este número total de tripulantes planeados para uma determinada base constitui um bom indicador do

tamanho da base. O número que se pode ler na designação de cada base (por exemplo 'B01'), adotada de ora em diante nesta Dissertação, corresponde ao número de ordem dessa base na lista de bases ordenada por ordem decrescente de tamanho.

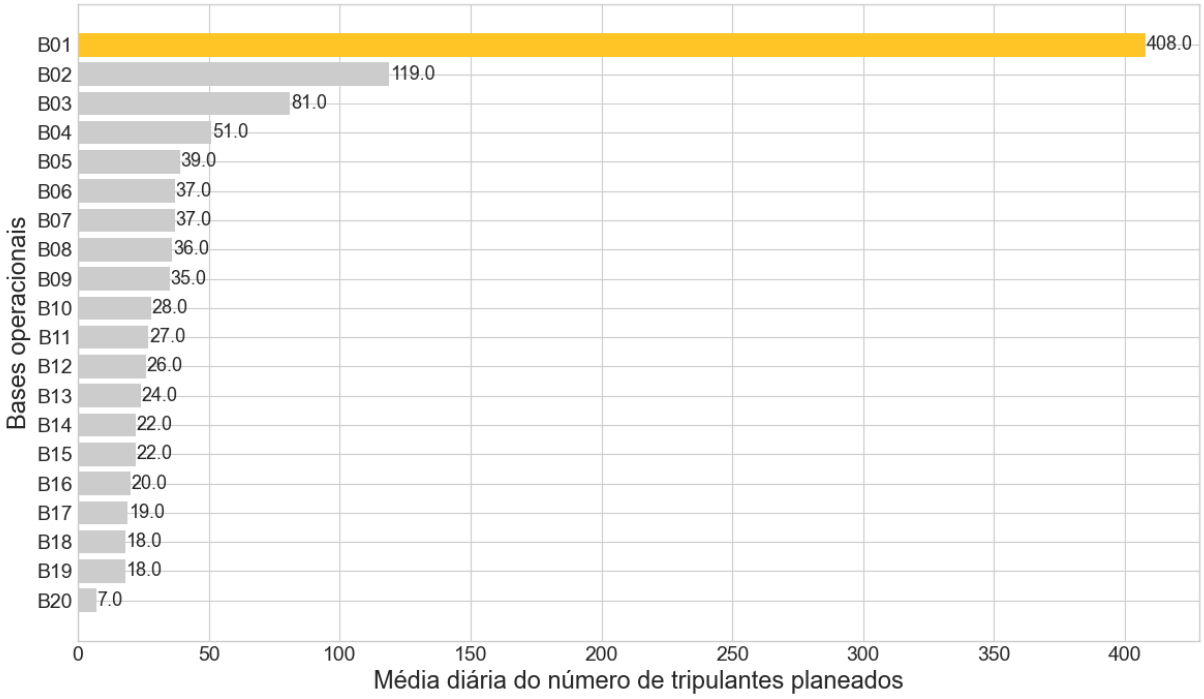

Figura 4.1: Média diária de tripulantes planeados do ano 2019.

Pode-se observar na Figura 4.1, que a base operacional B01 regista o maior número médio diário de tripulantes planeados, o que sugere ser uma base operacional relativamente grande. As restantes bases operacionais não se destacam tanto como a primeira. Nesta Dissertação, foi selecionada apenas a base operacional onde se verificaram os maiores registos médios diários de tripulantes planeados.

Uma vez que os dados encontram-se segmentados por tipo de função desempenhada (maquinistas e revisores) foram estudados 2 conjuntos de dados.

Assim, as próximas secções apenas irão apresentar o estudo para os *datasets* da base B01, segmentadas por maquinistas e revisores.

4.3.1 Análise exploratória das séries temporais

4.3.1.1 Caracterização das séries temporais

Estudos anteriores revelaram que o número de contratos cresce ao longo do tempo, e que existem menos turnos aos fins de semana. Para que o nível de cada série temporal

relativa ao número de tripulantes ausentes (maquinistas e revisores) por cada base não fosse afetada por estas variações, decidiu-se calcular a percentagem de absentismo por dia. Por outro lado, a percentagem de absentismo permite efetuar uma comparação mais intuitiva, `'absence_rate'`, facilitando a identificação de padrões, relações e desempenho relativo. Assim, definiu-se a seguinte métrica:

$$Absentismo\ (\%) = \frac{n_1}{n_1 + n_2} \times 100. \tag{4.2}$$

onde,

n_1 representa o número de tripulantes ausentes (maquinistas ou revisores); e

n_2 representa o número de tripulantes que foram trabalhar (maquinistas ou revisores).

A Tabela 4.5 resume os novos conjuntos de dados resultantes do procedimento anteriormente descrito.

Tabela 4.5: Estrutura dos conjuntos de dados com a introdução da variável `'absence_rate'`.

Componente	Nome	Conteúdo	Tipo
Índice	`date`	Data (formato yyyy-mm-dd hh:mm:ss)	*datetime64[ns]*
Atributos	`count_absence`	Número de tipulantes ausentes	*int*
	`count_duties`	Número de tripulantes que foram trabalhar	*int*
	`count_planned`	Número de tripulantes planeados para trabalhar	*int*
	`absence_rate`	Percentagem de ausências diária	*float*

Como referido no Subcapítulo 1.1, o planeamento tático tem como objetivo alocar a capacidade disponível com um horizonte de planeamento de 2 meses até um ano, isto significa que uma das responsabilidades do planeador é atribuir tarefas aos tripulantes e apresentar os dias em que as mesmas serão realizadas ao longo do ano. A atribuição das tarefas é feita em conformidade com os contratos que se encontram ativos e que abrangem esse período de tempo. Logo, no momento da previsão, pode-se converter a percentagem de ausências, de novo, em número de ausências multiplicando-a pelo número total de trabalhadores, Equação 4.1, planeados para esse dia.

Nas Figuras 4.2 e 4.3 representaram-se as séries referentes às percentagens diárias de absentismo desde janeiro de 2013 a dezembro de 2019, segmentadas por tipo de função profissional da base B01.

Relativamente ao gráfico da Figura 4.2, pode-se identificar de imediato a presença de alguns *outliers* ao longo do tempo. Existem dias em que a percentagem de absentismo apresenta alterações súbitas e não sistemáticas, o que torna difícil a identificação de possíveis padrões. Observando o gráfico da Figura 4.3, pode-se constatar a existência de "picos" periódicos que ocorrem a cada 12 meses, especialmente no mês de dezembro e no início do ano, de onde se depreende que há sazonalidade anual. Tendo em conta que

se filtraram as ausências provocadas por doença e por apoio à família, é de esperar que as séries tenham este comportamento, pois existe um aumento de gripe sazonal e outros vírus respiratórios nas alturas em que a temperatura é mais baixa.

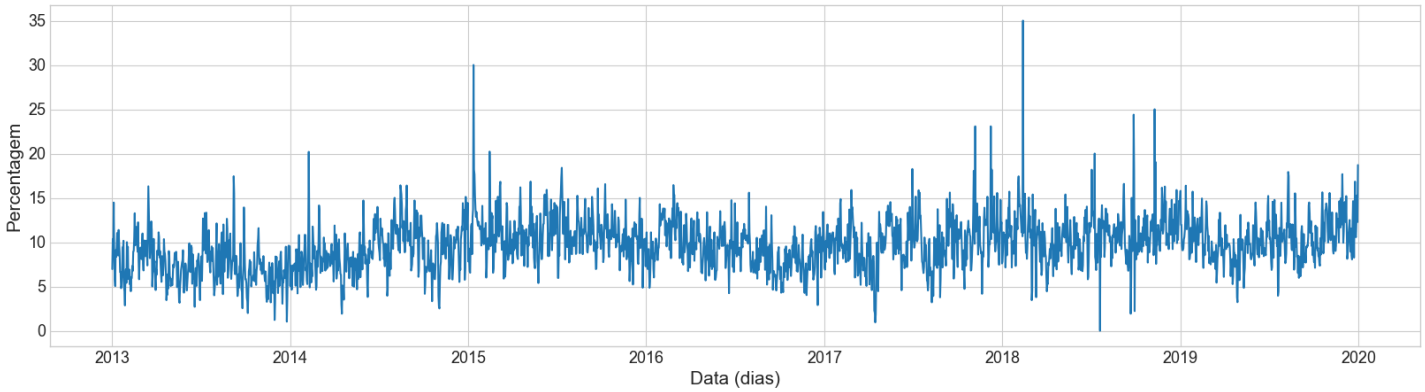

Figura 4.2: Série temporal da percentagem de ausências diárias dos maquinistas na base B01 desde janeiro de 2013 até dezembro de 2019.

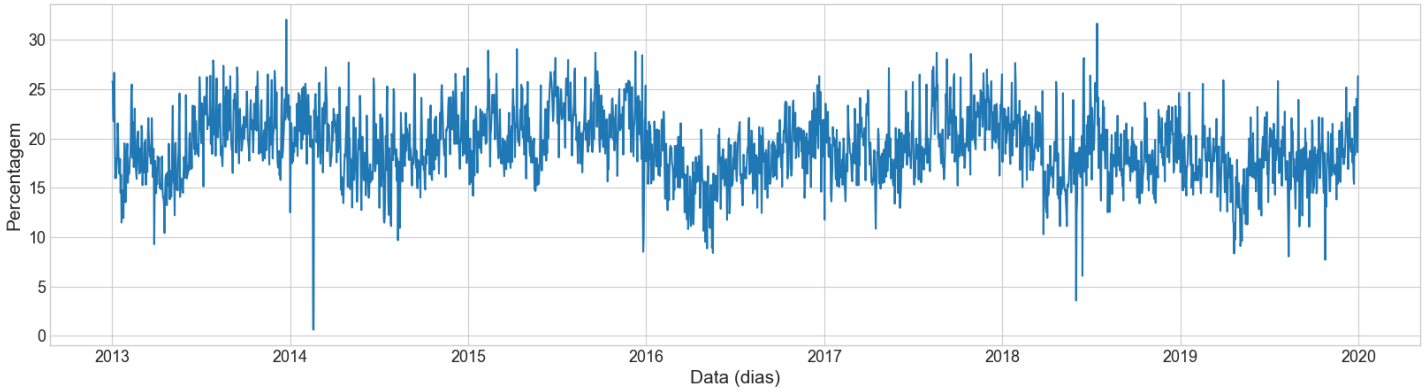

Figura 4.3: Série temporal da percentagem de ausências diárias dos revisores na base B01 desde janeiro de 2013 até dezembro de 2019.

De forma a estudar as duas séries temporais foi feita uma decomposição, com recurso ao método MSTL. Como a frequência dos dados é diária, é possível ter um padrão semanal a cada 7 dias e um padrão anual a cada 365 dias. Para o padrão semanal, sabe-se que, geralmente, o número de tripulantes ausentes é menor aos fins de semana, por haver menos turnos. Também é de esperar um padrão anual, o número de tripulantes ausentes atinge o maior pico nos meses mais frios do que nos meses mais quentes.

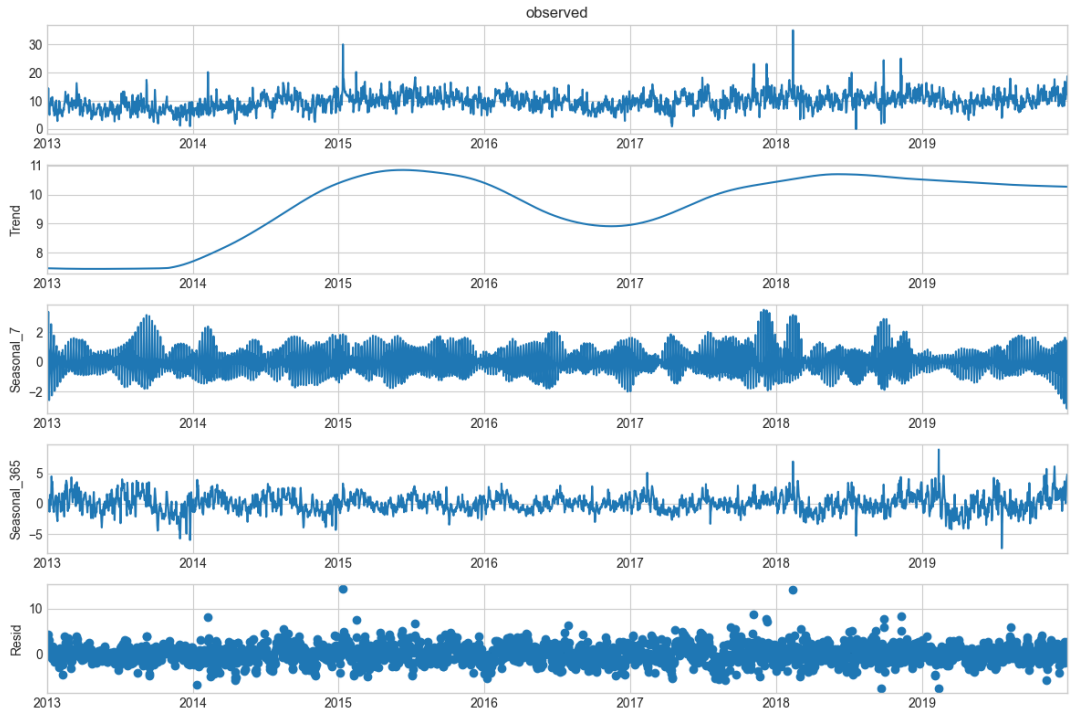

Figura 4.4: Decomposição MSTL da série temporal da percentagem de ausências diárias dos maquinistas na base B01 em tendência, sazonalidade semanal (`Seasonal_7`), sazonalidade anual (`Seasonal_365`) e componente residual.

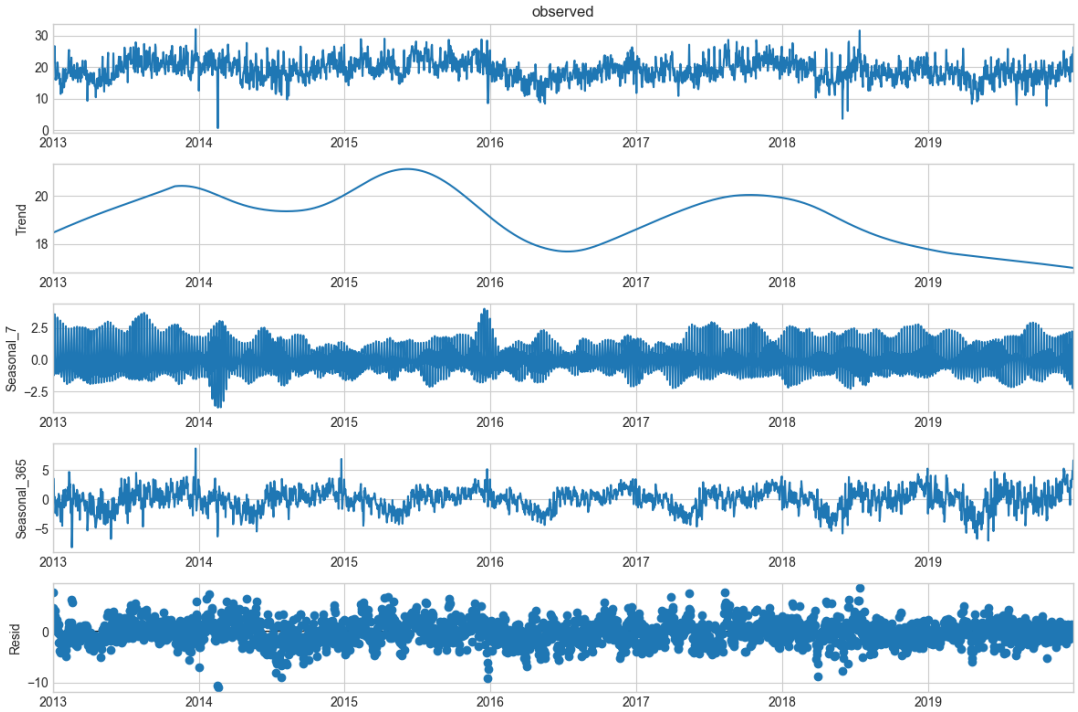

Figura 4.5: Decomposição MSTL da série temporal da percentagem de ausências diárias dos revisores na base B01 em tendência, sazonalidade semanal (`Seasonal_7`), sazonalidade anual (`Seasonal_365`) e componente residual..

Relativamente à tendência, na Figura 4.4 consegue-se visualizar uma tendência não linear crescente, enquanto que na Figura 4.5 tem-se uma tendência não linear decrescente.

Quanto às componentes da sazonalidade, ambas as séries possuem sazonalidade anual que se repete a cada 365 dias. Também se pode verificar que esta componente muda ao longo do tempo de forma semi-multiplicativa (no início e no fim), o que indica que a componente sazonal anual não é propriamente aditiva, sugerindo que é necessário uma transformação *Box Cox* em ambas as séries. Para inspecionar a componente da sazonalidade semanal, examinaram-se as 4 semanas do mês de janeiro (mês frio) e do mês de julho (mês quente) do ano 2019. Uma das vantagens da utilização da decomposição MSTL é o facto de permitir capturar a sazonalidade que muda ao longo do tempo.

Observando as Figuras 4.6 e 4.7, há um padrão repetitivo de picos ao longo da semana. No caso dos maquinistas, o menor pico verifica-se aos fins de semana no mês de janeiro de 2019, enquanto que no caso dos revisores, registam-se picos máximos e mínimos aos sábados e aos domingos, respetivamente. Relativamente à série dos revisores, no mês de julho, o padrão diário é semelhante ao padrão encontrado no mês de janeiro, exceto em certos dias da semana, em que o comportamento é diferente.

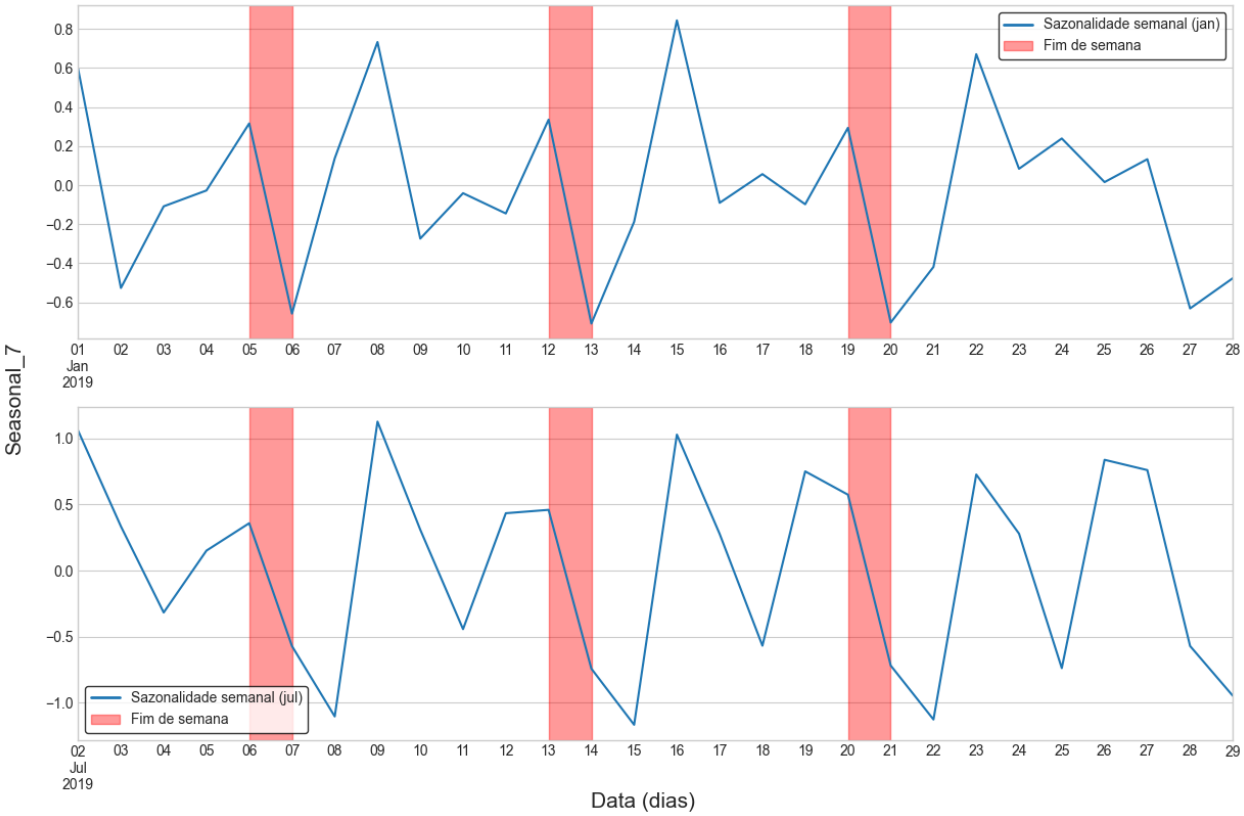

Figura 4.6: Componente sazonal semanal da série temporal da percentagem de ausências diárias dos maquinistas na base B01, extraída pela decomposição MSTL durante um período de 4 semanas em janeiro e julho do ano 2019.

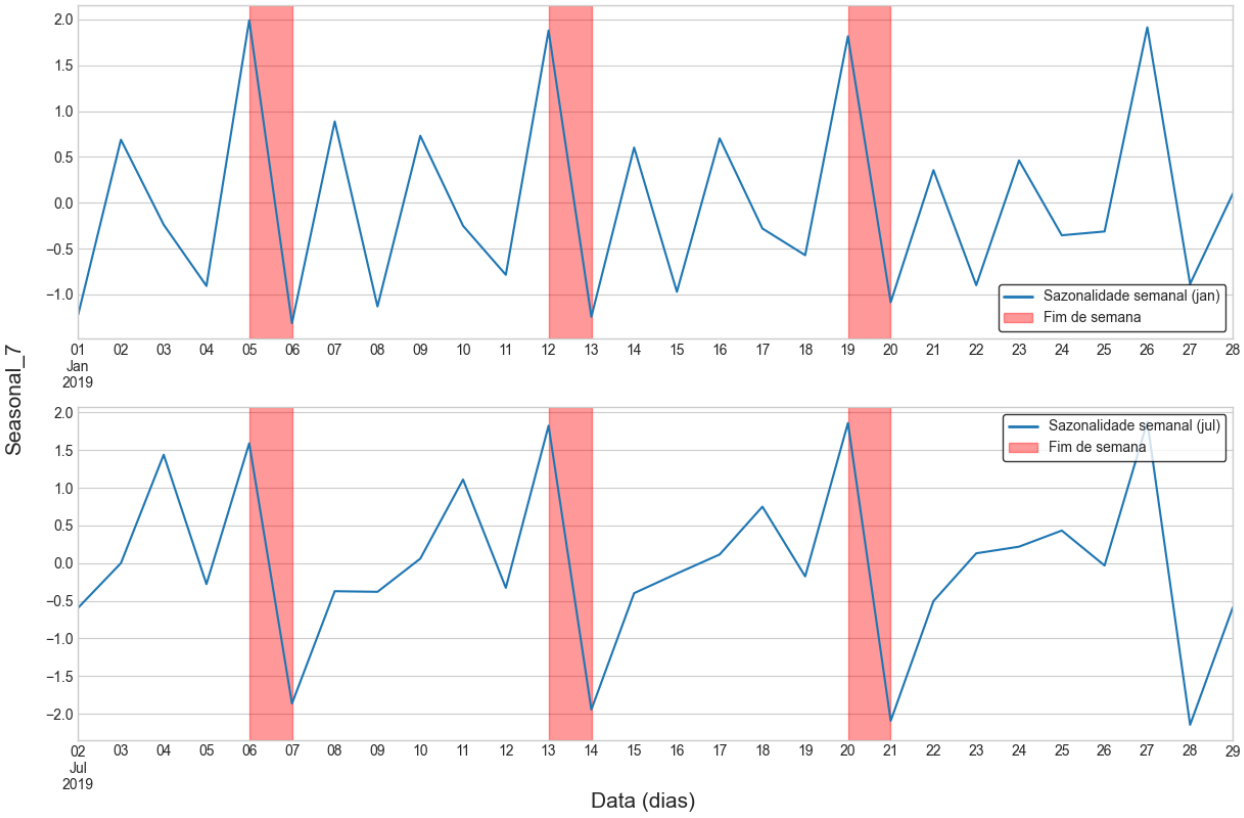

Figura 4.7: Componente sazonal semanal da série temporal da percentagem de ausências diárias dos revisores na base B01, extraída pela decomposição MSTL durante um período de 4 semanas em janeiro e julho do ano 2019.

Nas Figuras 4.8 e 4.9 representaram-se graficamente as funções de autocorrelação das duas séries temporais relativas às percentagens de ausências diárias dos maquinistas e revisores, na base B01, desde janeiro de 2013 até dezembro de 2019, respetivamente.

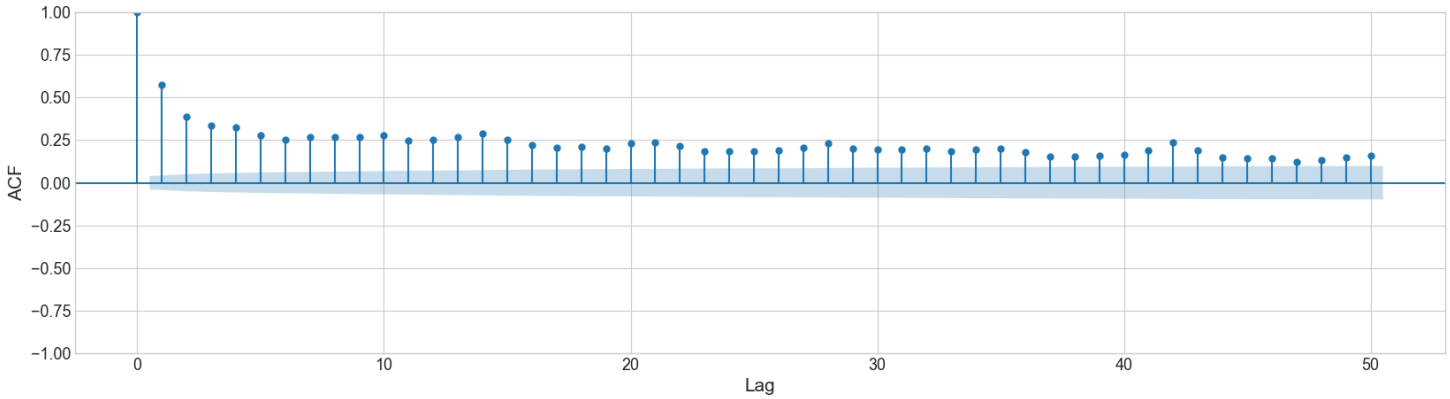

Figura 4.8: Correlograma da série temporal da percentagem de ausências diárias dos maquinistas na base B01. O eixo das abcissas representa o número de *lags*. A área sombreada a azul indica intervalo de confiança de 95%.

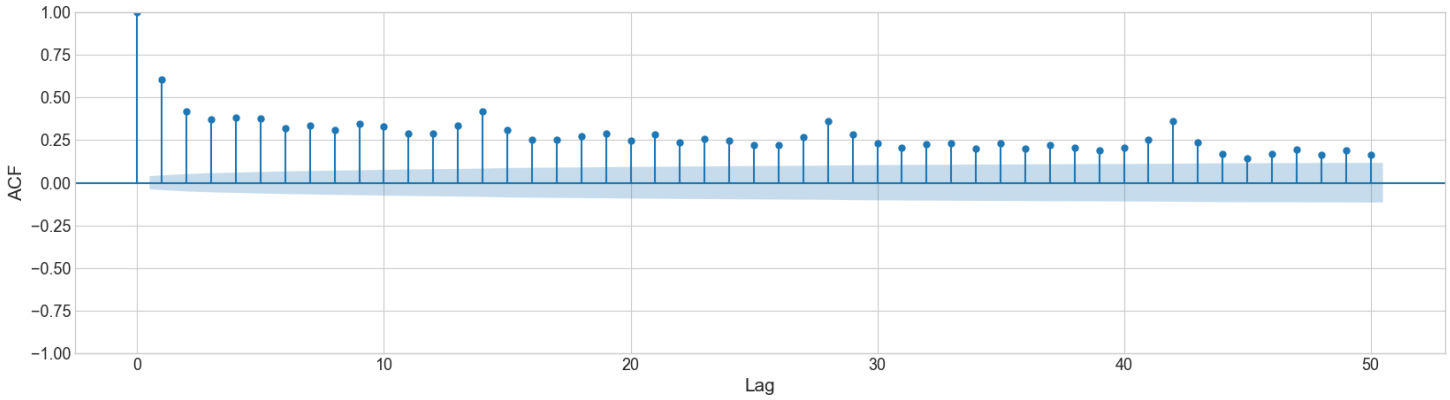

Figura 4.9: Correlograma da série temporal da percentagem de ausências diárias dos revisores na base B01. O eixo das abcissas representa o número de *lags*. A área sombreada a azul indica intervalo de confiança de 95%.

Conforme se pode observar nas Figuras 4.8 e 4.9, os valores só se aproximam dos limites do intervalo de confiança ao fim de muitos *lags*, concluindo serem bastante significativos. Também se pode verificar que certos coeficientes de autocorrelação são maiores em cada intervalo de tempo. Por exemplo, na Figura 4.9 os *lags* 14, 28 e 42, apresentam correlações visivelmente mais elevadas. Apesar deste comportamento não ser tão evidente na Figura 4.8, também se verifica nesta série. Estes comportamentos sugerem que as duas séries temporais exibem tendência e sazonalidade.

Para averiguar se as séries são ou não estacionárias, realizou-se o teste ADF e o teste KPSS para descobrir se as séries têm raiz unitária ou se são tendência estacionária:

- **Teste ADF:** Para a série temporal relativa à percentagem de ausências dos maquinistas, conclui-se que, a série temporal é estacionária ($p - value < 0.01$). Para a série temporal relativa à percentagem de ausências dos revisores, conclui-se que, a série temporal não é estacionária ($p - value \approx 0.016$).

- **Teste KPSS:** Para a série temporal relativa à percentagem de ausências dos maquinistas, conclui-se que, a série temporal é estacionária ($p - value < 0.01$). Para a série temporal relativa à percentagem de ausências dos revisores, conclui-se que, a série temporal é estacionária em tendência ($p - value \approx 0.025$).

Os resultados dos testes ADF e o KPSS são contraditórios quanto à estacionaridade dos dados. Por outro lado, existem evidências de que as duas séries não são estacionárias. Verificou-se que ambas as séries possuem sazonalidade semanal e anual através da decomposição, e, pela análise da ACF, verificaram-se autocorrelações elevadas durante muitos *lags*.

4.3.1.2 Identificação de *outliers*

Após a caracterização das séries temporais, procedeu-se à identificação de *outliers*. A identificação destes e compreensão dos fenómenos que estão na sua origem, permitem decidir a sua incorporação ou não no modelo e, por conseguinte, nas previsões a efetuar [37].

A Figura 4.10 representa a distribuição das duas séries temporais por meio do *boxplot*.

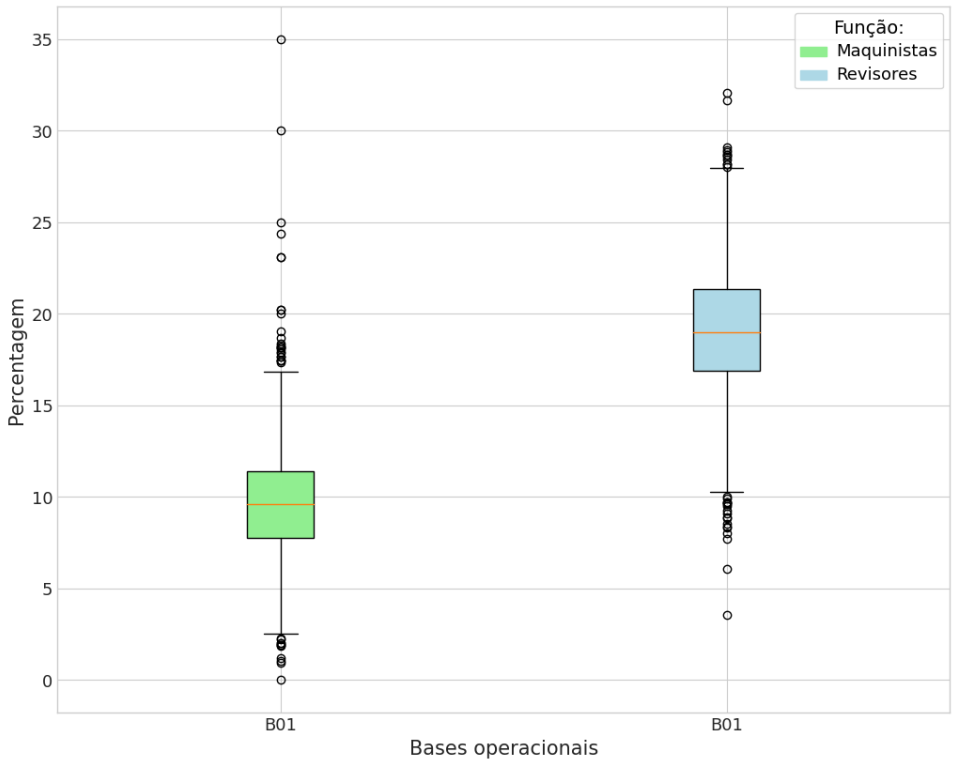

Figura 4.10: Distribuição das séries temporais relativas às percentagens de ausências de maquinistas e dos revisores da base B01.

Como se pode observar, os dados relativos aos revisores apresentam maior mediana e a sua distribuição encontra-se deslocada para valores mais elevados em relação à distribuição dos dados dos maquinistas. As séries apresentam distribuição aproximadamente simétrica (excluindo *outliers*).

Também é possível identificar alguns *outliers* nas duas séries. A remoção dos mesmos pode ter impacto no momento da previsão, por exemplo, *outliers* podem conter informações relevantes sobre eventos incomuns ou anomalias. Ao removê-los, o modelo poderá perder informações importantes sobre os padrões subjacentes na série temporal. Portanto, substituir valores discrepantes sem verificar a sua origem, pode ser uma prática perigosa [24].

Para perceber melhor estes fenómenos, representaram-se os *outliers* localizados no

tempo (com base na amplitude interquartílica[1]) nas Figuras 4.11 e 4.12:

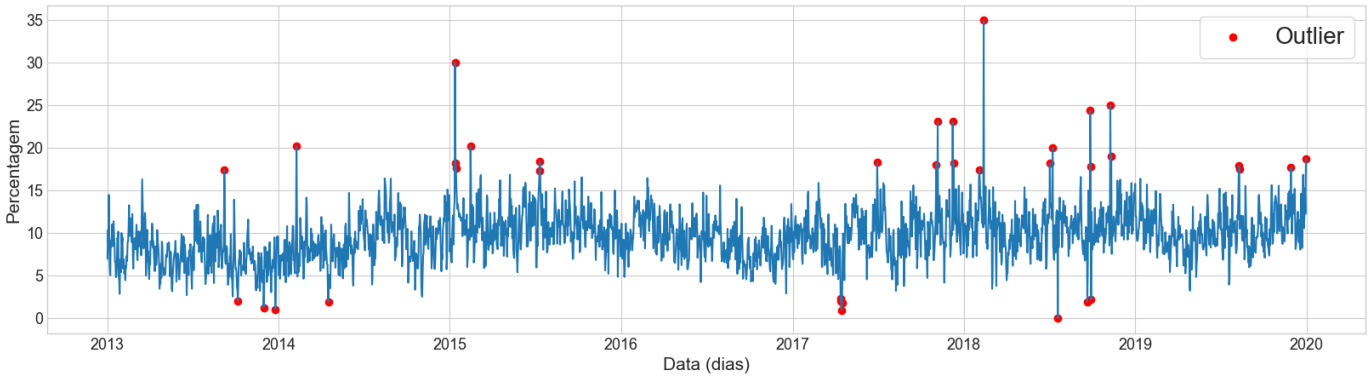

Figura 4.11: *Outliers* (pontos a vermelho) da série temporal relativa à percentagem de ausência diária dos maquinistas da base B01. Os pontos a vermelho são *outliers* identificados pelo critério do IQR.

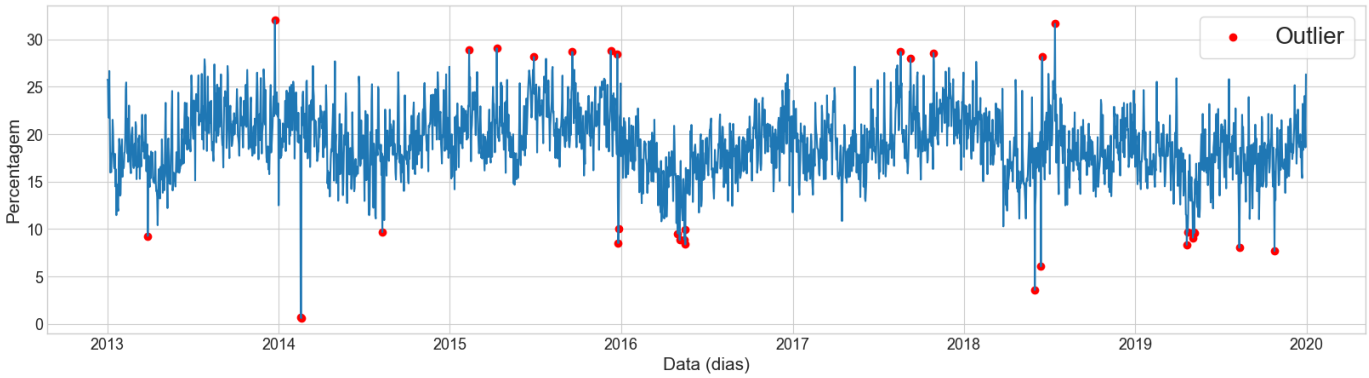

Figura 4.12: *Outliers* (pontos a vermelho) da série temporal relativa à percentagem de ausência diária dos revisores da base B01. Os pontos a vermelho são *outliers* identificados pelo critério do IQR.

Identificaram-se 36 e 32 *outliers* nas séries temporais relativas às percentagens de ausência diária dos maquinistas e revisores, respetivamente. Através das datas correspondentes aos *outliers*, decidiu-se averiguar as possibilidades para a ocorrência destes valores discrepantes, tendo sido concluído que:

- Em 2014, nos dias 17 e 18 de fevereiro (o que correspondem às observações x_{413} e x_{414}) da Figura 4.12, verificou-se que, em comparação com o conjunto inicial de dados da Tabela 4.3, este não contém registos do número de revisores que foram trabalhar nesses dias. Conclui-se que a utilização da interpolação linear introduziu dois *outliers* na série;

- Quanto aos restantes *outliers* das duas séries temporais, verificou-se que existem alguns dias cujo número de tripulantes que foram trabalhar ou se ausentaram

[1]Critério utilizado anteriormente na identificação de *outliers* no *boxplot*.

é relativamente pequeno em comparação com outros dias em que o número de registos é maior. Uma possível explicação pode ser erros do utilizador ao introduzir os dados no sistema ou podem ser acontecimentos simplesmente incomuns. Embora não se tenha a certeza de que esta sejam os principais motivos para o surgimento destas anomalias, é muito provável que estes acontecimentos tenham pelo menos contribuído para o sucedido.

Uma vez que se sabe que foram introduzidos erros, através da interpolação linear, nos dias 17 e 18 de fevereiro de 2014 na série temporal relativa à percentagem de ausências dos revisores, os *outliers* foram substituídos pela mediana móvel de 7 dias anteriores relativamente às datas, uma vez que esta medida de tendência central é mais robusta aos *outliers* do que a média. Portanto, os valores correspondentes a x_{413} e x_{414}, foram substituídos pelo valor 19.61. Relativamente aos restantes valores discrepantes, como não se tem certeza de que estes foram realmente os fatores que contribuíram para o efeito, os *outliers* não foram substituídos por nenhum método.

4.3.1.3 Efeitos dos feriados e eventos especiais

Os feriados nacionais e eventos especiais poderão ter impacto significativo no contexto do absentismo no trabalho. Geralmente, os feriados exibem fortes padrões sazonais nas séries temporais. Por exemplo, como se verificou observando as Figuras 4.4 e 4.5, os meses de janeiro e dezembro apresentam percentagens de absentismo mais elevadas do que os restantes meses. Para além de serem meses frios, também são considerados períodos festivos. Adicionalmente, os trabalhadores podem solicitar pedidos de folgas em datas próximas dos feriados, o que altera, por sua vez, o planeamento das operações.

Assim, decidiu-se incorporar no estudo feriados e eventos especiais como variáveis binárias de forma a analisar e compreender os padrões subjacentes entre o absentismo e os feriados. Para tal, recorreu-se ao *package holidays* [8]. Através da função `country_holidays()` obtiveram-se as datas correspondentes aos eventos festivos do país onde se situa o operador ferroviário, dentro do período entre janeiro de 2013 e dezembro de 2019. Com esta informação, criou-se uma lista com valores 1 e 0, representando um feriado ou não num determinado dia. No final, adicionou-se a lista a cada conjunto de dados e denominou-se a nova coluna por `'holidays'`[2]. Tendo em conta esta informação, procedeu-se à análise dos efeitos dos feriados.

Dado que nem todos os meses possuem feriados, realizou-se uma análise da distribuição dos dados apenas para os meses que contêm feriados.

Através de um *swarmplot*, Figuras 4.13 e 4.14, representaram-se as distribuições das percentagens de absentismo (entre o ano 2013 e 2019) divididas pelo mês e pela variável binária `'holidays'`, de cada série temporal. Ou seja, cada ponto representa um dia, desse mês, com uma percentagem de absentismo associada. As diferentes cores diferenciam

[2]Pode-se verificar todos os atributos que compõe os conjuntos de dados na Tabela I.1 do anexo I.

se esse dia corresponde a um feriado ou não. Os pontos encontram-se organizados de maneira a que não se sobreponham, de forma a destacar como os valores se agrupam dentro de cada categoria.

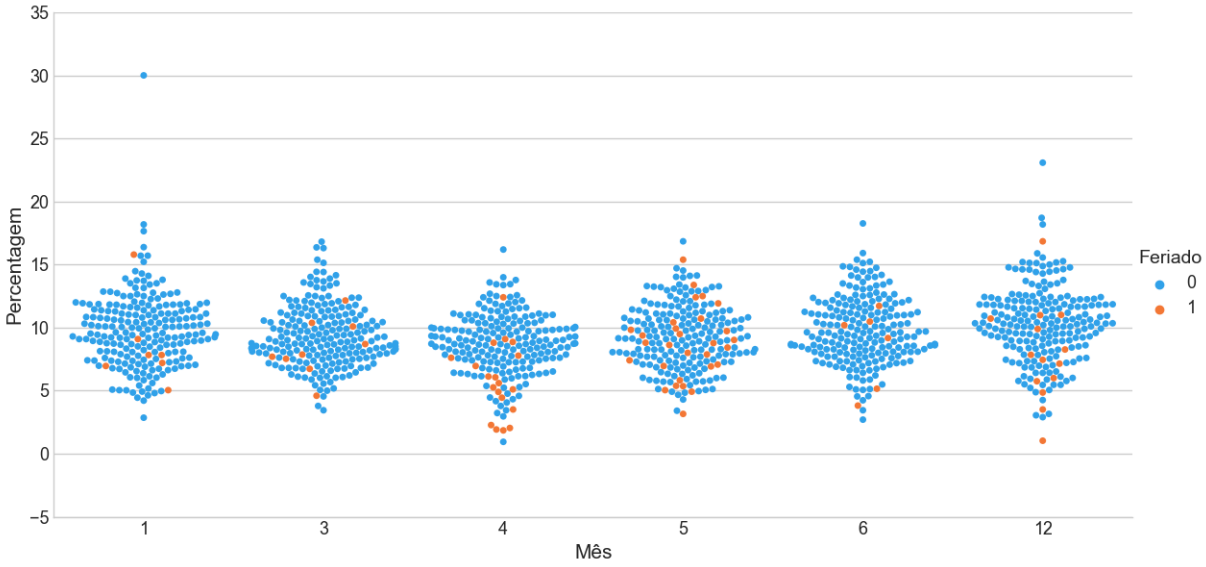

Figura 4.13: Distribuição da série temporal relativa às percentagens de ausências de maquinistas da base B01 de certos meses do período entre 2013 e 2019 (7 anos), divididas em função da observação corresponder a um feriado ou não.

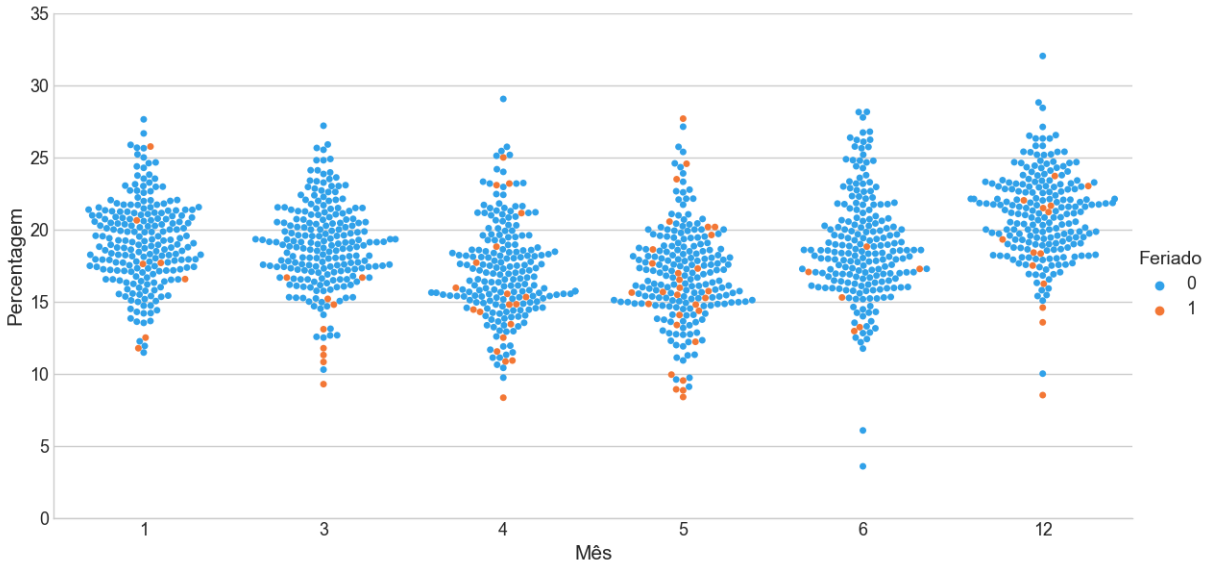

Figura 4.14: Distribuição da série temporal relativa às percentagens de ausências de revisores da base B01 de certos meses do período entre 2013 e 2019 (7 anos), divididas em função da observação corresponder a um feriado ou não.

Pelas Figuras 4.13 e 4.14, pode-se inferir que a densidade da maioria das observações, nos dias de feriados, encontram-se dispersas ao longo dos meses, sugerindo que não há um agrupamento evidente em períodos festivos num determinado mês. No entanto, no

mês 3 da Figura 4.14 é possível visualizar um agrupamento, indicando que, nesse mês, as percentagens de ausências dos revisores são, geralmente, menores nos feriados. Este comportamento seria expectável, uma vez que se sabe que há menos pessoas a trabalhar aos feriados. Consequentemente, existem menos ausências, o que se traduz em percentagens de ausências menores.

Antes da análise dos efeitos dos feriados, esperava-se que existissem mais agrupamentos de observações em dias de feriados. Embora as conclusões não sejam as mais satisfatórias, considerou-se englobar a variável `'holidays'` nas próximas etapas do estudo, pois possui alguma informação temporal para o problema.

4.4 Conclusões da análise exploratória

Dado que se tratam de séries temporais diárias longas que apresentam diversos padrões sazonais, é necessário usar métodos que lidem com essa sazonalidade complexa. Adicionalmente, a complexidade do problema aumenta consideravelmente na presença de sazonalidade não regular, ou seja, quando a sazonalidade se encontra associada a eventos, como a Páscoa, que ocorrem em datas diferentes consoante o ano.

Conforme se descreveu no capítulo 3.5, modelos como o SARIMA apenas são capazes de capturar um tipo de sazonalidade. Apesar do SARIMA conseguir lidar com períodos sazonais até 350, o custo computacional é elevado, resultando em cálculos lentos ou fazendo com que o sistema fique sem memória. Modelos como a regressão dinâmica hamónica, e algoritmos como o Prophet e o NeuralProphet, são adequados quando as séries apresentam vários padrões sazonais, e podem incluir como preditores os eventos/feriados como variáveis *dummy* para capturar a sazonalidade não regular.

Concluída a fase exploratória dos dados, procedeu-se à fase de previsão e análise dos seus resultados.

5

Apresentação e análise de resultados

Neste capítulo são descritos os resultados considerando o método da **janela fixa** — secção 5.1 — e o **método da janela móvel de previsão a *h*-passos** — secção 5.2, de modo a avaliar o desempenho relativo de cada procedimento e, com isso, perceber qual se adequa melhor a cada série temporal.

Os resultados obtidos através da utilização da RDH, do Prophet e do NeuralProphet, são apresentados no capítulo 3.5, às séries relativas às percentagens de ausências diárias dos maquinistas e dos revisores da base B01. A partir dessas previsões, os valores foram convertidos para número total de ausências diárias dos maquinistas e dos revisores, multiplicando pelo número total de trabalhadores, Equação 4.1, planeados para esses dias.

Como referido no capítulo anterior, serão adicionados os feriados (`'holidays'`) como informação externa, uma vez que se concluiu que esta variável possui alguma informação temporal. Nos estudos iniciais, compararam-se os resultados, com e sem a variável `'holidays'`, e concluiu-se, também, que a sua inclusão tem alguma influência na performance dos métodos utilizados. As métricas de desempenho sem a incorporação da variável, para as previsões da janela fixa, são apresentadas na Tabela III.1 do anexo III, para ilustrar o impacto desta variável[1].

Tendo em conta que as séries temporais apresentam mais do que um período sazonal, utilizaram-se séries de *Fourier* para replicar a sazonalidade semanal e anual. Apesar do Prophet e do NeuralProphet já terem incorporado séries de *Fourier* para modelar a sazonalidade, é possível modificar o seu número de termos e o seu período. No modelo RDH as séries de *Fourier* são utilizadas como regressores externos.

Outro aspeto importante a apontar é que os algoritmos Prophet e NeuralProphet, têm muitos hiperparâmetros. Também foi necessário ajustar os hiperparâmetros mais apropriados consoante cada série. Para o modelo RDH não foi necessário, pois recorreu-se à função `auto_arima()` [52].

[1] Também se adicionou na Tabela III.1, do anexo III, as métricas de desempenho com a incorporação da variável `'holidays'`, para facilitar a leitura.

A escolha dos hiperparâmetros e da definição de séries de *Fourier*, encontram-se nas Tabelas II.1 e II.2 do anexo II.

Os erros como o MAE, o RMSE e o MAPE, descritos na secção 3.3, foram seguidamente calculados para cada uma das séries, para efeitos de avaliação e comparação dos resultados produzidos. É importante referir que, como os dados se tratam de percentagens, as séries podem conter zeros. No entanto, existem alternativas para mitigar esse problema. Uma possibilidade é não englobar observações que tenham zero, em vez de incorporá-las no cálculo. Esta opção é válida no caso da série não possuir muitos zeros, sem que afete significativamente a interpretação do MAPE. Medidas como o MAPE são extremamente sensíveis na presença de zeros, pois tendem para infinito ou são indefinidos. Também se realizou uma análise de resíduos resultantes do algoritmo Prophet e modelo de RDH, para verificar se os mesmos capturaram adequadamente todas as informações disponíveis.

É fundamental mencionar que, tal como se observou no capítulo anterior, a componente sazonal anual muda ao longo do tempo de forma semi-multiplicativa, sugerindo que pode ser benéfico utilizar uma transformação *Box Cox* para ambas as séries, de modo a estabilizar a variância. Através da função `boxcox()` do *package Scipy* [60] determinou-se o melhor valor do λ, aplicando, de seguida, a transformação nos dados, antes do ajustamento. No entanto, a reversão nas previsões introduziu algum enviesamento. Este enviesamento resultou na deterioração das previsões, comparativamente com as previsões sem a aplicação da transformação. Logo, decidiu-se não aplicar a transformação nos dados.

De um modo geral, em termos de custo computacional, o Prophet provou ser computacionalmente mais rápido a estimar os parâmetros do modelo, em comparação com o RDH e o NeuralProphet, com um tempo médio de treino de 6.5 segundos. Por outro lado, a RDH mostrou-se ser relativamente lenta em termos computacionais em comparação com o Prophet e o NeuralProphet, com um tempo médio de treino de 12 minutos.

5.1 Avaliação das previsões com método da janela fixa

A avaliação da performance foi feita de forma tradicional, isto é, dividindo os dados em amostra de treino, para estimar os parâmetros, e amostra de teste, para avaliar a precisão. Os dados de cada série têm um total de 2556 observações, e decidiu-se dividir da seguinte forma:

- A **amostra de treino** tem uma dimensão de 2191 observações, o que corresponde ao período entre 1 de janeiro de 2013 e 31 de dezembro de 2018.
- A **amostra de teste** tem uma dimensão de 365 observações, o que corresponde ao período entre 1 de janeiro de 2019 e 31 de dezembro de 2019.

Quanto à presença de zeros na amostra de teste, não se identificaram quaisquer observações com valor de 0%, em ambas as séries. Segue-se a representação dos resultados produzidos por cada método, da amostra de teste (que corresponde ao período entre 1 de janeiro de 2019 e 31 de dezembro de 2019), das séries temporais relativas aos maquinistas e aos revisores da base B01:

5.1.1 Resultados da série relativa aos maquinistas da base B01

Observando os gráficos das Figuras 5.1, 5.2 e 5.3, de uma forma geral, os métodos acompanham relativamente bem a tendência dos valores reais das percentagens de absentismo da série.

Tal como se observou nas Figuras 4.6 e 4.7, existe um padrão repetitivo de picos ao longo da semana, sugerindo sazonalidade semanal. Contudo, esta sazonalidade muda ao longo do tempo. Como foram utilizadas séries de *Fourier* para capturar a sazonalidade, é de esperar que o comportamento da mesma seja fixa.

Existem picos bastante aleatórios, muitos deles descritos no capítulo 4.3.1.2, que o RDH, Prophet e NeuralProphet não conseguiram prever.

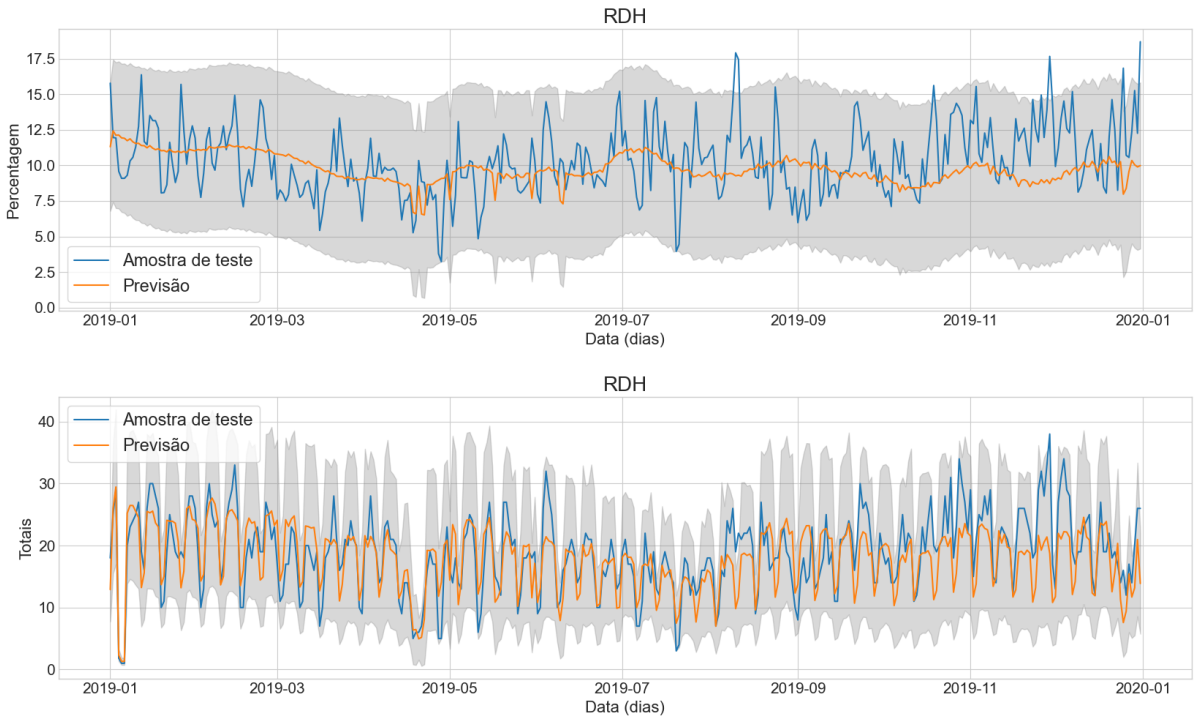

Figura 5.1: Previsões do modelo de RDH das percentagens de absentismo, em cima, e a conversão em número total de ausências dos maquinistas, em baixo, da base B01 com um intervalo de previsão a 95%, representado pela área a cinzento.

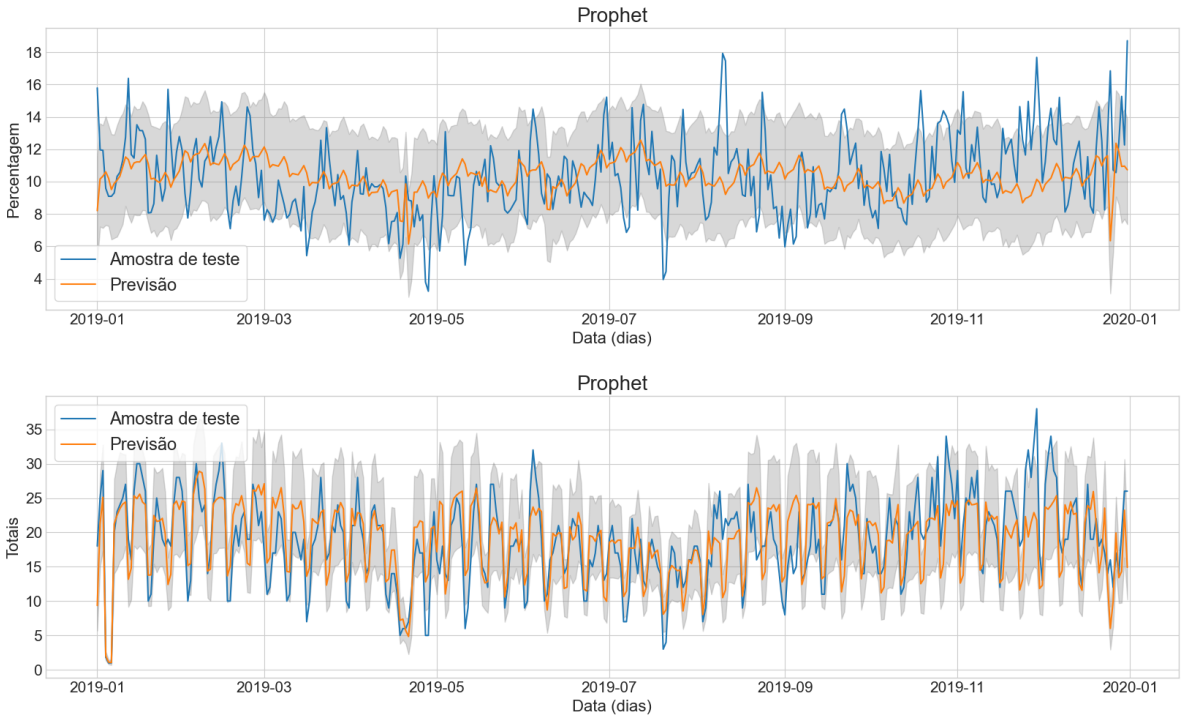

Figura 5.2: Previsões do Prophet das percentagens de absentismo, em cima, e a conversão em número total de ausências dos maquinistas, em baixo, da base B01 com um intervalo de previsão a 95%, representado pela área a cinzento.

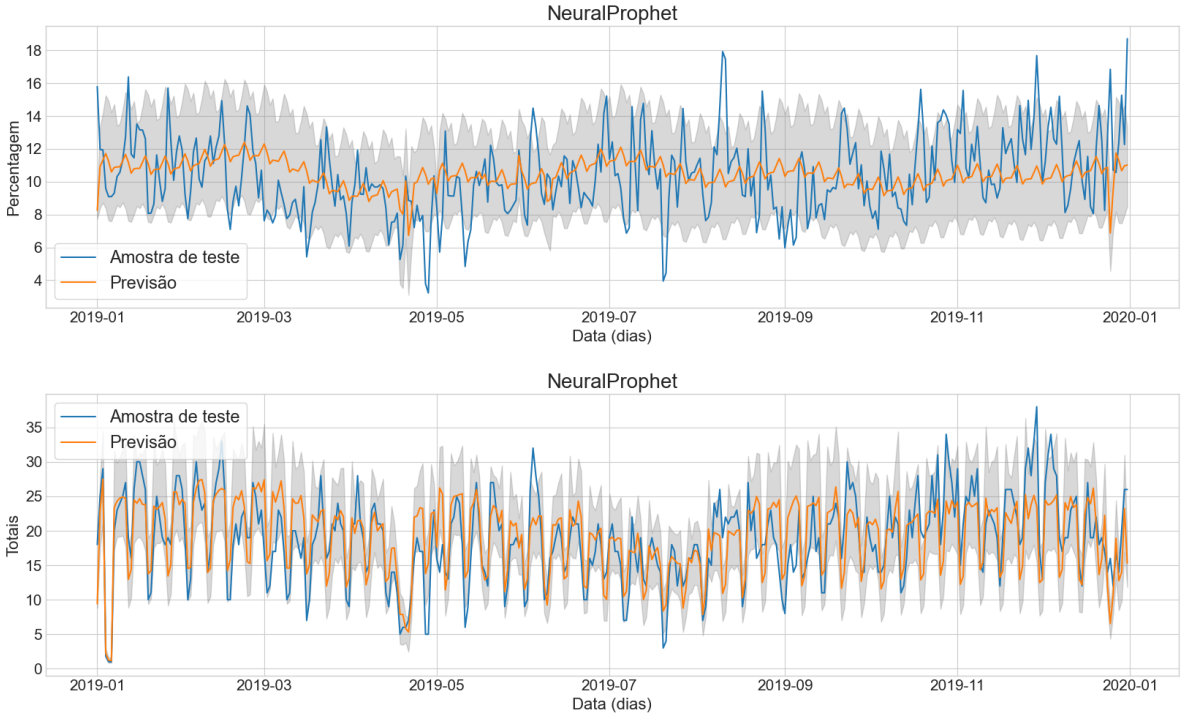

Figura 5.3: Previsões do NeuralProphet das percentagens de absentismo, em cima, e a conversão em número total de ausências dos maquinistas, em baixo, da base B01 com um intervalo de previsão a 95%, representado pela área a cinzento.

Analisando as métricas de desempenho da Tabela 5.1, pode-se afirmar que não se encontram diferenças significativas entre os métodos, o que indica que todos tiverem um desempenho bastante semelhante e, numa forma geral, os resultados são razoáveis.

Tabela 5.1: Resumo das métricas de desempenho — MAE, RMSE e MAPE — das previsões das percentagens de absentismo e do número total de ausências de maquinistas da base B01.

		Métrica		
Tipo de previsão	Método	MAE	RMSE	MAPE (%)
Percentagem	RDH	1.9853	2.5789	19.96
	Prophet	1.9591	2.5369	20.70
	NeuralProphet	1.9653	2.5112	21.09
Totais	RDH	3.3741	4.3282	19.96
	Prophet	3.3728	4.2718	20.70
	NeuralProphet	3.3798	4.2307	21.09

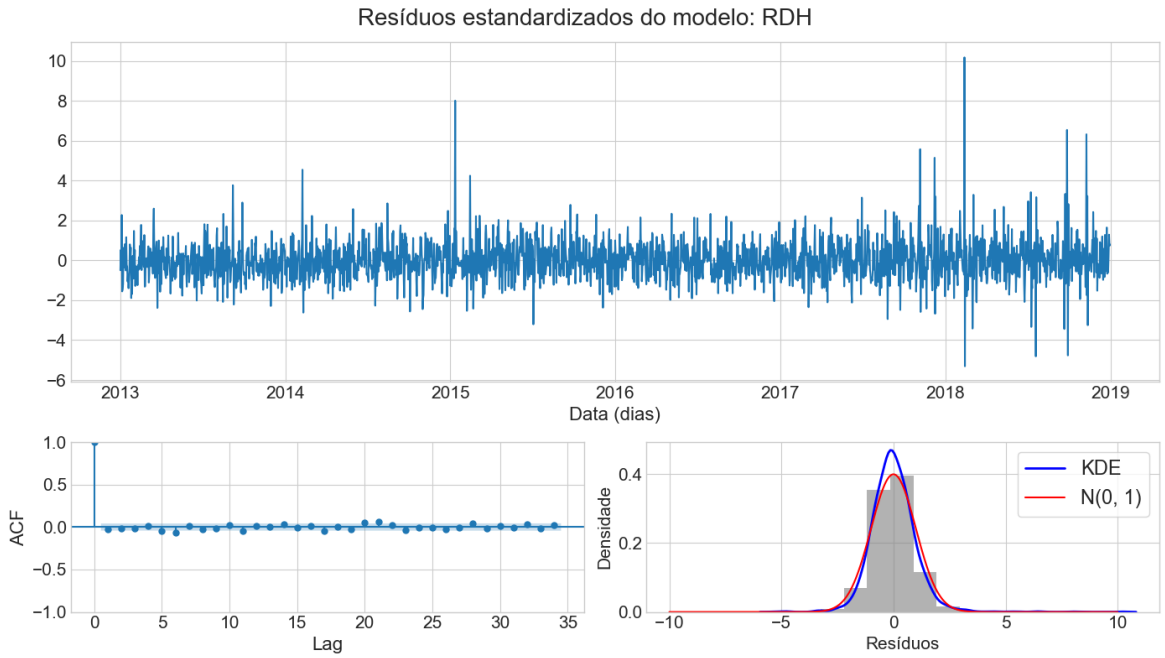

Figura 5.4: Diagnóstico dos resíduos estandardizados do modelo de RDH da série relativa às percentagens de ausências dos maquinistas da base B01. O primeiro gráfico representa os resíduos estandardizados ao longo do tempo. O segundo gráfico, do lado esquerdo, representa o correlograma dos resíduos. O terceiro gráfico, do lado direito, representa a distribuição dos resíduos.

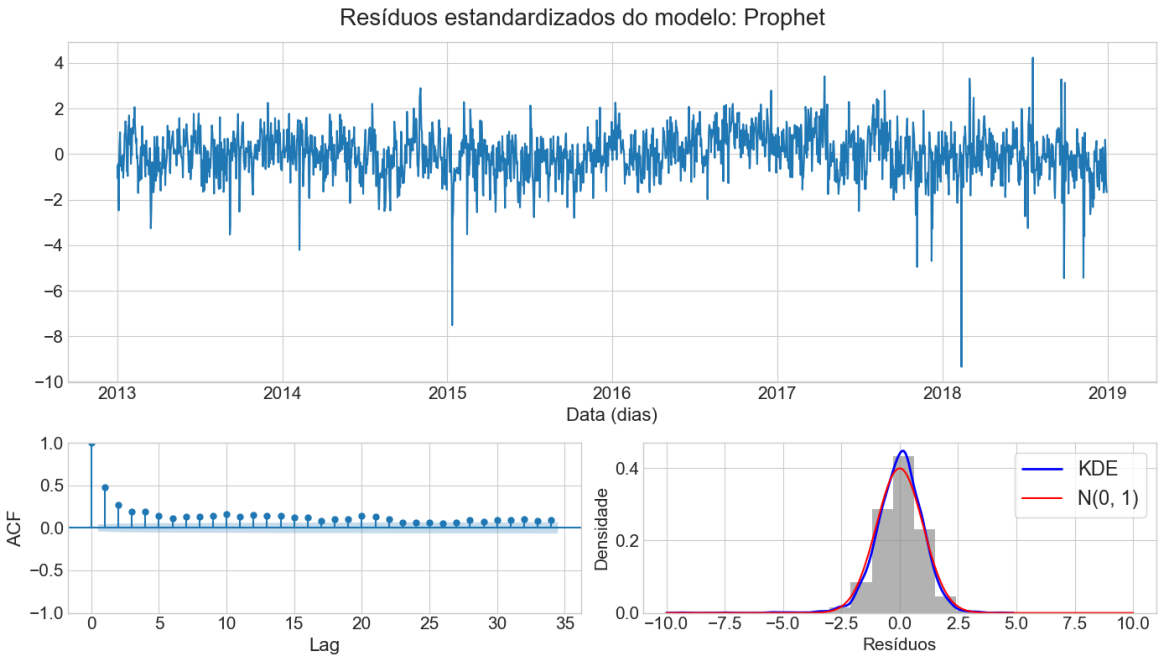

Figura 5.5: Diagnóstico dos resíduos estandardizados do Prophet da série relativa às percentagens de ausências dos maquinistas da base B01. O primeiro gráfico representa os resíduos estandardizados ao longo do tempo. O segundo gráfico, do lado esquerdo, representa o correlograma dos resíduos. O terceiro gráfico, do lado direito, representa a distribuição dos resíduos.

Os pressupostos da RDH e do Prophet são muito parecidos, no que toca à análise de resíduos. A representação dos resíduos produzidos pelo modelo de RDH ao longo do tempo — Figura 5.4 —, mostra que a variação dos resíduos permanece praticamente a mesma entre os dados históricos, com exceção de alguns valores discrepantes; logo, a variância residual pode ser tratada como constante, como explicado em [24]. Quanto aos resultados da ACF dos resíduos, a falta de correlação sugere que as previsões são boas, de acordo com [24]. Por fim, o histograma sugere que os resíduos aproximam-se de uma distribuição normal. Se ignorarmos os *outliers*, a cauda direita não parece tão longa. Consequentemente, os resultados sugerem que as previsões deste modelo podem ser boas, e como os resíduos aproximam-se de uma distribuição normal, infere-se que, provavelmente, os intervalos de previsão podem ser precisos. Por outro lado, a representação dos resíduos produzidos pelo Prophet ao longo do tempo — Figura 5.5 —, mostra que os resíduos não são estáveis, existe um ligeira tendência decrescente entre o ano 2017 e o final de 2018, o que indica que a média dos resíduos não é constante. Quanto aos resultados da ACF, pode-se verificar uma autocorrelação significativa durante muitos *lags*. Os histogramas sugerem, também, que os resíduos aproximam-se de uma distribuição normal. Como existe autocorrelação significativa nos resíduos, as previsões pontuais produzidas podem não ser tão boas como o modelo de RDH. Contudo, isto não significa que o modelo de RDH não possa ser melhorado.

Apesar do modelo de RDH não obter os melhores valores para as métricas MAE e

RMSE, este possui o MAPE mais baixo. Por outro lado, o Prophet é melhor do que o NeuralProphet, no que toca ao valor do MAE, contudo, é pior no valor do RMSE. Portanto, não é claro qual o método que se adequa melhor aos dados com base nas métricas. O que foi descrito, também se verifica sem a inclusão da variável `'holidays'` na Tabela III.1. Comparando os valores da Tabela 5.2 com a Tabela III.1, pode-se verificar que com a inclusão dos feriados, a performance é ligeiramente pior.

Tendo em consideração as métricas serem semelhantes, mesmo tendo em conta a análise de resíduos, o Prophet (sem a variável `'holidays'`) seria mais vantajoso na prática, uma vez que é muito mais rápido a estimar os parâmetros, e é mais fácil para o utilizador implementar.

5.1.2 Resultados da série relativa aos revisores da base B01

A análise é semelhante à àquela que foi feita na secção 5.1.1 no que toca às representações gráficas das estimativas das percentagens de absentismo e dos seus valores reais (consequentemente, o número de ausências). Porém, existem alguns pontos diferentes a mencionar.

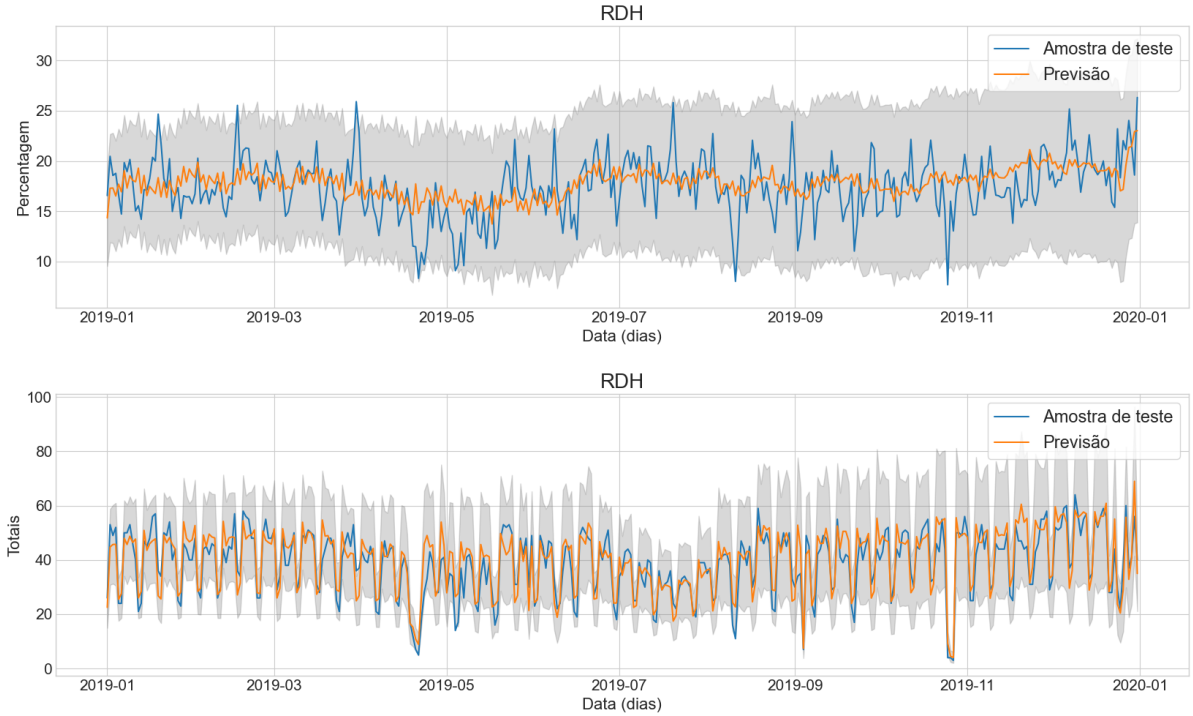

Figura 5.6: Previsões do modelo de RDH das percentagens de absentismo, em cima, e a conversão em número total de ausências dos revisores, em baixo, da base B01 com um intervalo de previsão a 95%, representado pela área a cinzento.

Figura 5.7: Previsões do Prophet das percentagens de absentismo, em cima, e a conversão em número total de ausências dos revisores, em baixo, da base B01 com um intervalo de previsão a 95%, representado pela área a cinzento.

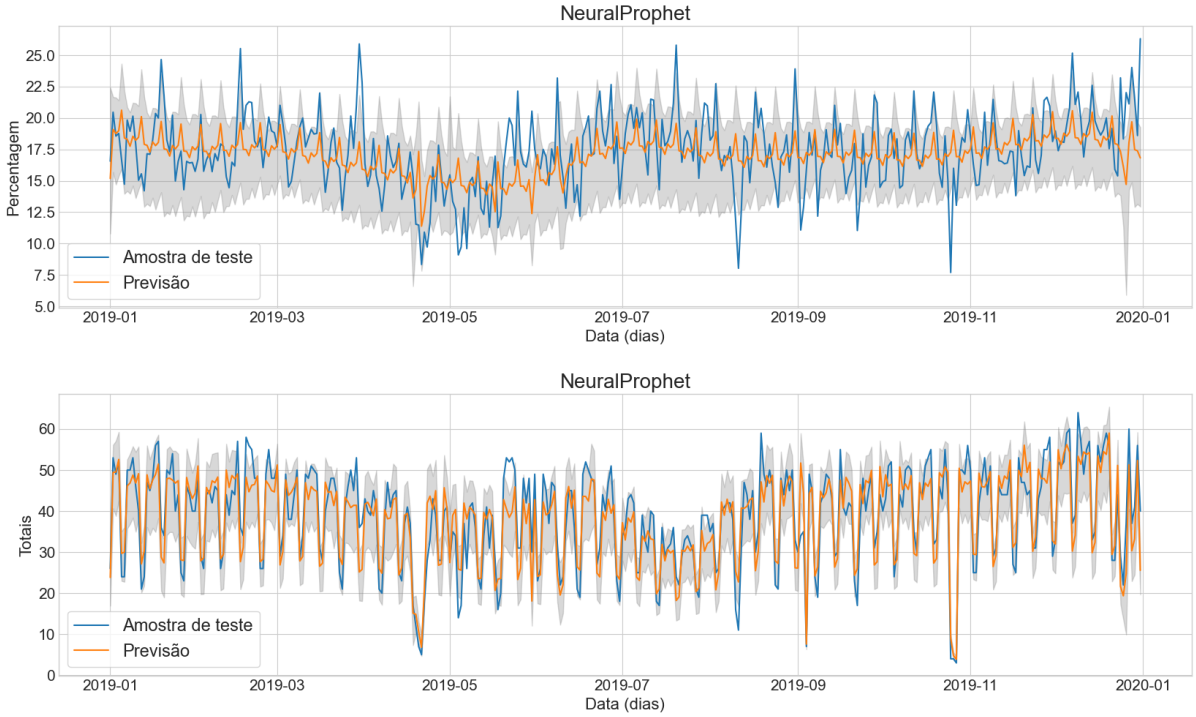

Figura 5.8: Previsões do NeuralProphet das percentagens de absentismo, em cima, e a conversão em número total de ausências dos revisores, em baixo, da base B01 com um intervalo de previsão a 95%, representado pela área a cinzento.

É evidente que a inclusão dos feriados também ajudou a capturar sazonalidade não regular. Por exemplo, é bastante claro um pico no final do mês de abril na série relativa à percentagens de absentismo dos revisores, o que corresponde ao período da Páscoa. O Prophet e o NeuralProphet conseguiram "acompanhar" claramente esse pico. Apesar de não ser tão evidente no modelo de RDH, esta variável também ajudou na performance desse modelo. Comparando os valores da Tabela 5.2 com a Tabela III.1, pode-se verificar que a não inclusão dos feriados, resultou numa performance pior.

Analisando as métricas de desempenho individualmente, pode-se afirmar, também, que não se encontram diferenças significativas, o que também indica que todos tiverem um desempenho bastante semelhante. No entanto, pode-se perceber pela Tabela 5.2, que o NeuralProphet, em geral, apresenta um MAE, um RMSE e um MAPE mais baixos que os restantes modelos.

Relativamente à comparação dos resultados entre as duas séries temporais, a diferença entre os seus valores do MAPE, do NeuralProphet, $21.09 - 12.86 = 8.23\%$, é relativamente grande. Isto indica que o mesmo está a produzir previsões mais precisas, com um menor erro absoluto percentual médio, na série temporal dos revisores.

Tabela 5.2: Resumo das métricas de desempenho — MAE, RMSE e MAPE — das previsões das percentagens e número total de ausências de revisores da base B01.

		Métrica		
Tipo de previsão	Método	MAE	RMSE	MAPE (%)
Percentagem	RDH	2.1460	2.7443	13.68
	Prophet	2.3115	2.9140	13.55
	NeuralProphet	2.0947	2.6675	12.86
Totais	RDH	4.5161	5.6775	13.68
	Prophet	4.8124	5.9001	13.55
	NeuralProphet	4.3069	5.2770	12.86

Quanto à análise dos resíduos, as conclusões são análogas à descrição anterior. A única conclusão diferente é referente à análise dos resíduos do Prophet. No diagnóstico anterior, verificou-se que os resíduos não eram estáveis ao longo do tempo, todavia, esse comportamento não se verifica na série relativa aos revisores — Figura 5.10. Apesar desta diferença, as previsões produzidas podem não ser tão boas como o modelo de RDH, uma vez que apresenta autocorrelação durante alguns *lags*.

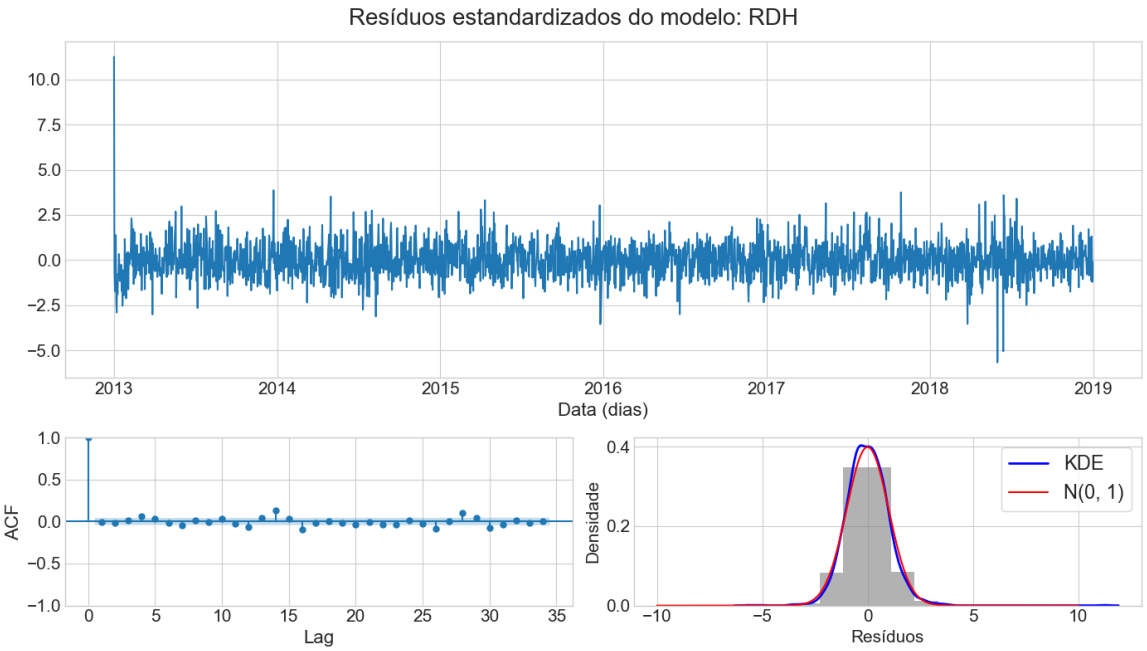

Figura 5.9: Diagnóstico dos resíduos estandardizados do modelo de RDH da série relativa às percentagens de ausências dos revisores da base B01. O primeiro gráfico representa os resíduos estandardizados ao longo do tempo. O segundo gráfico, do lado esquerda, representa o correlograma dos resíduos. O terceiro gráfico, do lado direito, representa a distribuição dos resíduos.

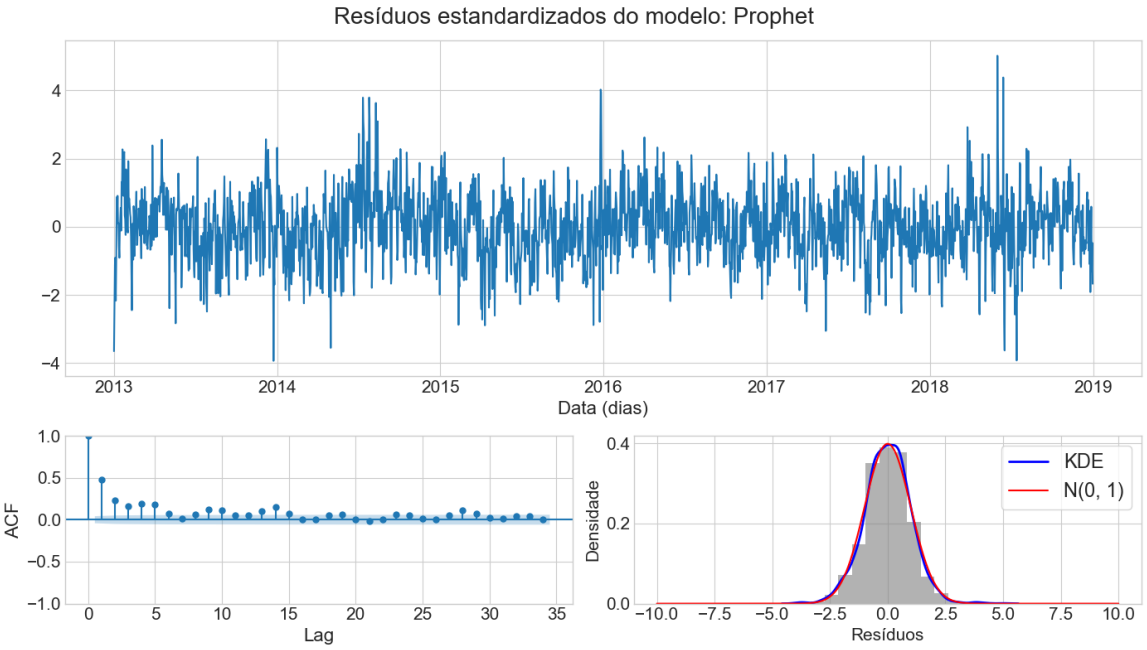

Figura 5.10: Diagnóstico dos resíduos estandardizados do Prophet da série relativa às percentagens de ausências dos revisores da base B01. O primeiro gráfico representa os resíduos estandardizados ao longo do tempo. O segundo gráfico, do lado esquerda, representa o correlograma dos resíduos. O terceiro gráfico, do lado direito, representa a distribuição dos resíduos.

Como o NeuralProphet (com a variável `'holidays'`) apresenta valores menores, em termos de métricas de desempenho, pode-se afirmar que o NeuralProphet adequa-se melhor aos dados.

5.2 Avaliação das previsões com o método da janela móvel de previsão a *h*-passos

Por fim, utilizando os métodos ajustados anteriormente, avaliou-se a sua performance através de *backtesting*, nas seguintes circunstâncias:

- Na série relativa aos maquinistas não se incluiu a variável `'holidays'`;
- Na série relativa aos revisores incluiu-se a variável `'holidays'`.

Efetuaram-se previsões a *h*-passos recorrendo ao método da janela móvel como descrito na secção 3.3.3. Para este estudo, decidiu-se efetuar previsões a 1, 7, 14, 30, 60, 90, 180 e 365 passos ao longo de um período de 365 dias. Uma vez que os dados respetivos ao ano 2020 não são os mais representativos de um período considerado "normal", realizou-se a separação da seguinte maneira:

- Cada **amostra de treino** tem uma dimensão de 1826 observações. A amostra da primeira iteração corresponde ao período entre 1 de janeiro de 2013 e 31 de dezembro de 2017, e o seu tamanho mantém-se ao longo do treino. Ou seja, mantém-se o número de observações, mas não são sempre as mesmas. Vai-se incluindo observações novas (do conjunto de teste) e removendo as mais antigas do conjunto de treino.
- Cada **amostra de teste** tem uma dimensão de 1 observação em cada iteração. Tendo em conta que se pretende fazer previsão a *h*-passos, segue-se a representação do período de cada subconjunto de teste ao longo de 365 iterações:

Tabela 5.3: Período de cada subconjunto de teste para diferentes horizontes de previsão.

h	Subconjunto de teste
1	[1/1/2018,...,31/12/2018]
7	[7/1/2018,...,6/1/2019]
14	[14/1/2018,...,13/1/2019]
30	[30/1/2018,...,29/1/2019]
60	[1/3/2018,...,28/2/2019]
90	[31/3/2018,...,30/3/2019]
180	[29/6/2018,...,28/6/2019]
365	[31/1/2018,...,30/12/2019]

Como vão ser treinados 365 vezes, para cada série temporal, a análise dos resíduos não faz muito sentido ser realizada. Como tal, apenas se vai verificar a capacidade de

generalização dos métodos através das métricas de desempenho ao longo de diferentes horizontes de previsão.

Quanto à presença de zeros nas séries, identificou-se uma observação igual a 0% em alguns dos subconjuntos de teste na série relativa aos maquinistas. Como tal, essas observações não foram contabilizadas no cálculo do MAPE desses subconjuntos de teste.

Para efeitos de simplificação, as próximas secções apenas irão apresentar os resultados para o número total de ausências dos maquinistas e dos revisores da base B01. Os resultados das percentagens de absentismo dos maquinistas e dos revisores da base B01 encontram-se nas Tabelas IV.1 e IV.2 no anexo IV.

Segue-se a apresentação dos resultados produzidos, de cada amostra de teste, da série temporal relativa aos maquinistas e aos revisores:

5.2.1 Resultados da série relativa aos maquinistas da base B01

Pela Figura 5.11, pode-se observar que os erros de previsão, como o MAE e o RMSE, têm o comportamento esperado, ou seja, de um modo geral, os seus erros aumentam à medida que o horizonte de previsão aumenta. No entanto, esse comportamento não se verifica no caso do MAPE. Para $h = 365$, o valor do MAPE é relativamente pequeno em comparação com os restantes h-passos. Isto poderá indicar que a escolha dos hiperparâmetros e do número de séries de *Fourier* pode não ter sido a melhor, uma vez que os métodos não generalizam ou não capturam os padrões intrínsecos da série para os diferentes períodos. Por exemplo, $h = 365$ encontra-se dentro do período da amostra de teste para a janela fixa da secção 5.1.1, como os hiperparâmetros dos métodos são os mesmos, isto pode significar que a escolha é apenas adequada para esse subconjunto. Por outro lado, este comportamento também se pode dever à natureza da série, por exemplo flutuações imprevisíveis do número real de ausências podem afetar os resultados de forma significativa.

Tabela 5.4: Métricas de desempenho da RDH, do Prophet e do NeuralProphet do número total de ausências dos maquinistas da base B01, para diferentes horizontes temporais.

		h-passos							
Método	Métrica	1	7	14	30	60	90	180	365
RDH	MAE	2.4904	2.9419	3.0632	3.1755	3.1863	3.2796	3.2329	3.4203
	RMSE	3.402	3.8863	4.0173	4.1272	4.1273	4.2086	4.1757	4.3463
	MAPE (%)	20.07	21.93	22.39	22.29	22.24	21.49	21.5	19.8
Prophet	MAE	2.6479	2.7585	2.7632	2.8235	2.9051	3.1606	3.3452	3.2702
	RMSE	3.4624	3.619	3.6321	3.6929	3.8007	4.0821	4.275	4.1379
	MAPE (%)	20.71	21.25	21.11	21.18	21.38	21.39	22.61	19.26
NeuralProphet	MAE	2.6021	2.6485	2.6845	2.8062	2.8411	3.0572	3.2655	3.2557
	RMSE	3.3956	3.4702	3.5233	3.6406	3.6653	3.8933	4.1423	4.0908
	MAPE (%)	20.25	20.56	20.57	20.95	21.05	20.83	22.2	19.17

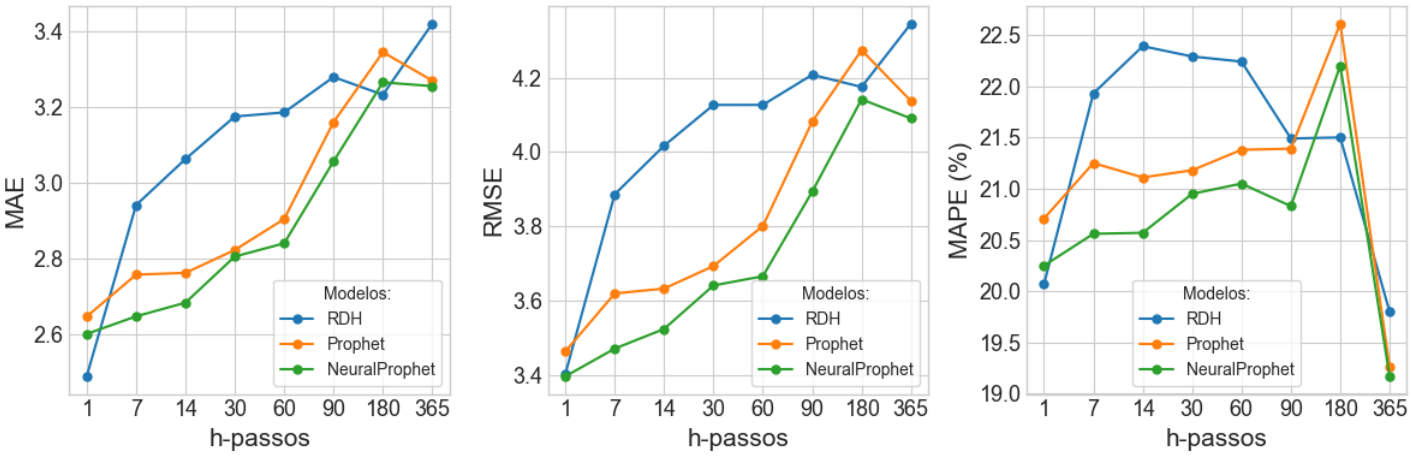

Figura 5.11: Métricas de desempenho — MAE, RMSE e MAPE — em função do horizonte de previsão para os diferentes modelos do número total de ausências dos maquinistas da base B01.

5.2.2 Resultados da série relativa aos revisores da base B01

Ao contrário dos resultados da secção 5.2.2, pode-se observar pela Figura 5.12 que os erros de previsão aumentam à medida que o horizonte de previsão aumenta, como seria de esperar.

Na secção 5.1.2 conclui-se que o NeuralProphet se adequava melhor através das métricas de desempenho. Porém, com a janela móvel, obteve-se uma performance pior quando comparada com a do modelo de RDH. Por exemplo, comparando os valores do MAPE para $h = 365$, o NeuralProphet obteve um MAPE de 24.04% e a RDH obteve um MAPE de 16.55% , sendo que a diferença entre os seus valores é bastante significativa. Esta degradação dos resultados pode-se dever ao número de observações na amostra de treino. Neste caso, os métodos foram ajustados com menos observações. Uma das desvantagens do NeuralProphet é que requer muitos dados, na fase de treino, para conseguir modelar padrões complexos. O Prophet também tem essa desvantagem, e, provavelmente, pode ser essa a razão para os resultados também sofrerem uma degradação.

Para além disso, tinha-se verificado pela Tabela 5.2 que os resultados tinham um desempenho muito semelhante, embora o NeuralProphet produzisse resultados ligeiramente melhores.

Tendo em consideração estes novos resultados, o modelo de RDH poderá ser o mais indicado para esta série.

Tabela 5.5: Métricas de desempenho da RDH, do Prophet e do NeuralProphet do número total de ausências dos revisores da base B01, para diferentes horizontes temporais.

		h-passos							
Método	Métrica	1	7	14	30	60	90	180	365
RDH	MAE	3.7163	4.1889	4.4051	4.7097	4.9625	5.1327	5.3987	5.4307
	RMSE	4.7654	5.2664	5.4915	5.823	6.1069	6.4186	6.6815	7.0082
	MAPE (%)	12.77	14.3	14.73	15.54	16.37	16.71	15.74	16.55
Prophet	MAE	4.0794	4.2813	4.4823	4.7926	5.192	5.5671	6.7241	7.8528
	RMSE	5.1486	5.3791	5.5736	5.9204	6.4199	6.849	8.2945	9.8451
	MAPE (%)	14.41	15.01	15.54	16.45	17.67	18.55	19.46	22.94
NeuralProphet	MAE	4.2669	4.4546	4.6961	5.0814	5.4897	5.9563	6.9181	8.0896
	RMSE	5.3966	5.6319	5.8999	6.3239	6.7828	7.2963	8.491	10.0735
	MAPE (%)	15.08	15.71	16.43	17.57	18.65	19.8	20.08	24.03

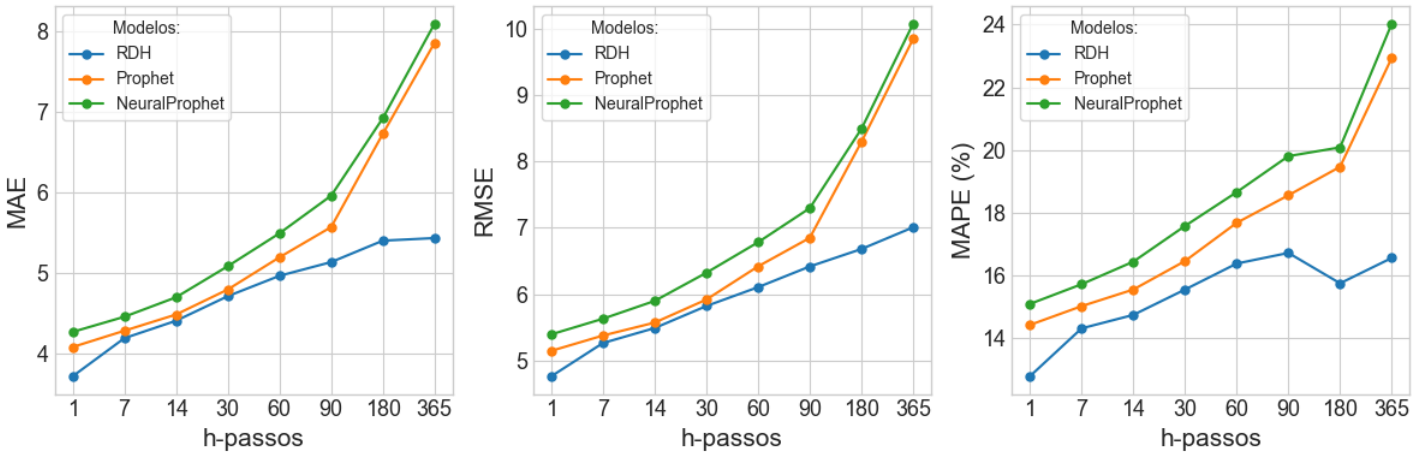

Figura 5.12: Métricas de desempenho — MAE, RMSE e MAPE — em função do horizonte de previsão para os diferentes modelos do número total de ausências dos revisores da base B01.

6

Conclusão e trabalho futuro

Este projeto teve como principal tema o desenvolvimento de um modelo preditivo de ausências não programáveis, capaz de estimar o número de maquinistas e revisores ausentes por doença e/ou apoio à família, para cada base operacional de um operador ferroviário norte da Europa, com o propósito de obter um modelo que conseguisse prever, com o menor erro possível, essas estimativas.

Em suma, foram apresentados métodos de previsão dos números de maquinistas e de revisores ausentes para a base B01. Mediante os resultados obtidos do capítulo 5, e face aos objetivos da Dissertação, conclui-se que, na secção 5.1, apesar das métricas de desempenho terem produzido resultados razoáveis e semelhantes, não se conseguiu perceber, com base nas métricas, que método se adequa melhor aos dados da série relativa aos maquinistas da base B01. Porém, em termos de custo computacional e implementação, o Prophet pode ser o mais indicado para esta série, obtendo um MAE de 3.3444, um RMSE de 4.2420 e um MAPE de 20.53%, sem a inclusão dos feriados, quando a série é convertida para número total de ausências. Apesar da análise dos resíduos indicar que as estimativas das previsões pontuais possam não ser as melhores, esta análise não deve ser utilizada para selecionar métodos de previsão. Para além disso, a análise também indica que, o seu ajustamento pode ser melhorado com outros hiperparâmetros. Relativamente à série dos revisores da base B01, conseguiu-se concluir que o NeuralProphet apresentava resultados ligeiramente melhores comparativamente com os resultados produzidos pelos outros métodos, obtendo um MAE de 4.3069, um RMSE de 5.2770 e um MAPE de 12.86%, quando a série convertida para número total de ausências.

Outra conclusão obtida através da comparação dos valores do MAPE das duas séries temporais, é que, de uma forma geral, os métodos estão a produzir previsões mais precisas na série dos revisores do que nos maquinistas. Por exemplo, comparando o Prophet ajustado à série dos maquinistas com o Neuralprophet ajustado à série dos revisores, a diferença entre os valores, $20.53 - 12.86 = 7.67\%$, é significativa.

Relativamente aos resultados da janela móvel, estes mostraram que os resultados do MAPE para $h = 365$ é bastante pequeno em comparação com os horizontes de previsão menos distantes. Isto poderá indicar que, provavelmente, a escolha dos hiperparâmetros e

definição das séries de *Fourier* podem ter prejudicado a capacidade de generalização dos métodos para diferentes subconjuntos da série, ou as flutuações imprevisíveis do número real de ausências podem ter afetado os resultados de forma significativa, relativamente à série dos maquinistas. Por outro lado, nos dados relativos aos revisores, o modelo de RDH obteve um desempenho bom ao longo dos diferentes horizontes de previsão em comparação com o NeuralProphet. Portanto, o modelo de RDH poderá ser o mais indicado para esta série, tendo em conta a sua capacidade de generalização.

Considerando esta análise, existem alguns pontos a destacar:

- A inclusão ou não da variável `'holidays'` não teve um impacto significativo;
- De um modo geral, os modelos obtiveram um desempenho semelhante.
- O MAPE é sensível na presença de valores próximos de zero. Tendo em conta que o problema se trata de percentagens e de número de ausências, o MAPE não é o mais indicado;
- Não foi possível verificar, verdadeiramente, a capacidade de generalização dos hiperparâmetros escolhidos com o método da janela móvel de previsão a h-passos, uma vez que não se utilizaram dados para além de 2019.

Apesar dos métodos utilizados apresentarem resultados satisfatórios, não se exclui a hipótese de que, com a utilização de outros algoritmos ou modelos, não se pudesse produzir resultados melhores.

Como referido no capítulo 1, esta Dissertação surge da necessidade da SISCOG melhorar a capacidade de prever ausências em relação a um estudo inicial, que resultou numa publicação [39]. Nesse estudo, os autores utilizaram algoritmos baseados em árvores de decisão, como o *Random Forests* e o *XGBoost*, e a regressão logística. De entre os algoritmos e modelos analisados, a regressão logística obteve melhores resultados, com uma precisão a rondar, em média, os 87%. Apesar destes resultados serem promissores, este estudo foi feito com base nas informações individualizadas das ausências dos maquinistas, de todas as bases operacionais do operador ferroviário. No final, as previsões foram agregadas para obter o número total de maquinistas ausentes. Portanto, estes resultados não podem ser comparáveis com o estudo efetuado nesta Dissertação. Apesar de não ser comparável, esta Dissertação revelou análises e resultados interessantes ao estudar o absentismo sob a forma de séries temporais. Por exemplo, relativamente à série dos maquinistas da base B01, para um horizonte de previsão a 7 dias, o valor médio do MAPE ronda os 21.25%, enquanto que o o valor do MAPE da regressão logística, para o mesmo horizonte de previsão, é de 25.14%. Para um horizonte de previsão a 30 dias , o valor médio do MAPE ronda os 21.47%, enquanto que o o valor do MAPE da regressão logística, para o mesmo horizonte de previsão, é de 25.03%.

6.1 Limitações e trabalho futuro

Quanto aos resultados obtidos é de salientar alguns aspetos que limitaram o processo, e possíveis recomendações para melhorar a performance dos métodos utilizados:

Ajuste de hiperparâmetros:

O desempenho do Prophet e do NeuralProphet, dependem da escolha adequada dos seus hiperparâmetros. Descobrir esses valores, que resultem num melhor desempenho e capacidade de generalização, não é uma tarefa simples. Uma estratégia é testar todas as combinações possíveis dos valores dos parâmetros de interesse [35]. Portanto, é recomendável utilizar métodos de otimização de hiperparâmetros.

Contudo, é um processo computacionalmente pesado, e por essa razão optou-se por selecionar manualmente os hiperparâmetros.

Investigação aprofundada dos *outliers*:

Apesar de se ter realizado uma análise dos *outliers* das séries e de se ter identificado algumas das causas para tais ocorrências, é necessário efetuar uma investigação mais aprofundada.

Como descrito no capítulo 4.3.1.2, alguns valores discrepantes podem-se dever a erros do utilizador ou podem ser acontecimentos simplesmente incomuns. Sendo incertos os principais motivos para os valores discrepantes, decidiu-se não substituir os *outliers* por nenhum método.

Estes valores discrepantes podem introduzir erro nos métodos, caso sejam erros genuínos. Consequentemente, estes afetam a precisão das estimativas. Por isso, é prudente averiguar se foram essas as causas para os *outliers* ou se os mesmos fazem realmente parte dos registos dos dados históricos.

Variáveis exógenas:

Explorar outras variáveis externas que possam influenciar o absentismo pode melhorar a performance dos métodos. No entanto, é importante que a seleção das variáveis seja feita com alguma prudência, pois nem todas as variáveis são relevantes, e isso pode afetar a precisão das estimativas.

No início deste estudo, exploraram-se variáveis relacionadas com os fatores meteorológicos. No entanto, acabou-se por não incluir essas variáveis nos métodos, pois, para além de serem muito incertos, num cenário realista, não se tem os valores reais dos fatores meteorológicos para horizontes temporais muito distantes. Por exemplo, se o utilizador quisesse prever o número de ausências do ano seguinte, não teria os valores reais para introduzir.

Um preditor que poderia ser útil seria os dias de semana como variáveis *dummy*, tal como sugerido em [24]. Para representar cada dia da semana, geravam-se 6 variáveis, cada

uma correspondente ao dia da semana (segunda, terça,...). Para uma determinada data, a variável assume valor 1 para o dia da semana correspondente e 0 para os restantes dias da semana. No caso do dia corresponder a um domingo, representava-se a linha por zeros.

Metodologia aplicada:

Cada série temporal possui características únicas. Como resultado, a metodologia aplicada para analisar e prever as séries temporais estudadas, nesta Dissertação, pode não ser adequada para as séries das outras bases operacionais que não foram selecionadas. A análise e a previsão de cada série exige uma abordagem diligente, levando em consideração as suas propriedades para garantir resultados precisos e interpretáveis.

Depois de desenvolver todos os métodos adequados para cada série temporal, seria interessante criar uma funcionalidade que permitisse aos planeadores inserir uma data ou um horizonte temporal que se pretende prever, de uma determinada base e tipo de função desempenhada, e o número ou uma lista de tripulantes planeados para trabalhar, nesse período. No final, a funcionalidade devolvia o número ou uma lista do número de pessoas ausentes dentro desse período.

Bibliografia

[1] S. A. Ali Shah et al. «An enhanced deep neural network for predicting workplace absenteeism». Em: *Complexity* 2020 () (ver pp. 11, 12).

[2] V. Araujo et al. «A hybrid approach of intelligent systems to help predict absenteeism at work in companies». Em: *SN Applied Sciences* 1 (2019-06). DOI: 10.1007/s42452-019-0536-y (ver p. 12).

[3] A. Asiri e M. Abdullah. «Employees absenteeism factors based on data analysis and classification». Em: *Special Issue in Communication and Information Technology* 12.1 (2019), pp. 119–127 (ver p. 11).

[4] K. Bandara, R. J. Hyndman e C. Bergmeir. «MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns». Em: *arXiv preprint arXiv:2107.13462* (2021) (ver p. 23).

[5] C. Boot et al. «Prediction of long-term and frequent sickness absence using company data». Em: *Occupational Medicine* 67.3 (2017), pp. 176–181 (ver pp. 13, 14).

[6] J. Brownlee. *How to Choose a Feature Selection Method For Machine Learning*. 2019-08. URL: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ (acedido em 2023-02-07) (ver p. xvi).

[7] R. B. Cleveland et al. «STL: A seasonal-trend decomposition procedure based on loess». Em: *Journal of Official Statistics* 6.1 (1990), pp. 3–33. URL: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf (ver p. 23).

[8] P. H. P. Contributors. *Python Holidays*. Accessed on Date September 5, 2023. 2023. URL: https://python-holidays.readthedocs.io/en/latest/ (ver p. 59).

[9] W. Contributors. *Feature (machine learning)*. 2022-08. URL: https://en.wikipedia.org/wiki/Feature_(machine_learning) (acedido em 2023-02-07) (ver p. xvi).

[10] W. Contributors. *Neuro-fuzzy*. 2023-01. URL: https://en.wikipedia.org/wiki/Neuro-fuzzy (acedido em 2023-02-08) (ver p. xvi).

[11] W. Contributors. *Unit root test — Wikipedia, The Free Encyclopedia*. [Online; accessed 14-8-2023]. 2022. URL: https://en.wikipedia.org/wiki/Unit_root_test (ver pp. 26, 27).

[12] I. Corp. *What is a neural network?* URL: https://www.ibm.com/topics/neural-networks (acedido em 2023-02-07) (ver p. xvi).

[13] databricks. *What is a DataFrame?* URL: https://www.databricks.com/glossary/what-are-dataframes (acedido em 2023-05-28) (ver p. xvi).

[14] D. A. Dickey e W. A. Fuller. «Distribution of the estimators for autoregressive time series with a unit root». Em: *Journal of the American statistical association* 74.366a (1979), pp. 427–431 (ver pp. 27, 28).

[15] Dominion. *THE COST OF ABSENTEEISM AND HOW TO CONTROL IT*. URL: https://www.dominionsystems.com/blog/managing-the-cost-of-absenteeism (acedido em 2022-11-17) (ver p. 4).

[16] Facebook. *Prophet: Automatic Forecasting Procedure*. URL: https://github.com/facebook/prophet (acedido em 2023-07-13) (ver p. 42).

[17] R. P. Ferreira et al. «Artificial neural network and their application in the prediction of absenteeism at work». Em: *International Journal of Recent Scientific Research* 9.1 (2018), pp. 2332–2334. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0901.1447 (ver pp. 12, 13).

[18] R. Gandhi. *Support Vector Machine — Introduction to Machine Learning Algorithms*. 2018-06. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (acedido em 2023-02-07) (ver p. xvi).

[19] P. Georgiadis e J. Williams. «Second UK train operator to cut services amid staff shortages». Em: *Financial Times* (2022-09). URL: https://www.ft.com/content/c0732e48-b493-4e86-9e98-5745f3dee22f (acedido em 2023-01-15) (ver p. 2).

[20] S. Glen. *ARMA model*. URL: https://www.statisticshowto.com/arma-model/ (acedido em 2023-01-30) (ver p. 36).

[21] Google. *Imbalanced data*. 2022-07. URL: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data (acedido em 2023-02-07) (ver p. xvii).

[22] C. R. Harris et al. «Array programming with NumPy». Em: *Nature* 585.7825 (2020-09), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (ver p. 45).

[23] J. D. Hunter. «Matplotlib: A 2D graphics environment». Em: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55 (ver p. 45).

[24] R. J. Hyndman e G. Athanasopoulos. *Forecasting: principles and practice*. 2ª ed. https://otexts.com/fpp2/. Melbourne, Australia: OTexts, 2018 (ver pp. 20–23, 29, 30, 33, 34, 38–41, 57, 67, 78).

[25] R. J. Hyndman e G. Athanasopoulos. *Forecasting: principles and practice*. 3ª ed. https://otexts.com/fpp3/. Melbourne, Australia: OTexts, 2021 (ver pp. 23, 30–33, 42).

[26] R. J. Hyndman e Y. Khandakar. «Automatic time series forecasting: the forecast package for R». Em: *Journal of statistical software* 27 (2008), pp. 1–22. DOI: 10.18637/jss.v027.i03 (ver p. 39).

[27] R. J. Hyndman e A. B. Koehler. «Another look at measures of forecast accuracy». Em: *International journal of forecasting* 22.4 (2006), pp. 679–688 (ver p. 33).

[28] Q. Kong, T. Siauw e A. Bayen. *Python Programming and Numerical Methods: A Guide for Engineers and Scientists*. Elsevier Science, 2020. ISBN: 9780128195499. URL: https://pythonnumericalmethods.berkeley.edu/notebooks/Index.html (ver p. xvii).

[29] L. Krippahl. *Notas das aulas de Aprendizagem Automática*. Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa. https://aa.ssdi.di.fct.unl.pt/teoricas.html. 2022 (ver p. xvii).

[30] D. Kwiatkowski et al. «Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?» Em: *Journal of econometrics* 54.1-3 (1992), pp. 159–178 (ver p. 28).

[31] E. Lima, T. Vieira e E. de Barros Costa. «Evaluating deep models for absenteeism prediction of public security agents». Em: *Applied Soft Computing* 91 (2020), p. 106236 (ver p. 13).

[32] D. d. S. Lopes. «Desenvolvimento de um modelo de previsão de preços no mercado Ibérico de eletricidade». Tese de mestrado. Instituto Superior de Engenharia do Porto, 2014 (ver pp. 17, 35, 36).

[33] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (ver p. ii).

[34] G. Maróti. «Operations research models for railway rolling stock planning». Em: *Dissertation Abstracts International* 68.01 (2006) (ver pp. 2, 3).

[35] G. Marques. *Notas das aulas de Aprendizagem Automática*. Instituto Superior de Engenharia de Lisboa do Instituto Politécnico de Lisboa. 2020 (ver pp. xvii, 78).

[36] A. Martiniano et al. «Application of a neuro fuzzy network in prediction of absenteeism at work». Em: *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. IEEE. 2012, pp. 1–4. ISBN: 978-1-4673-2843-2 (ver pp. 8, 10–12).

[37] A. Martins. *Notas das aulas de Métodos de Previsão*. Instituto Superior de Engenharia de Lisboa do Instituto Politécnico de Lisboa. 2018 (ver pp. 16–22, 24, 29, 36, 37, 57).

[38] G. Matos. «Optimisation of Periodic Train Timetables». Tese de mestrado. Instituto Superior Técnico, 2018-04 (ver p. 3).

[39] G. Matos, L. Albino e R. Saldanha. «Um modelo de aprendizagem automática para previsão de ausências de maquinistas». Em: *In Proceedings of 10º Congresso Rodoferroviário Português*. LNEC. Lisboa, 2022 (ver pp. 1, 5, 47, 77).

[40] NLTimes. «NS says staff shortages were at their worst in September». Em: *NL Times* (2022-12). URL: https://nltimes.nl/2022/12/19/ns-says-staff-shortages-worst-september (acedido em 2023-01-15) (ver p. 2).

[41] M. Ozturkmen. *Stationarity Tests in Time Series*. Kaggle notebook. Kaggle. 2021. URL: https://www.kaggle.com/code/mozturkmen/stationarity-tests-in-time-series (ver p. 27).

[42] J. Pais. «Métodos de Previsão». Tese de mestrado. Faculdade de Economia da Universidade do Porto, 2017 (ver pp. 18, 20, 24–26).

[43] F. Pedregosa et al. «Scikit-learn: Machine Learning in Python». Em: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (ver p. xvii).

[44] I. de Portugal. *Terminologia*. URL: https://www.infraestruturasdeportugal.pt/pt-pt/terminologia?letter=E (acedido em 2022-11-18) (ver p. 1).

[45] D. Potthof. «Railway Crew Rescheduling:Novel Approaches and Extensions». Tese de doutoramento. Erasmus University Rotterdam, 2010. ISBN: 978–90–5892–250–2 (ver pp. 3, 4).

[46] S. E. Said e D. A. Dickey. «Testing for unit roots in autoregressive-moving average models of unknown order». Em: *Biometrika* 71.3 (1984), pp. 599–607 (ver p. 27).

[47] S. Seabold e J. Perktold. «statsmodels: Econometric and statistical modeling with python». Em: *9th Python in Science Conference*. 2010 (ver p. 45).

[48] SISCOG. *Replaneamento em cenários de emergência*. URL: https://www.siscog.pt/pt/noticias/replaneamento-em-cenarios-de-emerg%C3%AAncia/ (acedido em 2022-11-18) (ver p. 4).

[49] SISCOG. *Software de apoio à decisão para planeamento e gestão otimizado de recursos*. URL: https://www.siscog.pt/pt/sobre-nos/ (acedido em 2023-01-09) (ver p. 1).

[50] SISCOG. *Soluções de Planeamento e Gestão de Recursos para transportes públicos*. URL: https://www.siscog.pt/pt/produtos (acedido em 2023-05-22) (ver pp. 5, 6).

[51] M. Skorikov et al. «Prediction of absenteeism at work using data mining techniques». Em: *2020 5th International Conference on Information Technology Research (ICITR)*. IEEE. 2020, pp. 1–6. DOI: 10.1109/ICITR51448.2020.9310913 (ver pp. 8–12).

[52] T. G. Smith et al. *pmdarima: ARIMA estimators for Python*. [Online]. 2017. URL: http://www.alkaline-ml.com/pmdarima (ver pp. 38, 40, 62).

[53] C. Soares. *Notas das aulas de Processamento de Streams*. Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa. 2022 (ver p. xvi).

[54] S. J. Taylor e B. Letham. «Forecasting at scale». Em: *The American Statistician* 72.1 (2018), pp. 37–45. DOI: https://doi.org/10.1080/00031305.2017.1380080 (ver p. 41).

[55] S. J. Taylor e B. Letham. *Prophet: Forecasting at Scale*. https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/. 2017 (ver p. 42).

[56] T. pandas development team. *pandas-dev/pandas: Pandas*. Versão latest. 2020-02. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134 (ver pp. 45–47).

[57] TIS. *Transporte ferroviário - Enquandramento*. URL: https://www.tis.pt/transporte-ferrovi%C3%A1rio.html (acedido em 2022-11-18) (ver p. 1).

[58] O. Triebe et al. «Neuralprophet: Explainable forecasting at scale». Em: *arXiv preprint arXiv:2111.15397* (2021) (ver p. 43).

[59] C. Valentin. *In-Depth Understanding of NeuralProphet through a Complete Example*. 2022-01. URL: https://towardsdatascience.com/in-depth-understanding-of-neuralprophet-through-a-complete-example-2474f675bc96 (acedido em 2023-09-13) (ver pp. 42, 43).

[60] P. Virtanen et al. «SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python». Em: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (ver p. 63).

[61] Z. Wahid et al. «Predicting Absenteeism at Work Using Tree-Based Learners». Em: 2019-01, pp. 7–11. DOI: 10.1145/3310986.3310994 (ver pp. 10, 11).

[62] C. Wong. «Machine learning and time series decomposition approaches for predicting airline crew non-availability». Tese de mestrado. The University of Texas at Austin, 2020-05 (ver p. 14).

[63] J. Xavier. «Análsise e previsão de séries temporais com modelos ARIMA e análise espectral singular». Tese de mestrado. Universidade Aberta, 2016 (ver pp. 18, 19, 28, 37).

I

Descrição dos atributos que constituem cada conjunto de dados

Tabela I.1: Todos os atributos que compõe os conjuntos de dados.

Componente	Nome	Conteúdo	Tipo
Índice	`date`	Data (formato yyyy-mm-dd hh:mm:ss)	*datetime64[ns]*
Atributos	`count_absence`	Número de tipulantes ausentes	*int*
	`count_duties`	Número de tripulantes que foram trabalhar	*int*
	`count_planned`	Número de tripulantes planeados para trabalhar	*int*
	`absence_rate`	Percentagem de ausências diária	*float*
	`holidays`	1 se for feriado, ou 0 caso contrário	*boolean*

II

Escolha dos hiperparâmetros e séries de *Fourier*

Tabela II.1: Resumo dos hiperparâmetros escolhidos para o Prophet e NeuralProphet, para as séries temporais dos maquinistas e revisores da base B01.

Método	Função desempenhada	Hiperparâmetro	Valor
Prophet	Maquinistas	`changepoint_prior_scale`	0.1
		`changepoint_range`	0.4
	Revisores	`changepoint_prior_scale`	0.05
		`changepoint_range`	0.8
NeuralProphet	Maquinistas	`epochs`	500
		`changepoint_range`	0.4
		`learning_rate`	0.001
		`batch_size`	32
	Revisores	`learning_rate`	0.01
		`batch_size`	64

Tabela II.2: Definição das séries de *Fourier*, para as séries temporais dos maquinistas e revisores da base B01.

Função	Método	Séries de Fourier
Maquinistas	RDH	Fourier(k = 15, m = 365.25)
		Fourier(k = 5, m = 30.44)
		Fourier(k = 2, m = 7)
Revisores	RDH	Fourier(k = 15, m = 365.25)
		Fourier(k = 5, m = 7)
	Prophet	Fourier(k = 20, m = 365.25)
		Fourier(k = 5, m = 7)

III

Resultados da RDH, do Prophet e do NeuralProphet com o método da janela fixa

Tabela III.1: Resumo das métricas de desempenho — MAE, RMSE e MAPE — das previsões das percentagens de absentismo e do número total de ausências de maquinistas e de revisores da base B01, com e sem a incorporação da variável `'holidays'`.

			Métrica c/ `'holidays'`			Métricas s/ `'holidays'`		
Função	Tipo de previsão	Método	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
Maquinistas	Percentagem	RDH	1.9853	2.5789	19.96	1.9654	2.5523	19.79
		Prophet	1.9591	2.5369	20.70	1.9353	2.4937	20.53
		NeuralProphet	1.9653	2.5112	21.09	1.9435	2.4656	20.90
	Totais	RDH	3.3741	4.3282	19.96	3.3741	4.3282	19.79
		Prophet	3.3728	4.2718	20.70	3.3444	4.2420	20.53
		NeuralProphet	3.3798	4.2307	21.09	3.362	4.3194	20.90
Revisores	Percentagem	RDH	2.1460	2.7443	13.68	2.1704	2.7882	13.78
		Prophet	2.3115	2.9149	13.55	2.3247	2.9235	13.66
		NeuralProphet	2.0947	2.6675	12.86	2.0971	2.6632	12.92
	Totais	RDH	4.5161	5.6775	13.68	4.5710	5.7749	13.78
		Prophet	4.8124	5.9001	13.55	4.8545	5.9549	13.66
		NeuralProphet	4.3069	5.2770	12.86	4.3191	5.2747	12.92

IV

Resultados da RDH, do Prophet e do NeuralProphet com o método da janela móvel de previsão a H-passos

Tabela IV.1: Métricas de desempenho da RDH, do Prophet e do NeuralProphet da percentagem de ausências dos maquinistas da base B01, para diferentes horizontes temporais.

		h-passos							
Método	Métrica	1	7	14	30	60	90	180	365
RDH	MAE	1.958	2.2468	2.3138	2.3364	2.2583	2.2412	2.1639	1.981
	RMSE	3.0603	3.2641	3.3659	3.3465	3.0701	3.0361	2.9752	2.5594
	MAPE (%)	20.07	21.93	22.39	22.29	22.24	21.49	21.5	19.8
Prophet	MAE	2.0581	2.137	2.1313	2.1551	2.1033	2.1685	2.2361	1.894
	RMSE	3.0945	3.1792	3.1859	3.2094	2.928	2.968	3.0417	2.4368
	MAPE (%)	20.71	21.25	21.11	21.18	21.38	21.39	22.61	19.26
NeuralProphet	MAE	2.0299	2.0763	2.0879	2.1416	2.067	2.1107	2.1849	1.8821
	RMSE	3.0663	3.1131	3.1372	3.1708	2.8702	2.8818	2.9626	2.4153
	MAPE (%)	20.25	20.56	20.57	20.95	21.05	20.83	22.2	19.17

Tabela IV.2: Métricas de desempenho da RDH, do Prophet e do NeuralProphet da percentagem de ausências dos revisores da base B01, para diferentes horizontes temporais.

		h-passos							
Método	Métrica	1	7	14	30	60	90	180	365
RDH	MAE	2.0913	2.3455	2.4254	2.5352	2.6398	2.66	2.599	2.5271
	RMSE	2.8448	3.1422	3.1888	3.2701	3.377	3.4114	3.1867	3.2315
	MAPE (%)	12.77	14.3	14.73	15.54	16.37	16.71	15.74	16.55
Prophet	MAE	2.2759	2.3747	2.4609	2.5861	2.7499	2.8868	3.2471	3.5895
	RMSE	3.0397	3.1314	3.1915	3.3089	3.486	3.6268	3.9199	4.4223
	MAPE (%)	14.41	15.01	15.54	16.45	17.67	18.55	19.46	22.94
NeuralProphet	MAE	2.3571	2.452	2.5664	2.726	2.8892	3.0667	3.3343	3.701
	RMSE	3.1266	3.218	3.3192	3.4673	3.6228	3.8096	4.0038	4.5327
	MAPE (%)	15.08	15.71	16.43	17.57	18.65	19.8	20.08	24.03