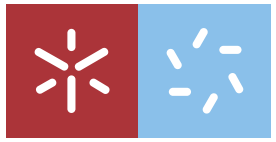
Desenvolvimento de Modelos de Previsão de Variáveis Climáticas

Cláudia Maria Ferreira Costa

UMinho | 2019

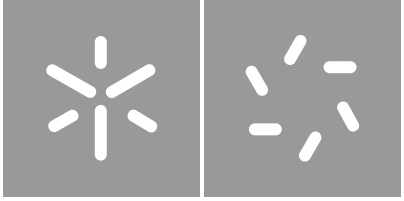


Universidade do Minho
Escola de Ciências

Cláudia Maria Ferreira Costa

Desenvolvimento de Modelos de Previsão de Variáveis Climáticas

outubro de 2019



Universidade do Minho
Escola de Ciências

Cláudia Maria Ferreira Costa

Desenvolvimento de Modelos de Previsão de Variáveis Climáticas

Tese de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Arminda Manuela Andrade Pereira Gonçalves
e do
Professor Doutor Marco André da Silva Costa

outubro de 2019

Agradecimentos

A luz que me deu, vai brilhar para sempre.
Para ti, pai.

À Universidade do Minho e a todos os professores agradecer a educação.

À Professora Doutora Arminda Manuela Gonçalves, agradecer toda a generosidade, a partilha de conhecimentos, toda a disponibilidade e as palavras de incentivo. Agradecer por, apesar da seriedade do trabalho, ter tornado todo o processo leve e divertido. Agradecer toda a amizade e preocupação sincera.

Ao Professor Doutor Marco Costa, agradecer todas as palavras de incentivo e apoio. Agradecer a generosa partilha de conhecimentos e toda a boa disposição. Agradecer toda a amizade e preocupação sincera.

À minha amada Mãe e aos meus amados irmãos, agradecer todo o amor incondicional. Agradecer serem o pilar seguro da minha vida. Agradecer todo o carinho, dedicação e sacrifícios. Nada seria possível e eu não seria nada sem vós. Ao meu amado pai, por nos iluminar sempre.

Ao meu amor, por me ver com olhos mais bondosos do que os meus próprios e a toda a minha querida família agradecer todo o amor, todo o apoio, todo o carinho e força.

A todos os meus queridos amigos, agradecer todo o companheirismo e toda a amizade pura. Agradecer toda a dedicação e todo o amor.

Este trabalho foi financiado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), do Programa Operacional Competitividade e Internacionalização (POCI) e pelo Orçamento da Fundação para a Ciência e Tecnologia (FCT).

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

Num mundo onde as mudanças climáticas e os crescentes conflitos sociais são uma realidade, é essencial uma gestão adequada dos recursos naturais escassos. A análise de séries temporais de dados meteorológicos tem assumido um interesse crescente em muitas áreas, em particular no problema da irrigação. Este estudo realizado no contexto do projeto "TO CHAIR - Os Desafios Óptimos na Irrigação", financiado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), do Programa Operacional Competitividade e Internacionalização (POCI) e pela Fundação para a Ciência e a Tecnologia (FCT), tem como principal objetivo identificar os modelos de previsão mais adequados para modelar séries meteorológicas que têm impacto no processo de evapotranspiração e da humidade no solo, por forma a planear de forma mais eficiente o uso da água nos sistemas de irrigação. Para isso, é necessário estimar e prever variáveis meteorológicas (velocidade média do vento, temperatura mínima e máxima do ar e precipitação) em tempo real (diário) para uma determinada localização, sendo, neste caso, numa quinta em Carrazeda de Ansiães, situada no distrito de Bragança, no Norte de Portugal. Os dados em estudo são registos diários observados no período de 1 de janeiro de 2010 até ao dia 23 de abril de 2019. Assim, neste estudo, apresenta-se uma comparação de dois métodos de previsão, os modelos TBATS (transformação Box-Cox, erros ARMA, tendência e componentes sazonais trigonométricas) e os modelos de regressão linear com erros correlacionados. Estes modelos foram selecionados devido à sua capacidade para modelar flutuações sazonais fortemente presentes nos dados meteorológicos, em particular, em lidar com séries temporais com padrões sazonais complexos.

Palavras-chave: Irrigação, Séries temporais, Variáveis meteorológicas, Previsão, TBATS, Regressão com erros correlacionados.

Abstract

In a world where climate change and growing social conflicts are a reality, proper management of scarce natural resources is essential. There is a growing interest in time series analysis of meteorological data in many areas, in particular regarding the problem of irrigation. This study is carried out in the context of project "TO CHAIR - Optimum Challenges in Irrigation" – funded by the European Regional Development Fund (ERDF), the Competitiveness and Internationalization Operational Program (COMPETE 2020) and the Foundation for Science and Technology (FCT) – and its main objective is to identify the most suitable forecasting models for modeling weather series that have an impact on the evapotranspiration process and on soil humidity, in order to more efficiently plan the use of water in irrigation systems. For this, it is necessary to estimate and forecast meteorological variables (average wind speed, minimum and maximum air temperature and precipitation) in real time (daily) for a given location: in this case, a farm in Carrazeda de Ansiães, in the district of Bragança in the north of Portugal. The data under study consist of daily records observed from January 1, 2010 to April 23, 2019. This study presents a comparison of two forecasting methods, the TBATS models (Box-Cox transformation, ARMA errors, trend and trigonometric seasonal components) and the linear regression models with correlated errors. These models were selected due to their ability to model seasonal fluctuations strongly present in meteorological data, in particular when dealing with time series with complex seasonal patterns.

Keywords: Irrigation, Time series, Meteorological variables, Forecasting, TBATS, Regression with correlated errors.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AIC - *Akaike's Information Criterion* (em português, Critério de Informação de Akaike)

AICc- *Corrected Akaike's Information Criterion* (em português, Critério de Informação de Akaike Corrigido)

Ampli. - Amplitude

AR - *Autoregressive* (em português, Autorregressivo)

ARIMA - *Autoregressive Integrated Moving Average* (em português, Autorregressivo Integrado de Médias Móveis)

ARMA - *Autoregressive Moving Average* (em português, Autorregressivo e de Médias Móveis)

Assi. - Assimetria

BATS - *Box-Cox Transformation, ARMA errors, Trend, and Seasonal components* (em português, Transformação Box-Cox, erros ARMA, tendência e componentes sazonais)

BI - Bolsa de Investigação

BIC - *Bayesian Information Criterion* (em português, Critério de Informação Bayesiano)

Curt. - Curtose

CV - Coeficiente de Variação Amostral

EAM - Erro Absoluto Médio

EEAM - Erro Escalado Absoluto Médio

EM - Erro Médio

EPAM - Erro Percentual Absoluto Médio

EQM - Erro Quadrático Médio

ETS - *Exponential Trignometric Smoothing* (em português, Alisamento Exponencial Trigonométrico)

FAC - Função de Autocorrelação

FACP - Função de Autocorrelação Parcial

FCT - Fundação para a Ciência e a Tecnologia

FEDER - Fundo Europeu de Desenvolvimento Regional

IC - Intervalo de Confiança

IPMA - Instituto Portugês do Mar e da Atmosfera

MA - *Moving Average* (em português, Médias Móveis)

Max. - Máxima

Med. - Mediana

Min. - Mínima

MR - Modelo de Regressão

OLS - *Ordinary Least Squares* (em português, Estimador de Mínimos Quadrados)

OMM - Organização Mundial de Meteorologia

POCI - Programa Operacional Competitividade e Internacionalização

Quart. - Quartil

REQM - Raiz do Erro Quadrático Médio

SARIMA - *Seasonal Autoregressive Integrated Moving Average* (em português, Autorregressivo Integrado de Médias Móveis Sazonal)

SCov - *Structural Models with Covariates* (em português, Modelos Estruturais com Covariáveis)

SES - *Simple Exponential Smoothing* (em português, Alisamento Exponencial

Simples)

STL - *Seasonal-Trend Decomposition by Loess* (em português, Decomposição Sazonal e de Tendência usando Loess)

TBATS - *Trignometric, Box-Cox Transformation, ARMA errors, Trend, and Seasonal components* (em português, componentes Transformação Box-Cox, erros ARMA, tendência e componentes sazonais trigonométricas)

Temp. Máxima - Temperatura Máxima

Temp. Mínima - Temperatura Mínima

TETS - *Trignometric Exponential Smoothing* (em português, Alisamento Exponencial Trigonométrico)

TSCov - *Trignometric Structural Models with Covariates* (em português, Modelos Trigonométicos Estruturais com Covariáveis)

Capítulo 1

Introdução

Num mundo onde a mudança climática e os crescentes conflitos sociais são uma realidade, uma gestão adequada dos recursos existentes é vital. É necessário discutir o uso da água, procurando encontrar as melhores soluções técnicas, para melhorar a eficiência do seu uso, em particular nos sistemas de rega, em resposta às preocupações ambientais e de sustentabilidade.

A maioria dos sistemas de rega no mercado baseiam-se no controlo on-off sem técnicas de previsão associadas. O sistema dispara o ciclo de rega quando um valor crítico mínimo da humidade do solo é detetado e suspende-o quando um máximo definido é atingido (por vezes, perto da saturação). O excesso de água no solo é, frequentemente, uma consequência desta técnica sendo responsável por um desperdício de água significativo. Desta forma, a modelação matemática para compreender o comportamento da humidade no solo permite, entre outras, o planeamento eficiente do uso da água dos sistemas de rega (Lopes, 2018). De acordo com o Instituto Português do Mar e da Atmosfera (IPMA), a 30 de setembro de 2017, cerca de 81% do território de Portugal Continental se encontrava em seca severa, 7,4% em seca extrema, 10,7% em seca moderada e 0,8% em seca fraca. O ano de 2017 foi um ano extremamente seco e, considerando os dados de 1 de janeiro de 2017 a 27 de dezembro de 2017, encontrava-se entre os 4 anos mais secos desde 1931 (todos ocorreram após 2000) e com uma precipitação total anual média de 60% menor do que é considerado normal. A quantidade de água doce disponível no nosso planeta é de apenas 2,5% do volume total disponível, sendo que cerca de 70% desta é usada na Agricultura. Assim, muito deve ser feito para se economizar água, uma vez que esta é vital para o nosso planeta.

Este estudo foi realizado no âmbito de uma bolsa de investigação (BI) do Projeto POCI-01-0145-FEDER-028247, intitulado “TO CHAIR - Os Desafios Óptimos na Ir-

rigação", financiado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), do Programa Operacional Competitividade e Internacionalização (POCI) e pelo Orçamento da Fundação para a Ciência e Tecnologia (FCT). O projeto envolve 4 Universidades, a Universidade do Minho, a Universidade de Aveiro, a Universidade do Porto e a Universidade Trás-os-Montes e Alto Douro, envolvendo 19 investigadores e vários bolseiros. Este projeto pretende compreender e analisar o comportamento da humidade no solo através da modelação matemática e estatística com o objetivo de traçar um planeamento eficiente dos recursos hídricos. Para tal, é necessário desenvolver modelos de previsão para diferentes variáveis climáticas, em tempo real (por dia) para um local específico, no caso em estudo, uma quinta em Carrazeda de Ansiães, situada no Norte de Portugal. Neste estudo são, assim, analisados dados da Universidade de Trás-os-Montes e Alto Douro e do Instituto Português do Mar e da Atmosfera (IPMA), com o objectivo de estimar e prever adequadamente as perdas de água por evapotranspiração.

A evaporatranspiração é o processo pelo qual a água é transferida da terra para a atmosfera por evaporação do solo e por transpiração das plantas. A quantificação da evaporatranspiração é fundamental para as atividades ligadas à gestão das bacias hidrográficas, na modelação meteorológica e hidrológica e, em particular, no planeamento da irrigação das culturas agrícolas. Os modelos estatísticos mais utilizados para analisar as séries meteorológicas são os modelos lineares (generalizados) e os modelos de séries temporais.

O principal objetivo do projeto é identificar os modelos de previsão mais adequados para modelar as séries meteorológicas que têm impacto no processo de evapotranspiração, procedendo-se ao seu estabelecimento e comparação em termos da sua capacidade explicativa e preditiva. Estas séries temporais, geralmente, apresentam um comportamento com forte tendência e sazonalidade de alta frequência.

Dada a natureza das séries e o seu comportamento, o estudo apresentado nesta dissertação apresenta uma comparação de dois modelos, os modelos TBATS que incorporam componentes de transformação Box-Cox, erros ARMA, tendência e componentes sazonais trignométricas e os modelos de regressão linear com erros correlacionados. Estes modelos foram selecionados dada a sua capacidade para lidar com dados com padrões sazonais complexos. De referir que para a aplicação das metodologias aos dados recorreu-se ao *software* R.

Os fatores meteorológicos de extrema importância para a Agricultura e, por sua vez, utilizados para os planos de irrigação são: a radiação solar, a temperatura máxima do ar, a temperatura mínima do ar, a precipitação e a velocidade do vento.

Estas variáveis são utilizadas para calcular a evapotranspiração de referência. Neste projeto são estudadas apenas quatro destas variáveis meteorológicas. A temperatura máxima do ar (ºC), a temperatura mínima do ar (ºC), a precipitação (mm) e a velocidade média do vento (m/s). Tratam-se de registos diários observados no período de 1 de janeiro de 2010 até dia 23 de abril de 2019 para uma quinta, chamada de Senhora da Ribeira, em Carrazeda de Ansiães no distrito de Bragança no Norte de Portugal.

Quanto à organização do conteúdo da dissertação, é apresentada uma breve revisão da literatura (Capítulo 2), referente aos métodos de previsão de séries temporais, alguns exemplos de aplicação destes métodos, bem como, alguns critérios necessários à avaliação dos métodos. Nos Capítulos 3 e 4 descrevem-se, de forma introdutória, os conteúdos teóricos relacionados com os métodos de previsão de séries temporais, os métodos que são aplicados aos dados e as medidas de avaliação utilizadas para selecionar o método mais adequado ao estudo realizado. A aplicação prática das metodologias, apresentadas nos capítulos anteriores, aos dados é exposta no Capítulo 5 e 6. É apresentada uma análise exploratória dos dados (diários) no Capítulo 5, seguindo-se a aplicação das duas metodologias selecionadas: o modelo TBATS e o modelo de regressão com erros correlacionados no Capítulo 6. É realizado um estudo comparativo da capacidade preditiva dos dois métodos aplicados, através das medidas de avaliação apresentadas. As considerações finais são apresentadas no Capítulo 7 e no Capítulo 8 são apresentadas as propostas para trabalho futuro.

Capítulo 2

Revisão de Literatura

”Em locais de escassez de água, o foco não é apenas a gestão do abastecimento de água, mas o uso eficiente dos recursos hídricos disponíveis. Uma vez que, mais de 70% da água que consumimos em todo o mundo vai para a agricultura, que haverá cerca de 11 mil milhões de bocas para alimentar até 2100, melhorar a eficiência de irrigação terá um impacto potencialmente significativo no uso global de água” (Alexandratos et al., 2019).

Um artigo da Agência Europeia do Ambiente (2012) refere que ”Um dos domínios em que as novas práticas e políticas podem dar um contributo significativo em matéria de ganhos de eficiência na utilização dos recursos hídricos é o da irrigação das culturas. Em países do sul da Europa, como a Grécia, Itália, Portugal, Chipre e Espanha, e no sul de França, as condições áridas ou semiáridas impõem o recurso à irrigação. No entanto, não é necessário que a utilização de água na irrigação seja tão intensiva. Atualmente, já se conseguem ganhos de eficiência na utilização da água em toda a Europa, quer através da eficiência do transporte (a percentagem de água captada e fornecida aos campos) e a eficiência da utilização no terreno (a água efetivamente utilizada numa cultura, em comparação com a quantidade total de água fornecida a essa cultura). Na Grécia, por exemplo, a melhoria da eficiência das redes de transporte e distribuição permitiu obter um ganho de eficiência estimado em 95% na utilização da água em comparação com os métodos de irrigação anteriormente utilizados.”

No artigo de Wang e Cai (2009) é possível ler-se que ”O planeamento da irrigação, determina o tempo e a quantidade de água necessária a aplicar a uma terra cultivada durante a estação de crescimento”. O artigo explora o uso de previsões de variáveis meteorológicas no planeamento da irrigação usando métodos de simulação e otimização, comparando os ganhos registados com essas previsões para aqueles

registados usando um esquema de irrigação recuperado, o que indica o real comportamento do agricultor. Muitos estudos anteriores abordaram o uso de previsões meteorológicas no planeamento da irrigação. Uma grande ênfase da pesquisa foi identificar o horizonte de previsão, (i.e., sazonal, duas semanas, 7 dias) que é o mais provável para melhorar o planeamento. Venäläinen et al. (2005) examinaram a utilidade das previsões climáticas sazonais a partir de modelos atmosféricos numéricos para suplementar ou substituir dados meteorológicos medidos para a modelação da humidade e irrigação do solo. Descobriram que os erros nas previsões sazonais da chuva podem ter um efeito importante na previsão da irrigação.

Atualmente, já existem inúmeras plataformas *online* que fazem excelentes previsões das variáveis meteorológicas, como a temperatura máxima e mínima, a precipitação, a velocidade do vento, entre muitas outras. Como exemplo, a nível nacional, o IPMA é um instituto público que assume as responsabilidades ao nível do território nacional nos domínios do mar e da atmosfera. A nível internacional, já existem muitos *sites*, por exemplo o *site APIXU*, que abrangem dados meteorológicos de todo o planeta. As previsões são feitas com base em dados históricos de muitos anos que asseguram que as previsões meterológicas sejam precisas. Estes *sites* de previsão utilizam métodos de interpolação, modelos matemáticos e geofísicos.

Na literatura, facilmente se encontram inúmeras aplicações de análise de séries temporais, na saúde, economia, finanças, engenharia, agricultura, entre outras. Uma série temporal é uma sequência de valores de uma variável registada com igual período de tempo. Os desenvolvimentos teóricos das séries temporais começaram muito cedo com o desenvolvimento dos processos estocásticos. A primeira aplicação de modelos autorregressivos a dados foi resultante do trabalho de Yule e J.Walker nos anos 20 e 30 do século passado. Inspirados por estes trabalhos, Box e Jenkins (1970) desenvolveram uma abordagem prática para a construção de modelos autorregressivos integrados e de médias móveis, designados por *Autorregressive Integrated Moving Average Models* (ARIMA).

Na dissertação de Lima (2018) é estudado um conjunto de séries temporais do segmento do retalho com fortes tendências e padrões sazonais. O objetivo da tese consistiu em avaliar a precisão de vários métodos de previsão, na área da modelação de séries temporais, aplicados a dados do segmento do retalho. Apresenta um estudo comparativo da precisão entre os modelos autorregressivos e de médias móveis *Autorregressive Moving Average Models* (ARMA) e respetivas extensões, os modelos de decomposição clássica associados a modelos de regressão linear múltipla e métodos de alisamento exponencial. A escolha recaiu sobre estes modelos, uma

vez que, apresentam capacidade de modelar tendências e flutuações sazonais. Para avaliar a capacidade preditiva do método foram utilizadas diversa medidas de avaliação. Entre as quais, o erro quadrático médio (EQM), a raiz do erro quadrático médio (REQM), o erro percentual absoluto médio (EPAM), o erro escalado absoluto médio (EEAM) e estatística de U de Theil.

Alpuim e El-Shaarawi (2008) referem que estimar e testar tendências em séries temporais é um procedimento comum em diferentes campos da aplicação da Estatística como, por exemplo, a epidemiologia, a econometria e a estatística ambiental. Geralmente, a estimação é feita com recurso a modelos de regressão que incluem tempo e/ou funções de tempo como variáveis independentes e, por vezes, em conjunto com outras variáveis. Contudo, acontece frequentemente os resíduos serem autocorrelacionados o que pode levar a estimações incorretas da variância dos estimadores de mínimos quadrados ordinários (OLS).

No artigo Alpuim e El-Shaarawi (2009) são examinadas médias mensais de séries temporais de temperatura em duas cidades Europeias, Lisboa (1856-1999) e Praga (1841-2000). Neste artigo, ondas seno e cosseno foram incluídas como variáveis independentes para descrever o padrão sazonal da temperatura e as ondas seno e cosseno multiplicadas pelo tempo foram usadas para descrever o aumento da temperatura correspondente aos diferentes meses. O modelo também tem em consideração a estrutura autorregressiva, AR (1), encontrada nos resíduos. Um teste da significância das variáveis que descrevem a variação do aumento de temperatura mostra que Lisboa e Praga tiveram um aumento de temperatura diferente de acordo com o mês. Os meses de Inverno mostram um aumento maior que os meses de Verão.

Na dissertação Monteiro (2017) pode ler-se que "O aquecimento global e, em particular, o dos oceanos é um assunto de extrema importância, atendendo à atualidade das alterações climáticas e da consequente necessidade de adaptação das populações face a este fenómeno. As séries temporais de variáveis físicas, como por exemplo a temperatura à superfície do mar, permitem estudar e interpretar estes fenómenos ao procurar modelos matemáticos para os descrever e prever, quando possível". Nesta dissertação são utilizadas técnicas estatísticas de análise de séries temporais no estudo da temperatura à superfície do mar. Para cada região foram explorados modelos de séries temporais, nos quais se incluem os modelos de decomposição clássica e de Holt-Winters e, ainda, os modelos ARIMA sobre a série residual, resultante do ajustamento de uma tendência linear e de uma componente sazonal à série temporal inicial. A seleção do melhor modelo foi efetuada por medidas de ajustamento disponíveis para este tipo de análise.

Um dos fatores na determinação das previsões corretas para um determinado fenómeno é a seleção do modelo apropriado. Os modelos mais comuns são os modelos ARMA, ARIMA e SARIMA (Box e Jenkins (1970), Lee e Ko (2011), Pappas et al. (2008), Chen et al. (1995)) e modelos de alisamento (ou suavização) exponencial (Taylor (2003) e Kostenko e Hyndman (2008)). A utilização destes modelos é perfeitamente apropriada, desde que as séries a trabalhar não apresentem padrões sazonais complexos. A solução para este problema é o modelo *Trignometric, Box-Cox Transformation, ARMA errors, Trend, and Seasonal components*(TBATS), introduzido há alguns anos por De Livera et al. (2011).

Murat et al. (2018) fazem previsões de séries temporais diárias meteorológicas usando os modelos ARIMA e de regressão. Trabalharam séries temporais diárias da temperatura do ar e de precipitação, recolhidas entre 1 de janeiro de 1980 e 31 de dezembro de 2010, em quatro *sites* europeus, provenientes de diferentes regiões. Utilizaram para modelar e prever, os métodos Box-Jenkins e Holt-Winters sazonal autorregressivo de médias móveis, modelos ARIMA com regressores externos na forma de termos de Fourier e modelos de regressão, incluindo componentes de tendência e sazonalidade. Referem que a previsão de eventos futuros, de variáveis meteorológicas, com base em séries temporais históricas é de grande importância para a modelação agrofísica (Lamorski et al. (2013), Baranowski et al. (2015) e Murat et al. (2016)). Muitas vezes os métodos de previsão de séries temporais são baseados nestas análises de dados históricos. Estes métodos assumem que os padrões do passado podem ser usados para prever eventos futuros. Falam no facto, de nos últimos anos, um dos modelos mais utilizados ser o autorregressivo integrado de médias móveis (ARIMA) e que têm como objetivo o estudo rigoroso e cuidado das observações do passado de uma série temporal para desenvolver modelos aproximados que permitam a previsão de valores futuros para as séries. Estes modelos têm três constantes de controlo: a irregular, a tendência e sazonalidade. Os modelos de erros correlacionados surgem da incapacidade dos modelos ARIMA não conseguirem lidar com sazonalidades com período superior a 200. Nestes modelos, são adicionados regressores externos na forma de termos de Fourier para modelar a sazonalidade.

Apesar dos modelos escolhidos não conseguirem prever exatamente os valores da temperatura e precipitação, eles dão informação que ajuda a devenvolver estratégias para um planeamento sustentável e apropriado para a Agricultura, ou serem apenas utilizados como uma ferramenta suplementar no planeamento de estratégias. De forma a mitigar as limitações dos modelos de espaço de estados quanto à previsão de séries temporais com padrões sazonais complexos, tais como problemas de sazonali-

dade de alta frequência, períodos sazonais múltiplos, sazonalidade com períodos não inteiros, efeitos de duplo calendário, os modelos de espaço de estados tradicionais sofreram algumas modificações. De Livera et al. (2011) introduziram duas novas estruturas, BATS como acrónimo para os principais recursos do modelo: *Box-Cox Transformation, ARMA errors, Trend, and Seasonal components* e TBATS como acrónimo para os principais recursos do modelo: *Trignometric, Box-Cox Transformation, ARMA errors, Trend, and Seasonal components*. Estas modificações vão desde a introdução de transformações Box-Cox, representações de Fourier com coeficientes a variar no tempo e correções de erros ARMA. Expressões analíticas e avaliação por máxima verosimilhança, conduzem a uma abordagem mais simples e compreensiva para previsão de séries temporais com padrões sazonais complexos. Uma das maiores vantagens desta abordagem é a redução da carga computacional na estimação da Máxima Verosimilhança e a enorme aplicabilidade e versatilidade, como ilustram em três estudos empíricos diferentes. Ainda, realçam o facto de a formulação trigonométrica proposta ser apresentada como um meio de decompor séries temporais complexas e mostram que esta decomposição leva à identificação e extração das componentes sazonais que, por si só, não seriam apresentadas no gráfico da série temporal.

Brożyna et al. (2018) apresentam um artigo que descreve a capacidade do modelo TBATS, que não possui restrições de sazonalidade, de criar previsões detalhadas e de longo prazo. Referem que, quer se pretenda prever preços de ações, taxas de desemprego ou temperaturas, é importante escolher o modelo que melhor descreve o fenómeno no passado, que será melhor para prever no futuro. Dependendo da sua natureza, uma série temporal é constituída por: tendência, movimentos sazonais, movimentos cíclicos e componente irregular. As séries temporais são frequentemente analisadas usando os dados agregados para obter uma única sazonalidade e um período de previsão adequado. Por exemplo, dados mensais para o próximo ano podem ser previstos com base em dados mensais de algumas dezenas de anos anteriores; e os dados para as próximas semanas podem ser previstos com base em dados de algumas dúzias de semanas ou meses prévios. O artigo pretende apresentar a capacidade do modelo TBATS trabalhar com sazonalidade usando séries temporais específicas, por exemplo, com dados horários recolhidos ao longo de um período de vários anos e, assim, gerar uma previsão de médio prazo com a especificidade de uma previsão de curto prazo. Para esta análise, foram usados dados horários sobre a procura de energia elétrica na Polónia, a partir de um período de 14 anos, que permitiu observar e incorporar três padrões sazonais diferentes.

Outro exemplo, em Naim et al. (2018), estes apresentam um estudo comparativo entre o modelo BATS e TBATS para previsões a curto prazo de séries com padrões sazonais complexos. Referem que, uma série temporal pode ter um padrão sazonal único e/ou complexo. Padrões complexos incluem sazonalidade de período não inteiro, vários períodos sazonais e efeitos de duplo calendário. Os métodos de séries temporais tradicionais, como o método Naïve, o método *Drift*, o método Holt, o método Holt com *drift* e ARIMA são usados com sucesso para modelar séries univariadas. Os métodos ETS e SARIMA são métodos muito utilizados e eficientes quando as séries temporais apresentam um único padrão sazonal, mas não conseguem um desempenho satisfatório quando existe um padrão sazonal complexo. Relembram que, atualmente, as séries temporais com padrões sazonais complexos são um fenómeno comum. Por exemplo, o consumo de eletricidade, o consumo de gás natural para uma organização, a taxa de chegada em *call centers* por hora ou por dia, etc., não têm periodicidade regular, mas sim uma sazonalidade dinâmica. Considerando este facto, o objetivo do artigo foi desenvolver um modelo de previsão univariado de curto prazo mais eficaz usando os métodos BATS e TBATS para prever padrões sazonais complexos.

No que respeita a estudos comparativos de análise, modelação e previsão da velocidade do vento, Benth e Benth (2010) propuseram um modelo ARMA para a série temporal do vento de uma localização espacial única e estimaram com dados na amostra (*in-sample*) registados em três regiões diferentes de parques eólicos, no estado de Nova Iorque. A série é constituida por observações medidas de três em três horas e utilizaram médias de velocidade de vento diárias. Conseguiram demonstrar que existem grandes discrepâncias no comportamento das observações médias diárias e das observações recolhidas de três em três horas. A avaliação baseada em observações *out-of-sample*, reflete que os modelos propostos são confiáveis e que podem ser usados em diferentes aplicações, como exemplo, previsão do tempo, geração de energia produzida pelo vento, entre outras. Compararam o poder de predição da velocidade do vento de três em três horas com o das médias diárias. Para tal, calcularam o erro de predição quadrático médio (MSPE).

Noutro exemplo, Jain (2018) implementa um modelo ARIMA para a previsão da velocidade do vento. Os resultados são comparados com o modelo respetivo, i.e., um modelo ETS (*Exponential Trignometric Smoothing*). O autor pretende fazer previsão das condições do tempo na Índia. Após a abordagem ARIMA, surge a necessidade da decomposição sazonal. Para tal, utilizou o modelo TBATS introduzido por De Livera et al. (2011). Segue com a abordagem ETS. Como resultado,

identifica que a abordagem ARIMA é mais eficiente, já que captura de forma mais eficiente o comportamento dinâmico das propriedades da velocidade do tempo, quando comparado com o modelo ETS. Apesar da análise preliminar concluir que o modelo ARIMA é melhor em detrimento do ETS, considera de extrema necessidade uma investigação mais profunda.

Na tese de doutoramento de Puindi (2018) percebe-se que modelos estruturais de espaço de estados são bastante eficientes para modelar séries com padrões sazonais complexos. Contudo, quando o objetivo é a previsão, sabe-se que qualquer informação adicional, disponível em variáveis de influência externa, poderá melhorar as previsões. Neste contexto, trabalhos sobre modelos de previsão de séries temporais com sazonalidade complexa e que integram os efeitos das covariáveis são praticamente inexistentes. Refere ainda que, apenas existe um modelo formulado para lidar com séries temporais de sazonalidade complexa, trata-se do modelo TBATS. Esta tese contribuiu para a formulação de modelos estruturais dinâmicos com a integração dos efeitos das covariáveis. Foram construídos dois modelos estruturais baseados na formulação de múltiplas fontes de aleatoriedade. O primeiro modelo, SCov, redefine métodos tradicionais de alisamento exponencial sazonal simples. O segundo modelo, denominado por TSCov é uma extensão do modelo TBATS formulado para acomodar as séries temporais com sazonalidade complexa. Os modelos são formulados através das três componentes não observáveis: nível, tendência e sazonalidade que são consideradas aleatórias e variantes no tempo.

Shu et al. (2014) referem que os modelos ARIMA são os modelos mais utilizados em séries temporais. Com o objetivo de aumentar o nível de precisão do modelos, no estudo que apresentam, sugerem uma abordagem para minimiar os resíduos através da modificação com recurso às séries de Fourier.

Capítulo 3

Séries Temporais

Uma série temporal é um conjunto de observações medidas sequencialmente ao longo do tempo. Os desenvolvimentos teóricos das séries temporais começaram muito cedo com o desenvolvimento dos processos estocásticos. Inspirados pelos trabalhos de Yule e de J. Walker, Box e Jenkins (1970) desenvolveram uma abordagem prática para a construção de modelos autorregressivos integrados e de médias móveis, designados por *Autorregressive Integrated Moving Average Models* (ARIMA).

Uma das principais características das séries temporais é a importância da ordem em que as observações são registadas. Regra geral observações sucessivas observadas ao longo tempo são correlacionadas. Estas observações podem ser registadas de forma contínua ou de forma discreta, no tempo. A priori, a medida de diversas quantidades, como a temperatura do ar, podem ser feitas de forma contínua mas muitas vezes as medidas são recolhidas de forma discreta no tempo, ou seja, recolhidas em determinados momentos de tempo, com igual espaçamento (dias, semanas, meses, anos, etc). Para além disso, as séries temporais também podem ser classificadas em univariadas, se são constituídas por apenas uma só variável, e em multivariadas quando, em cada instante, se observam mais do que uma variável. Na prática, os dados registados de forma contínua no tempo sofrem uma discretização em intervalos de tempo iguais para poderem ser trabalhados, nomeadamente em termos computacionais; tendo sempre em consideração que o intervalo desta amostragem seja pequeno, evitando a perda significativa de informação. Em todas as áreas, existem fenómenos cuja análise do comportamento ao longo do tempo é importante, bem como, perceber o como essas variações se comportam e qual é o desenvolvimento desses fenómenos ao longo do tempo. Assim, na atividade de determinada indústria, na qualidade de sono de uma pessoa, nas flutuações de preços, fenómenos meterológicos, entre muitos outros, a análise, estimação e previsão de séries temporais é fundamental. Na

meteorologia a análise das séries temporais é especialmente importante. A análise temporal permite melhorar o conhecimento de fenómenos meteorológicos tornando posível uma melhor perceção do tempo para determinada área geográfica ou a possibilidade de prever fenómenos extremos. A análise de dados meteorológicos é útil em áreas como a Agricultura, no controlo de trânsito, nas mudanças climáticas, em cálculos estruturais na Engenharia, na estimação de recursos naturais, entre muitos outros.

3.1 Objetivos da Análise Temporal

Considera-se como principal objetivo da análise de séries temporais, o desenvolvimento de modelos matemáticos que expliquem/descrevam de forma plausível os dados de um determinado fenómeno, Shumway e Stoffer (2017).

Tendo em conta todas as características de uma série temporal, existem imensos propósitos na sua utilização. Chatfield (2000) descreve que os principais objetivos passam pela:

a) Análise descritiva dos dados: descrever o comportamento dos fenómenos associados à série, de forma a obter e compreender o mecanismo que a gerou, através da análise dos dados usando estatísticas sumárias e/ou métodos gráficos;

b) Modelação: encontrar um modelo estatístico adequado para descrever a evolução da série temporal, seja ele um modelo univariado ou um modelo multivariado;

c) Previsão: prever o comportamento futuro da série, ou seja, estimar os valores futuros da série. Estas previsões podem ser a 1-passo, previsões para a observação seguinte ou previsões a multi-passos, para várias observações seguintes;

d) Controlo: previsões adequadas permitem que sejam aplicadas medidas para controlar um determinado processo, sendo úteis em processos industriais, económicos, entre outros.

3.1.1 Componentes de uma Série Temporal

Chatfield (2000) descreve que, de maneira resumida, a decomposição da variação de uma série temporal pode ser efetuada considerando quatro componentes. A componente Tendência (T), a componente sazonal (S), a componente cíclica (C) e

a componente irregular/residual (I). Estas componentes podem ser combinadas de maneiras diferentes, de forma multiplicativa ou aditiva, respetivamente, por

$$Y_t = T \times C \times S \times I, \tag{3.1}$$

$$Y_t = T + C + S + I. \tag{3.2}$$

Um modelo aditivo é apropriado quando a magnitude das oscilações sazonais não varia com o nível da série. Por sua vez, se estas oscilações aumentam ou diminuem proporcionalmente com a tendência da série, então um modelo multiplicativo é o mais adequado (Wheelwright et al., 1998).

Por vezes, a transformação dos dados é necessária, nomeadamente a logarítmica, de forma a converter um modelo multiplicativo num modelo aditivo, i.e.,

$$\log Y_t = \log T_t + \log S_t + \log I_t.$$

Desta forma, constrói-se um modelo multiplicativo por ajustamento de um modelo aditivo ao logaritmo dos dados, que não deve ser aplicado a séries temporais de valores negativos ou nulos. Os modelos de decomposição aditivo e multiplicativo não são as únicas formas de decomposição de séries temporais. Estes modelos podem ser misturados, originando outros modelos que incluem relações tanto aditivas como multiplicativas, e.g., um *modelo multiplicativo com erros aditivos*, i.e.,

$$Y_t = T_t \times S_t + I_t.$$

A decomposição de séries temporais pode ser feita de forma totalmente automática e simples no ambiente R, como por exemplo a função *decompose*, que permite avaliar as diversas componentes separadamente e, assim, ajudar a identificar o comportamento individual das mesmas.

A **Tendência (T)** é o padrão, a inclinação, que uma série temporal apresenta ao longo do tempo. A tendência pode ser positiva ou negativa, linear ou não linear, dependendo se apresenta um padrão crescente ou um padrão decrescente, ao longo do tempo. Se uma série temporal não exibir um comportamento crescente ou decrescente, a série diz-se estacionária em média.

A **Sazonalidade (S)** são as flutuações regulares que se repetem, aumentos e diminuições, e ocorrem na série durante um período de tempo específico. Estas flutuações devem-se a fatores sazonais, que apresentam padrões de comportamento semelhantes e que surgem em muitas séries temporais. Geralmente, este período é

anual mas também pode ser semanal, mensal, trimestral, ou seja, padrões semelhantes observados em épocas do ano. Chatfield (2003) refere como exemplo prático, o padrão de vendas de gelado é sempre alto no Verão. A sazonalidade pode ser aditiva, se não depende do nível da série, ou multiplicativa , quando é proporcional ao nível. O ajustamento sazonal aplica-se quando se pretende remover esta fonte de variação na série que pode estar a omitir outras componentes relevantes, como a tendência.

A **Componente Cíclica (C)** é a componente representada pelos movimentos oscilatórios de longo prazo, com periodicidade não regular. Esta componente é difícil de prever, uma vez que, é recursiva e não periódica, ou seja, os tempos entre os picos é não regular. Esta componente, frequentemente, é ignorada para séries "curtas".

A **Componente Irregular/Aleatória (I)** é uma componente irregular de natureza aleatória também designada por ruído branco ou resíduo. Esta componente representa tudo o que não se consegue modelar ou definir.

3.2 Processos Estocásticos

Um processo estocástico é um fenómeno estatístico que evolui no tempo de acordo com leis probabilísticas. O processo estocástico é a extensão do conceito de variável aleatória.

Definição 1 *Um processo estocástico é qualquer coleção ou família de variáveis aleatórias definidas por* $\{Y(t), t \in T\}$*, em que* $Y(t)$ *é uma variável aleatória (ou conjunto de variávei aleatórias) com contradomínio* S*, denominado por espaço de estados.* T *é um conjunto de índices ordenados representado o tempo e é denominado por espaço de parâmetros (Alpuim, 2003).*

Formalmente, define-se uma série temporal como o conjunto de observações de um processo estocástico $\{Y(t), t \in T\}$, nos instantes $t_1, t_2, \ldots, t_n$. As observações são observadas em intervalos de tempo regulares, considerando-se o t inteiro (i.e., $t = 0, \pm 1, \pm 2, \ldots$). Sendo que, para $T = \mathbb{Z}$ ou $T = \mathbb{N}$ o processo é de tempo discreto. Para $T = \mathbb{R}$ ou $T = \mathbb{R}^+$ o processo é de tempo contínuo.

Existem duas formas de caracterizar um processo estocástico. Uma das formas é especificar a distribuição de probabilidade conjunta das suas n variáveis aleatórias $(Y(t_1), \ldots, Y(t_n))$ para todos os inteiros n e pontos $t_1, \ldots, t_n$. Mas esta maneira de caracterizar o processo estocástico é de extrema dificuldade e, na prática, não viável.

A alternativa passa pela caracterização do processo estocástico a partir das funções determinísticas a ele associadas. Estas funções determinísticas são particularmente importantes para a caracterização do comportamento do processo, ou seja, os

momentos do processo. Em particular, descrever o primeiro (valor médio) e segundo momentos (função de autocovariância) designados por

– $\mu(t) = E[Y(t)]$, para $t = 0, \pm1, \pm2, \ldots,$

– $\gamma(t_1, t_2) = E[(Y(t_1) - \mu(t_1))(Y(t_2) - \mu(t_2))]$, $t = 0, \pm1, \pm2, \ldots$.

A variância $\sigma^2(t) = Var[Y(t)]$ para $t = 0, \pm1, \pm2, \ldots$, é um caso particular da função autocovariância quando $t_1 = t_2$.

Na inferência estatística sobre a estrutura de um processo estocástico, baseada nos valores observados é usual assumir-se algumas suposições. A mais importante das suposições é a estacionariedade. Os processos estocásticos dividem-se em estacionários e não estacionários.

3.2.1 Processos estacionários

A ideia mais simples de estacionariedade é a de que a lei de probabilidade subjacente ao comportamento do processo não se altera ao longo do tempo. Um processo estacionário é um processo estocástico, em que os valores da variável aleatória se distribuem no tempo, em torno de um valor médio constante e a independência entre os valores observados em diferentes instantes t é assumida. Neste tipo de processos a média e a variância são constantes, não sendo, assim, necessário o conhecimento do valor na origem da série (Alpuim, 2003).

Para um processo estacionário $\{Y(t), t \in T\}$ com variância finita, $\forall t \in T$

Definem-se dois tipos de estacionariedade. Estacionariedade forte ou estritamente estacionário e estacionariedade fraca ou de segunda ordem (Shumway e Stoffer (2017)).

Definição 2 *Um processo estacionário $\{Y(t), t \in T\}$ diz-se estritamente estacionário (ou fortemente estacionário) se a distribuição conjunta de $(Y(t_1), ..., Y(t_n))$ é igual à distribuição conjunta de $(Y(t_1 + \delta), ..., Y(t_n + \delta))$ qualquer que seja o n-úplo $(t_1, ..., t_n)$ e para qualquer δ, i.e.,*

$$F_{(Y(t_1),...,Y(t_n))}(y_1, ..., y_n) = F_{(Y(t_1+\delta),...,Y(t_n+\delta))}(y_1, ..., y_n), \text{ em todos os pontos } (y_1, ..., y_n). \tag{3.3}$$

A condição de estacionaridade restrita implica o conhecimento de todas as distribuições marginais, o que torna a verificação muito difícil na prática, além disso, não é aplicável à maioria das séries temporais (Alpuim, 2003). Desta forma, é usual aplicar

a definição mais simples que impõe condições apenas nos dois primeiros momentos da série temporal (estacionaridade de segunda ordem).

Definição 3 *Um processo $Y(t), t \in T$ diz-se estacionário de segunda ordem (ou fracamente estacionário) se todos os momentos até à segunda ordem de $(Y(t_1), ..., Y(t_n))$ existem e são iguais aos momentos correspondentes até à segunda ordem de $(Y(t_1 + \delta), ..., Y(t_n + \delta))$. Desta forma, um processo $Y(t)$ é fracamente estacionário (de segunda ordem) se:*

1. *$\mu(t) = \mu$, o valor médio não depende de t;*
2. *$\sigma^2(t) = \sigma^2$, a variância não depende de t;*
3. *$Cov[Y(t_1), Y(t_2)] = \gamma(|t_2 - t_1|)$, a covariância depende apenas do desfazamento $t_2 - t_1$.*

Um processo estritamente estacionário, com variância finita é também um processo estacionário de segunda ordem. A recíproca pode não se verificar.

3.2.2 Funções de autocovariância, autocorrelação e autocorrelação parcial

A função de autocovariância é a função de covariância entre realizações de um processo estocástico observadas em diferentes horizontes de tempo e desfasadas em k unidades de tempo (*lag*).

Definição 4 *A função de autocovariância, para um processo estacionário, é definida por*

$$\gamma_k = Cov[Y_t, Y_{t+k}] = E[(Y_t - \mu)(Y_{t+k} - \mu)].$$

Se o processo é de tempo contínuo, γ_k é definida para $k \in \mathbb{R}$, se processo é de tempo discreto γ_k é definida para $k \in \mathbb{Z}$.

O estimador da função de autocovariância para um processo estacionário de segunda ordem é dado por

$$\hat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y}).$$

Quanto maior o valor de k, maior o desvio entre o valor estimado $\hat{\gamma}_k$ e o valor da função γ_k. Desta forma, deve-se estimar $\hat{\gamma}_k$ apenas para os primeiros $\frac{n}{4}$ valores de k (Chatfield, 2000).

A função de autocovariância deve respeitar as seguintes propriedades:

1. $\gamma_0 = Cov[Y_t, Y_t] = Var[Y_t] = \sigma^2$;

2. $\gamma_k = \gamma_{-k}$;

3. $|Y_k| \leq \gamma_0$ como consequência da desigualdade de *Cauchy-Schwarz*, $|E[XY]| \leq \sqrt{E(X^2)E(Y^2)}$;

4. É semidefinida positiva, $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\gamma(|t_i - t_j|) \geq 0$, onde $\alpha_1, ..., \alpha_n$ representam um conjunto de números reais e $t_1, ..., t_n$ os instantes de tempo.

A função de autocorrelação mede a correlação entre realizações de um processo estocástico observadas em diferentes horizontes de tempo e desfasadas em k unidades de tempo (*lag*). Desta forma, interpreta-se ρ_k como uma medida da semelhança entre cada realização e a mesma realização deslocada k unidades de tempo iguais (Murteira et al., 2000).

Definição 5 *A função de autocorrelação (FAC), para um processo estacionário, é definida por*

$$\rho_k = Corr[Y_t, Y_t + k] = \frac{Cov[Y_t, Y_{t+k}]}{\sqrt{Var[Y_t]Var[Y_{t+k}]}} = \frac{Cov[Y_t, Y_{t+k}]}{Var[Y_t]} = \frac{\gamma_k}{\gamma_0}.$$

O estimador da função de autocorrelação para um processo estacionário de segunda ordem é representado pela seguinte expressão

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}.$$

A função de autocorrelação deve respeitar as seguintes propriedades:

1. $\rho_0 = Corr[Y_t, Y_t] = 1$;

2. $\rho_k = \rho_{-k}$;

3. $|\rho_k| \leq 1$, como consequência da desigualdade de *Cauchy-Schwarz*, $|E[XY]| \leq \sqrt{E(X^2)E(Y^2)}$;

4. É semidefinida positiva, $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\rho(|t_i - t_j|) \geq 0$, onde $\alpha_1, ..., \alpha_n$ representam um conjunto de números reais e $t_1, ..., t_n$ os instantes de tempo.

Definição 6 *A função de autocorrelação parcial (FACP), para um processo estacionário ou conjunto de autocorrelações parciais de desfasamento (lag) k é dado por* $\{\phi_{kk} : k = 1, 2, \ldots\}$ *onde*

$$\phi_{kk} = Corr[Y_t, Y_{t+k}|Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}] = \frac{|P_k^*|}{|P_k|},$$

e P_k^* é a matriz $k \times k$ de autocorrelações onde a última coluna é substituída por $[\rho_1 \ \rho_2 \ \ldots \ \rho_k]^T$. A matriz P_k é dada por

$$P_k = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & 1 \end{bmatrix}.$$

obtendo-se as seguintes propriedades:

1. $\phi_{11} = \rho_1$;
2. $\phi_{22} = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2}$;
3. $\phi_{33} = \frac{\rho_3(1-\rho_1^2)+\rho_1(\rho_1^2+\rho_2^2-2\rho_2)}{(1-\rho_2)(1+\rho_2-2\rho_1^2)}$.

O correlograma teórico é a representação gráfica de ρ_k em função de k. A análise do correlograma é uma ferramenta de extrema utilidade na identificação de várias caraterísticas de uma série temporal e constitui um auxiliar importante na escolha do modelo que lhe é mais adequado. O aumento de k traduz-se no decréscimo de ρ_k e, consequentemente, de γ_k , ou seja, excetuando casos especiais, o aumento de k implica o decréscimo de γ_k e de ρ_k, isto é

$$k \to +\infty, \ \ \rho_k \to 0 \text{ e } \rho_k \to +\infty.$$

3.2.3 Ruído Branco

Definição 7 *O processo estacionário designado de ruído branco, $\{\varepsilon_t, t \in \mathbb{Z}\}$, o qual é definido como a sequência de variáveis aleatórias independentes e identicamente distribuídas, de média e variância constantes. Desta forma, um processo estocástico é um ruído branco se satisfaz as condições:*

1. $E(\varepsilon_t) = \mu_\varepsilon (usualmente, \mu_\varepsilon = 0)$;

2. $Var(\varepsilon_t) = \sigma_\varepsilon^2$;

3. $Cov(\varepsilon_t, \varepsilon_{t+k}) = \gamma_k = 0, \quad k = \pm 1, \pm 2, \ldots$.

Como a média e a função de autocovariância não dependem do tempo, o processo é estacionário de segunda ordem.

Adicionalmente, se as variáveis aleatórias seguem uma distribuição Normal, i.e., $\varepsilon_t \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$, então o processo ε_t é designado por ruído branco Gaussiano. Este processo é muito útil na construção de modelos estocásticos, embora dificilmente se observe em séries reais. Um bom modelo de previsão deverá produzir erros de previsão com comportamento semelhante ao de um ruído branco, dada a imprevisibilidade inerente ao ruído branco.

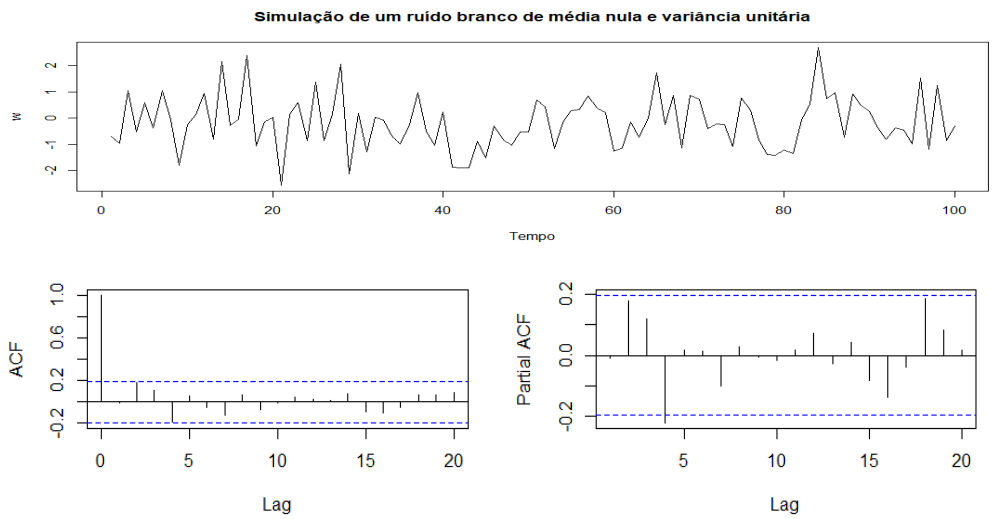


Figura 3.1: Simulação de um ruído branco e repectivas FAC e FACP empíricas.

Na Figura 3.1 encontra-se representada uma trajetória de um processo de ruído branco e as respectivas FAC e FACP estimadas.

3.3 Modelos para Processos Estacionários

3.4 Processos Médias Móveis (MA)

Seja ε_t um processo ruído branco de média nula e variância constante.

Definição 8 *O processo* $\{Y_t,\ t \in T\}$ *é considerado um processo de médias móveis*

(MA), com ordem q, se

$$Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{q-1} + \mu_t,$$

onde μ_t *é a média da série temporal* Y_t, $\theta_i \in \mathbb{R}, i = 1, \ldots, q$. *O modelo é representado por MA(q). A média de* Y_t *é dada por*

$$E(Y_t) = E(\varepsilon_t) + \sum_{j=1}^{q} \theta_j E(\varepsilon_{t-j}) = \mu_t.$$

Se $\mu_t = 0$, *então* $E(Y_t) = 0$, *então a variância é dada por*

$$Var(X_t) = Var(\varepsilon_t) + \sum_{j=1}^{q} \theta_j^2 Var(\varepsilon_{t-j}) = (1 + \theta_1^2 + \ldots + \theta_q^2)\sigma_\varepsilon^2.$$

A função de autocovariância (com $\theta_0 = 1$*) é dada por*

$$\gamma_k = Cov(Y_t, Y_{t+k}) = \begin{cases} 0\,, & k > q \\ \sigma_\varepsilon^2 \sum_{j=0}^{q-k} \theta_j\theta_{j+k} & k = 0, \ldots, q. \end{cases}$$

Como a média e a variância são constantes, a função de autocorrelação (FAC) para $k \geq 0$ *é dada por*

$$\rho_k = \begin{cases} 1 & , k = 0 \\ \frac{\sum_{j=0}^{q-k} \theta_j\theta_{j+k}}{\sum_{j=0}^{q} \theta_j^2} & , k = 1, \ldots, q. \\ 0 & , k > q + 1 \end{cases}$$

O processo é estacionário de segunda ordem, com média e variância constantes e a autocovariância independente do instante t, para qualquer valor possível de $\theta_1, \ldots, \theta_q$. Se a distribuição do processo aleatório, ε_t, for Normal (Gaussiana) diz-se que o processo é estritamente estacionário.

Utilizando o operador atraso B, definido por

$$B^j Y_t = Y_{t-j} \;\; \forall j = 0, \ldots, q.$$

Define-se o polinómio $\Theta(B)$, de ordem q em B, por

$$\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q.$$

Assim,

$$Y_t = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)\varepsilon_t = \Theta_q(B)\varepsilon_t.$$

Existem 2^q modelos possíveis para $\Theta(B) = 0$ que terão a mesma função de autocorrelação, mas apenas um terá raízes da equação fora do círculo unitário. Ou seja, o processo de Médias Móveis de ordem q ($MA(q)$) é invertível, quando todas as q soluções $(\theta_1, \ldots, \theta_q)$ de $\Theta(B) = 0$ sejam $|\theta_i| > 1$, $i = 1, \ldots, q$ (Metcalfe e Cowpertwait (2009)).

3.5 Processos Autorregressivos (AR)

Definição 9 *Um processo estacionário $\{Y_t, t \in T\}$ é considerado um processo autorregressivo (AR), com ordem p, se*

$$Y_t = \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \varepsilon_t \tag{3.4}$$

onde, ε_t é um processo de ruído branco e ϕ_i , $i = 1, \ldots, p$, são constantes reais.

O modelo é representado por AR(p). Os processos autorregressivos são modelos úteis para explicar séries temporais em que o valor da variável, para um determinado instante t, pode ser explicado por p valores passados, $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ adicionado de um erro aleatório ε_t associado. O processo ε_t é discreto puramente aleatório com média $\mu_\varepsilon = 0$ e variância σ_ε^2 constante. Os parâmetros autorregressivos são valores reais, i.e., $\phi_i \in \mathbb{R}$, com $i = 1, \ldots, p$.

Utilizando o operador atraso, B, tem-se

$$Y_t = \phi_1 B + ... + \phi_p B^p + \varepsilon_t \iff (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)Y_t = \varepsilon_t,$$

que pode ser reescrito como

$$\Phi(B)Y_t = \varepsilon_t \iff Y_t = \frac{\varepsilon_t}{\phi(B)} = \Phi(B)\varepsilon_t.$$

Como no caso do modelo MA, existem 2^p soluções possíveis para $\Phi(B) = 0$ e apenas um modelo tem todas as raízes desta equação fora do círculo unitário, tornando o modelo invertível. O polinómio $\Phi(B)$ é chamado de polinómio característico do processo e as suas raízes determinam quando o processo é estacionário. O processo é estacionário para o modelo $\Phi(B) = 0$ se apresenta as suas raízes fora do círculo unitário. Assim, o processo $AR(p)$ é invertível, quando todas as p soluções $(\alpha_1, \ldots, \alpha_p)$

de $\Phi(B) = 0$ sejam $|\phi_i| > 1$, $i = 1, \ldots, p$ (Metcalfe e Cowpertwait (2009)).

Para um processo estacionário $AR(p)$ tem-se a

- função de autocovariância: $\gamma_k = E(Y_t, Y_{t+k}) = \sum_{i=1}^{p} \phi_i \gamma_{k-i} \ , k > 0$, e a
- função de autocorrelação: $\rho_k = \sum_{i=1}^{p} \alpha_i \rho_{k-i} \ , k > 0$.

3.6 Processos Autorregressivos e de Médias Moveis (ARMA)

Os modelos autorregressivos e de médias móveis designam-se por ARMA(p,q) e incluem tanto termos autorregressivos como de médias móveis.

Definição 10 *Um processo estacionário $\{Y_t, t \in T\}$ é considerado um processo autorregressivo e de médias móveis (ARMA), com ordem p e q, se*

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-1}, \tag{3.5}$$

ou a equação

$$\Phi_p(B) Y_t = \Theta_q(B) \varepsilon_t, \tag{3.6}$$

onde ε_t é um ruído branco de média nula, independente de Y_{t-k} para todo o $k \geq 1$, $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ e $\Theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ são os polinómios autorregressivo e de médias móveis de ordens p e q, respetivamente e ϕ_i, $i = 1, \ldots, q$ e θ_i, $i = 1, \ldots, p$ são constantes reais.

3.7 Modelos para Processos não Estacionários

A maioria das séries temporais, na prática, é não estacionária. Quanto tal acontece, é necessário remover dos dados as fontes de variação não estacionárias, tais como a tendência e sazonalidade, de forma a permitir o ajustamento de um modelo estacionário. Se uma série é não estacionária na variância é necessário proceder à sua estabilização. Para tal, pode utilizar-se um método de transformação paramétrica, conhecido como transformação de Box-Cox, sendo as mais comuns a transformação inversa, a transformação logarítmica e a transformação quadrática. Se a série temporal observada for não estacionária em média, pode aplicar-se uma (ou várias) diferenciação (regular) à mesma. Assim, $\nabla^d Y_t = (1 - B)^d Y_t$, com $d \geq 1$, é a série estacionária depois de diferenciada d vezes, substituindo Y_t por $\nabla^d Y_t$ na equação (3.6),

obtém-se um modelo designado por ARIMA capaz de descrever séries não estacionárias. Este tipo de modelo é designado de modelo "integrado", uma vez que o modelo estacionário ajustado aos dados diferenciados deve ser somado ou "integrado". Estes modelos podem, também, à semelhança dos modelos ARMA, ser generalizados para incluir termos sazonais, dando origem aos modelos SARIMA.

3.8 Processos Autorregressivos Integrados e de Médias Móveis (ARIMA)

Uma das abordagens mais utilizadas em séries temporais é a modelação autorregressiva integrada de médias móveis (ARIMA), popularizada por Box e Jenkins (1970) e Box e Tiao (1975) para previsão de séries temporais. Esta abordagem representa a combinação de três modelos matemáticos, usando modelos autorregressivos (AR), integrados (I) e de médias móveis (MA) para séries temporais. "Autorregressivo" porque representa os *lags* da série diferenciada que aparece na equação de previsão, "médias móveis" são os *lags* dos erros de previsão e "integrados" referentes aos níveis de diferenciação que tornam a série temporal estacionária.

Esta metodologia reune diferentes métodos já existentes de maneira a produzir modelos capazes de descrever o comportamento de uma grande variedade de séries temporais de forma parcimoniosa e cujos resultados têm grande poder preditivo, quando comparados com outros métodos (Morrettin e Toloi (2006)). A ideia é a de que cada valor da série temporal pode ser explicado pelos seus valores anteriores, ou seja, de que existe uma certa dependência entre valores próximos no tempo. Desta forma, o modelo é representado por ARIMA(p, d, q), onde p é parâmetro autorregressivo (número de *lags* da série diferenciada), d é o nível de diferenciação que torna a série temporal estacionária em média, chamado de parâmetro de integração. O número de *lags* dos erros de previsão é chamado de parâmetro de médias móveis q.

Os modelos $AR(p)$, $MA(q)$ e $ARMA(p, q)$ impõem a condição de estacionaridade, consequentemente isto implica que séries temporais que não sejam estacionárias sejam alteradas, removendo a fonte de não estacionaridade, seja a tendência ou a sazonalidade. Esta alteração, é resultado da diferenciação que reduz a não estacionaridade integrada no modelo (Ehlers, 2009). Neste caso temos a diferenciação que reduz a não estacionariedade integrada no modelo

$$W_t^d = \nabla^d Y_t = (1 - B)^d Y_t.$$

O processo autorregressivo ARIMA(p, d, q) é dado pela equação

$$\Phi(B)(1-B)^d Y_t = \Theta(B) I_t.$$

A diferenciação transforma a série original numa série estacionária, e a condição de invertibilidade também é imposta a este tipo de modelos. Esta equação é adequada para modelos não sazonais, eficaz para séries que possuam uma tendência que seja retirada pela diferenciação.

Assim, um modelo ARIMA(p, d, q) pode explicar a dependência temporal de várias formas:

a) A série temporal é diferenciada com ordem d para torná-la estacionária. Se $d = 0$, as observações são modeladas diretamente, se $d = 1$, as diferenças entre as observações consecutivas são modeladas;

b) A dependência do tempo do processo estacionário Y_t é modelada pela inclusão de modelos autorregressivos p;

c) O termo de médias móveis q considera a observação de erros anteriores.

Com a combinação destes três modelos, obtém-se o modelo ARIMA. Assim, a forma geral dos modelos ARIMA é dada por

$$Y_t = c + \sum_{i=1}^{p} \phi_i \, Y_{t-i} + \sum_{j-1}^{q} \theta_j \, \varepsilon_{t-j}, \tag{3.7}$$

onde Y_t é um processo estocástico estacionário, c a constante que determina o nível da série temporal, ε_t é o erro ou termo de perturbação do ruído branco, ϕ_i o coeficiente autorregressivo, com $i = 1, \ldots, p$ e θ_j o coeficiente de médias móveis, com $j = 1, \ldots, q$. Para uma série temporal sazonal, estas etapas podem ser repetidas de acordo com o período do ciclo, independentemente do intervalo de tempo.

3.8.1 Processo Autorregressivo Integrado e de Médias Móveis Sazonal (SARIMA)

Em algumas séries não estacionárias, as variações induzidas pelos fatores sazonais podem dominar as variações da série original. Para estas séries temporais que possuem uma componente sazonal, utilizam-se os modelos ARIMA sazonais, designados por modelos autorregressivos integrados e de médias móveis sazonal SARIMA.

Os Modelos ARIMA sazonais são obtidos pela inclusão no modelo ARIMA(p, d, q) da componente sazonal obtendo um modelo SARIMA de ordem $(p, d, q)(P, D, Q)_s$.

Nestes modelos, a estacionariedade é obtida através de múltiplas diferenciações do padrão sazonal S. A diferenciação sazonal é representada por ∇_S^D em que D é o número de diferenciações sazonais. Os dois tipos de diferenciação podem ser usados simultaneamente $\nabla_d \nabla_S^D$.

O modelo SARIMA é representado da seguinte forma

$$\Phi(B)\mathbf{\Phi}(B^S)\nabla_d \nabla_S^D Y_t = \Theta(B)\mathbf{\Theta}(B^S)e_t.$$

Os valores das ordens d e D removem, respetivamente, a tendência e a sazonalidade. Geralmente, para tornar a série temporal estacionária estes valores não são superiores a 1. As restantes ordens, p, P, q e Q, podem ser determinadas com base nas funções de autocorrelação e autocorrelação parcial.

O método Box-Jenkins é o tradicionalmente utilizado, trata-se de um processo iterativo de três etapas (identificação, estimação e diagnóstico) que identifica o modelo através da análise da representação gráfica dos dados e respetivas FAC e FACP empíricas. Se a série temporal não é estacionária, esta deve ser diferenciada gradualmente até que seja considerada estacionária. Assim, o valor d do modelo é obtido.

Desta forma, existem três etapas básicas no processo para se obter um modelo ARIMA ou SARIMA:

1. Identificação dos modelos possíveis

   a) Análise da função de autocorrelação (FAC) e função de autocorrelação parcial (FACP) para identificar os termos não sazonais;

   b) Verificar se a série temporal é estacionária. Caso não o seja, a série deverá sofrer uma transformação como a diferenciação para a tornar estacionária. Para os modelos autorregressivos(AR) é mandatório que o módulo das raizes do polinómio AR sejam superiores à unidade. No caso dos modelos de média móveis (MA), é mandatório que sejam invertíveis e para tal, também é necessário que as raízes do polinómio (MA) se encontrem fora do círculo unitário;

   c) Ao diferenciar as séries temporais sazonais percebe-se a ordem de diferenciação necessária para tornar a série estacionária e identificar o modelo ARMA apropriado.

2. Estimação

Nesta etapa, os parâmetros do modelo são estimados. Atualmente, com todos os avanços das tecnologias, a estimação dos parâmetros é realizada com o recurso a ferramentas computacionais como *softwares* STATA, SPSS e R, entre outros.

3. Avaliação

   Esta etapa envolve um processo rigoroso de avaliação por aplicação de testes estatísticos para cada um dos modelos. Modelos diferentes podem comportar-se de forma semelhante e as formulações alternativas devem ser retidas para posterior avaliação da capacidade de predição e previsão dos modelos.

Existem diversos métodos para avaliar os modelos. Os mais utilizados são a representação gráfica dos resíduos do modelo estimado no sentido de detetar eventuais valores que possam afetar o modelo, ou existência de problemas de autocorrelação e a representação da FAC e FACP dos resíduos para verificar a adequabilidade do modelo. Os resíduos do modelo deverão ter uma distribuição Normal e devem apresentar um comportamento idêntico ao de um ruído branco e satisfazer o pressuposto de não correlação.

Estas condições (Normalidade e não correlação) podem ser analisadas recorrendo às seguintes formas:

- Testar a normalidade recorrendo a métodos gráficos e testes estatísticos. Quanto às representações gráficas, a representação do histograma e o $QQ-plot$, o histograma deverá apresentar um comportamento da função densidade da distribuição Normal e o gráfico $QQ-plot$, representação gráfica dos quantis reais e dos teóricos, este deve apresentar um conjunto de pontos que se localizem sobre um reta. Os testes estatísticos tradicionais são os testes de Shapiro-Wilk e o teste de Kolmogorov-Smirnov.

- O teste Ljung-Box é aplicado aos resíduos da série temporal, após o ajustamento de um modelo. São analisadas as h autocorrelações dos resíduos e estas não devem ser significativamente diferentes de zero. As hipóteses de teste são

$$H_0 : \rho_1 = \rho_2 = ... = \rho_h = 0 \quad vs \quad H_1 : \exists k, \in \{1, \ldots, h\} : \ \rho_k \neq 0$$

e a estatística de teste Q_h é definida por

$$Q_h = n(n+2) \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{h-m},$$

e segue aproximadamente uma distribuição Qui-Quadrado com $h - m$ graus de liberdade, $\hat{\rho}_k$ é a autocorrelação e m é o número de parâmetros estimados.

Teoricamente, seria possível ter qualquer período de sazonalidade numa série, uma vez que o número de parâmetros a serem estimados não depende da ordem sazonal. No entanto, a diferenciação sazonal de ordem muito alta não faz muito sentido, note-se o exemplo para dados diários, estes envolvem a comparação do que aconteceu hoje com o que aconteceu há exatamente um ano e não existe nenhuma restrição de que o padrão sazonal é bom. Na prática os modelos SARIMA não têm capacidade para lidar com séries temporais com longos períodos sazonais $s \geq 200$. As versões sazonais do modelos ARIMA são utilizados em casos com períodos sazonais mais pequenos, como por exemplo para um período de 12 meses.

Para este tipo de casos, De Livera et al. (2011) sugerem a abordagem da regressão dinâmica com termos de Fourier, onde o padrão sazonal é modelado usando termos de Fourier com dinâmicas de séries temporais de curto prazo permitidas no erro ARMA, que serão apresentados no capítulo seguinte.

Capítulo 4

Metodologias

Existem diversas abordagens para modelar séries temporais com um padrão sazonal único. Entre estes, existem os modelos de regressão com erros correlacionados (Alpuim e El-Shaarawi (2009)), modelos de alisamento exponencial (Winters, 1960), modelos ARIMA sazonais (Box e Jenkins (1970)), os modelos de espaço de estados (Harvey e Fernandes (1989)) e as inovações aos modelos de espaço de estados (Hyndman et al., 2008).

4.1 Modelos de Regressão com Erros Correlacionados

Uma abordagem importante na modelação de séries temporais, particularmente de séries meteorológicas, é o ajustamento de modelos de regressão. Os modelos de regressão são originalmente baseados em modelos lineares, mas que permitem que componentes de tendência e sazonalidade sejam incorporados nos modelos. A componente tendência é considerada determinística e pode ser explicada através de funções polinomiais do tempo. No caso da componente sazonal, esta pode ser modelada através de uma descrição qualitativa do padrão sazonal utilizando variáveis indicatrizes ou utilizando funções harmónicas, com ondas sinusoidais. Para os modelos de regressão serem válidos, os termos do erro devem ser não correlacionados, apresentarem média nula e variância constante (i.e., um ruído branco). O modelo de regresão linear simples traduz a relação entre uma variável independente e uma variável dependente da seguinte forma

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \quad t = 1, 2, \ldots, n.$$

O modelo de regressão linear simples representa uma relação linear entre a variável dependente Y_t e a variável independente X_t, β_0 e β_1 são os parâmetros ou coeficientes de regressão e ε_t é o erro associado ao valor esperado de Y_t. O coeficiente β_0 representa a ordenada na origem e representa o valor esperado da variável de Y_t quando a variável explicativa é nula, β_1 representa o declive e é a variação do valor esperado de Y_t por cada acréscimo unitário na variável X_t. O modelo de regressão linear múltipla representa a relação entre uma ou mais variáveis independentes e uma variável dependente da seguinte forma

$$Y_t = \beta_0 + \beta_1 X_t^1 + \beta_2 X_t^2 \ldots + \beta_p X_t^p + \varepsilon_t.$$

O modelo de regressão assume uma relação linear entre a variável dependente Y_t e as p variáveis independentes $X^1, X^2, \ldots, X^p$, em que $\beta_0, \beta_1, \ldots, \beta_p$ são os parâmetros ou coeficientes de regressão e ε_t é o termo de erro (um ruído branco). Este termo captura possíveis fatores que influenciam Y_t, para além das p variáveis explicativas. O coeficiente β_0 representa o valor esperado da variável de Y_t quando as variáveis explicativas são simultaneamente nulas e $\beta_j, j = 1, \ldots, p$, é a variação do valor esperado de Y_t por cada acréscimo unitário na variável X^j, permanecendo constantes as restantes variáveis explicativas. A forma matricial do modelo é

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

onde,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{(n\times 1)}, X = \begin{bmatrix} 1 & X_1^1 & \cdots & X_1^p \\ 1 & X_2^1 & \cdots & X_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n^1 & \cdots & X_n^p \end{bmatrix}_{(n\times(p+1))}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{((p+1)\times 1)}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n\times 1)}.$$

Os modelos de regressão para séries temporais permitem a inclusão da informação de observações anteriores, mas não a inclusão de outras informações que também possam ser relevantes. Estes modelos de regressão permitem a inclusão de muitas informações relevantes das variáveis explicativas, mas não permitem captar a dinâmica subtil da série temporal que pode ser analisada, por exemplo, com os modelos ARIMA. Assim, surgem os modelos de regressão com erros correlacionados. Estes modelos permitem que os erros da regressão apresentem autocorrelação. Para representar esta mudança de perspectiva, o termo ε_t é substituido por d_t na equação.

Assim, os erros d_t seguem um modelo ARIMA, (Alpuim e El-Shaarawi (2008)), e o modelo é dado por

$$Y_t = \beta_0 + \beta_1 X_t^1 + \beta_2 X_t^2 \ldots + \beta_p X_t^p + d_t, \;\; t = 1, \ldots, n.$$

Como exemplo, se d_t segue um modelo ARIMA$(1,1,1)$ a equação é $(1-\phi_1 B)(1-B)d_t = (1+\theta_1 B)\varepsilon_t$, onde ε_t é um ruído branco.

Regressão com erros correlacionados e termos de Fourier

Em séries temporais, observações sucessivas usualmente estão correlacionadas. Desta forma, o ajustamento de modelos de regressão a dados de séries temporais pode levar a falhas nos pressupostos dos modelos, nomeadamente, a falha do pressuposto de indepedência dos erros. A existência de autocorrelação torna os estimadores de mínimos quadrados não eficientes, ainda que se mantenham, em geral, consistentes e não enviesados.

Alpuim e El-Shaarawi (2008) mostram que, sob determinadas condições da matriz de planeamento $\mathbf{X}$, os estimadores dos mínimos quadrados e os estimadores de máxima verosimilhança são assintoticamente equivalentes e totalmente eficientes. Desta forma, o conjunto das p variáveis explicativas no tempo t, $\mathbf{X}_t^T = (X_t^1, X_t^2, ..., X_t^p)$, deve verificar uma relação linear recursiva do tipo

$$\mathbf{X}_t = \Psi \mathbf{X}_{t-1}, \tag{4.1}$$

onde Ψ é uma matriz $p \times p$ de coeficientes constantes. Na maioria das vezes, as variáveis independentes usadas em modelos lineares verificam esta relação recursiva, como acontece nos casos de tendências lineares e polinomiais, ondas sinusoidais, variáveis indicatrizes, etc. (Alpuim e El-Shaarawi (2009)).

Referem ainda que para um modelo linear que verifique a condição (4.1), o vetor dos estimadores dos mínimos quadrados é assintoticamente Normal com matriz de variâncias e covariâncias dada por

$$Var[\hat{\beta}] = \sigma_a^2 \left[\Phi(B)\mathbf{X}^T \Phi(B)\mathbf{X}\right]^{-1}, \tag{4.2}$$

onde σ_a^2 representa a variância do ruído branco a_t, $\mathbf{X}$ representa a matriz de planeamento, isto é, a matriz que contém o conjunto das variáveis independentes, e $\Phi(B)\mathbf{X}$ representa a matriz onde cada elemento é obtido através da aplicação do operador $\Phi(B)$ ao elemento correspondente da matriz $\mathbf{X}$. O elemento genérico da

matriz $\Phi(B)\mathbf{X}$ é dado por

$$X_t^{j*} = \Phi(B)X_t^j = X_t^j - \phi_1 X_{t-1}^j - \cdots - \phi_k X_{t-k}^j, \tag{4.3}$$

para $j = 1, \ldots, p$ e $t = k+1, \ldots, n$. Na prática, os valores dos coeficientes autorregressivos, $\phi_1, \ldots, \phi_k$, e da variância do ruído branco, σ_a^2, são desconhecidos. No entanto, para amostras grandes (n elevado), estes podem ser substituídos por estimadores consistentes, o que permite realizar um teste assintótico à significância de cada variável explicativa, baseado na distribuição Normal (Alpuim e El-Shaarawi (2009)).

Assim, ao modelo de regressão é aplicada a série de Fourier para modelar o padrão sazonal usando termos de Fourier com dinâmicas de séries temporais de curto prazo permitidas no erro, considerando-se o seguinte modelo:

$$y_t = \beta_0 + \sum_{i=1}^{K} \left[\alpha_i sin\frac{2\pi it}{m} + \beta_i cos\frac{2\pi it}{m} \right] + d_t, \tag{4.4}$$

onde d_t é um processo ARIMA, α_i e β_i correspondem aos coeficientes de Fourier e m o período sazonal. O termo K representa o número de pares de senos e cosenos de Fourier (equivalente a incluir variáveis indicatrizes). Este termo é determinado fazendo variar o valor de i (usualmente $i = 1, \ldots, 5$) e escolhe-se o modelo com o valor de K que, otimiza um critério de avaliação dos modelos, geralmente baseado nas medidas do erro de previsão.

Deve ter-se em atenção que uma curva cosseno com um certo período, deve ser sempre incluída ou eliminada juntamente com a curva seno correspondente com o mesmo período e vice-versa. Esta exigência é feita de forma a que o modelo verifique a equação (4.1), garantindo que os estimadores dos mínimos quadrados sejam ótimos e que a fórmula para a variância dos estimadores (4.2) possa ser aplicada (Alpuim e El-Shaarawi (2009)).

Da mesma forma que a variância dos estimadores precisa de alterações quando existe autocorrelação nos erros, também os intervalos de previsão devem ser calculados tendo isso em consideração. Para tal, basta substituir na expressão original do intervalo de previsão a matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ por $(\Phi(B)\mathbf{X}^T\Phi(B)\mathbf{X})^{-1}$.

O intervalo de confiança a $(1-\alpha)100\%$ para a previsão pontual de Y_k tem limite inferior dado por

$$\hat{Y}_k - t_{1-\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_k^T((\Phi(B)\mathbf{X}^T\Phi(B)\mathbf{X})^{-1}\mathbf{x}_k\right)},$$

e limite superior dado por

$$\hat{Y}_k + t_{1-\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2\left(1+\mathbf{x}_k^T((\Phi(B)\mathbf{X}^T\Phi(B)\mathbf{X})^{-1}\mathbf{x}_k\right)},$$

onde $\hat{\sigma}^2$ é a variância estimada dos erros, e $t_{1-\frac{\alpha}{2};n-p-1}$ é o quantil $(1-\alpha/2)$ de uma distribuição t-Student com $n-p-1$ graus de liberdade.

Dokumentov et al. (2015) referem que as maiores vantagens desta abordagem são: a elevada capacidade do modelo permitir qualquer dimensão de sazonalidade para dados com mais do que um período sazonal, a inclusão de termos de Fourier com diferentes frequências, o padrão sazonal é suave para valores pequenos de K e os termos de curto prazo são fáceis de lidar com a aplicação dos erros ARMA.

A desvantagem desta abordagem, quando comparada com a abordagem tradicional dos modelos ARIMA, é a de que a sazonalidade é assumida como fixa. O padrão sazonal não pode mudar ao longo do tempo. Mas, na prática, a sazonalidade geralmente é constante, o que não se torna uma grande desvantagem, exceptuando as séries temporais muito longas.

4.2 Alisamento Exponencial

O alisamento exponencial é um método que foi proposto no final da década de 50 (Brown, 1959; Holt, 1957; Winters, 1960) e motivou alguns dos processos de previsão mais populares e utilizados para a formulação de modelos.

Os métodos de alisamento exponencial constituem um conjunto de métodos de previsão capazes de acompanhar mudanças no declive, no nível e no padrão sazonal. Desta forma, são técnicas estatísticas que descrevem a série baseando-se no estudo das mudanças que se produzem na mesma, em vez de construir um modelo matemático explicíto que tenha gerado os dados que a constituem. As previsões produzidas usando os métodos de alisamento exponencial são combinações ponderadas de observações passadas, com os pesos decaindo exponencialmente à medida que as observações se tornam mais distantes. Ou seja, quanto mais recente a observação, maior o peso associado. Desta forma, o nome alisamento exponencial reflete que no cálculo da previsão, os pesos de observações passadas diminuem exponencialmente, Makridakis et al. (2008).

Estes métodos são amplamente utilizados uma vez que geram modelos simples, que se adequam a uma grande variedade de séries temporais e que possuem uma capacidade de previsão confiável (Hyndman e Athanasopoulos, 2013), o que apresenta grande importância para diversas aplicações práticas.

O método de alisamento exponencial mais básico é o método de Alisamento Exponencial Simples para o qual apenas é necessária a estimação de um parâmetro. Como métodos mais complexos, destacam-se o método linear de Holt, que usa dois parâmetros, e o método de Holt-Winters com três parâmetros. O método de alisamento exponencial simples aplica-se a séries temporais sem tendência e não sazonais. Na presença de alguma destas componentes, tendência ou sazonalidade, os métodos mais adequados são os métodos lineares de Holt ou de Holt-Winters.

4.3 Alisamento Exponencial Simples

O alisamento exponencial simples (*Simple Exponential Smoothing* (SES)), desenvolvido por Brown (1959), é um método que usa valores exponencialmente alisados l_t, e $t = 1, 2, \ldots, n$. Estes valores, são designados de nível da série, representam a média ponderada dos dados, em que o peso das observações mais recentes é maior e decai exponencialmente para valores mais antigos, Morettin e Toloi (2006). O nível da série é dado por

$$l_t = \alpha Y_t + (1 - \alpha) l_{t-1},$$

onde, α é a constante de alisamento associada ao nível e $\in [0, 1]$. No processo de alisamento exponencial considera-se primeiro a componente tendência, que é uma combinação o nível (l_t) e do declive (b_t). A previsão é $\hat{Y}_{t+h|t}$ no instante t a h passos de y_t, a tendeência de previsão T_h e o parâmetro de amortecimento $0 < \phi < 1$. O método de alisamento exponencial simples assume que os dados têm origem de m modelo com média constante, ou seja, que os dados tenham sido gerados por um modelo da forma $y_t = \mu_t + \varepsilon_t$, onde ε_t é um ruído branco e o nível μ_t pode mudar lentamente com o tempo, no entanto, a constante μ fornece um modelo razoável da série temporal. Dado l_{t-1}, onde l_t é o estimador do nível no tempo t e y_t, o método de alisamento exponencial simples atualiza o estimador de nível através da seguinte equação

$$l_t = \alpha y_t + (1 - \alpha) l_{t-1},$$

onde $0 \leq \alpha \leq 1$ é o parâmetro de alisamento que determina o peso dado a cada uma das componentes para gerar previsões. Quanto maior o seu valor, maior o peso que a observação y_t recebe, ou seja, o estimador torna-se mais sensível a mudanças no nível. A previsão de h passos à frente é igual ao nível da última observação, ou seja

$$\hat{Y}_{t+h|t} = l_t, \quad h > 0$$

A vantagem deste método está na sua aplicação simples e flexível. É ainda possível adequar o valor de α para obter um melhor ajustamento. Uma das suas desvantagens, é não ser aplicável em séries que apresentem tendência ou sazonalidade.

4.4 Alisamento Linear de Holt

Holt (1957) desenvolveu um método que permite lidar com séries temporais com tendência, que é uma extensão do método de alisamento exponencial simples. Este método, designa-se por método linear de Holt, ou método de Holt. O método de Holt é recomendado para séries temporais que apresentam tendência linear (positiva ou negativa). Esta técnica apresenta semelhanças ao alisamento exponencial simples, com a atualização de duas componentes em cada período t, o nível e declive. Três equações caracterizam este modelo (Hyndman e Athanasopoulos, 2013):

$$\text{Nível: } l_t = \alpha Y_t + (1 - \alpha)(l_{t-1} + b_{t-1}),$$

$$\text{Declive: } b_t = \beta(l_t - l_{t-1}) + (1 - \beta) b_{t-1},$$

$$\text{Previsão: } \hat{Y}_{t+h} = l_{t+h}b_t, \quad h = 1, 2, \ldots$$

onde as constantes de alisamento α e $\beta \in]0, 1]$.

Este modelo apresenta as mesmas vantagens do modelo anterior, com o acréscimo da capacidade para efetuar modelação de séries temporais com tendência. Por outro lado, a desvantagem principal é a dificuldade em determinar os valores mais adequados para as constantes de alisamento.

4.5 Alisamento de Holt-Winters

Quando existe sazonalidade nos dados, os métodos descritos anteriormente não apresentam um bom comportamento. Neste caso, surge o método de alisamento de Holt-Winters. O método de Holt-Winters é uma extenção do método de Holt, proposta para os casos em que os dados apresentam não só tendência, mas também sazonalidade. Desta forma, o método compreende a equação de previsão e três equações de atualização uma para o nível, uma para o declive e outra para a sazonalidade. Existem dois métodos de Holt-Winters, o método Holt-Winters aditivo e o método de Holt-Winters multiplicativo, dependendo do tipo de sazonalidade (aditiva ou multiplicativa).

Holt-Winters Aditivo

O método de decomposição aditivo define-se por

$$Y_t = T_t + S_t + \varepsilon_t.$$

A preferência pelo método aditivo ocorre quando as variações sazonais são independentes do nível da série. As equações que constituem este método são:

$$\text{Previsão: } \hat{Y}_{t+h} = l_t + hb_t + s_{t-s+h_s^+}, \quad h = 1, 2, \ldots,$$

$$\text{Nível: } l_t = \alpha(Y_t - s_{t-s}) + (1-\alpha)(l_{t-1} + b_{t-1}),$$

$$\text{Declive: } b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},$$

$$\text{Efeito Sazonal: } s_t = \gamma(Y_t - l_t) + (1-\gamma)s_{t-s},$$

onde $0 \leq \gamma \leq 1 - \alpha$ e $h_s^+ = [(h-1) \bmod s] + 1$.

Para se iniciar o método aditivo de Holt-Winters é necessário definir os valores iniciais para o nível, o declive e para os índices de sazonalidade. As equações são as

seguintes:

$$\hat{l}_s = \frac{1}{s}\sum_{i=1}^{s} Y_i,$$

$$\hat{b}_s = \frac{1}{s^2}\left(\sum_{i=s+1}^{2s} Y_i - \sum_{i=1}^{s} Y_i\right),$$

$$\hat{s}_i = Y_i - \hat{l}_s, \text{para } i = 1, \ldots, s.$$

Holt-Winters Multiplicativo

O modelo de decomposição multiplicativo define-se por

$$Y_t = T_t \times S_t + \varepsilon_t.$$

O método multiplicativo é preferível nos casos em que as variações sazonais se alteram de forma proporcional ao nível da série. As equações que constituem este método são:

$$\text{Previsão: } \hat{Y}_{t+h} = (l_t + hb_t)s_{t-s+h_s^+}, h = 1, 2, \ldots,$$

$$\text{Nível: } l_t = \alpha\frac{Y_t}{s_{t-s}} + (1-\alpha)(l_{t-1} + b_{t-1}),$$

$$\text{Declive: } b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},$$

$$\text{Efeito Sazonal: } S_t = \gamma\frac{Y_t}{l_t} + (1-\gamma)_{s_{t-s}}.$$

Para se iniciar o método multiplicativo de Holt-Winters, é necessário definir os valores iniciais para o nível, declive e para os índices de sazonalidade. As equações são as seguintes:

$$\hat{l}_s = \frac{1}{s}\sum_{i=1}^{s} Y_i,$$

$$\hat{b}_s = \frac{1}{s^2}\left(\sum_{i=s+1}^{2s} Y_i - \sum_{i=1}^{s} Y_i\right),$$

$$\hat{s}_i = \frac{Y_i}{\hat{l}_s}, \text{ para } i = 1, \ldots, s.$$

A escolha ótima das constantes de alisamento (α, β, γ) é feita por métodos numéricos de otimização de modo a minimizarem o erro quadrático médio das previsões a $1 - passo$.

4.6 Modelos de Alisamento Exponencial para Dados com Sazonalidade Complexa

Os modelos de alisamento exponencial com um único fator sazonal, consideram-se ótimos para uma classe de inovações dos modelos de espaço de estados ((Ord et al., 1997), (Hyndman et al., 2002)). Os modelos de alisamento exponencial são os mais utilizados para previsão. Estes modelos admitem a possibilidade do cálculo da verosimilhança, da derivação de intervalos de predição consistentes e da seleção do modelo baseado nos critérios de informação.

Os modelos sazonais são frequentemente aplicados na estrutura das inovações dos modelos de espaço de estados que incluem aqueles subjacentes aos bem conhecidos métodos de Holt-Winters aditivo e multiplicativo. Taylor (2003) incorporou, na versão linear de Holt-Winters, uma segunda componente sazonal, do seguinte modo

$$\begin{aligned} y_t &= l_{t-1} + b_{t-1} + s_t^{(1)} + s_t^{(2)} + d_t, \\ l_t &= l_{t-1} + b_{t-1} + \alpha d_t, \\ b_t &= b_{t-1} + \beta d_t, \\ s_t^{(1)} &= s_{t-m_1}^{(1)} + \gamma_1 d_t, \\ s_t^{(2)} &= s_{t-m_2}^{(2)} + \gamma_2 d_t, \end{aligned} \tag{4.5}$$

onde m_1 e m_2 são os períodos dos ciclos sazonais e d_t é uma variável aleatória de ruído branco que representa o erro de predição (ou perturbação). As componentes l_t e b_t representam o nível e a tendência da série temporal no tempo t, respectivamente. A componente $s_t^{(i)}$ representa a i-ésima componente sazonal no tempo t. Os coefficientes α, β, γ_1 e γ_2 são os parâmetros de alisamento. As variáveis de estado iniciais são dadas por l_0, b_0, $\left\{s_{1-m_1}^{(1)}, \ldots, s_0^{(1)}\right\}$ e $\left\{s_{1-m_2}^{(2)}, \ldots, s_0^{(2)}\right\}$.

Os parâmetros e os valores iniciais devem ser estimados, mas quando o número de componentes sazonais é elevada esta estimação é complexa. Este problema é parcialmente resolvido observando a existência de uma redundância quando m_2 é um inteiro múltiplo de m_1. Considerando uma série temporal $\{y_t\}$, que consiste em sequências repetidas de constantes $c_1, \ldots, c_{m_1}$, uma para cada período no menor ciclo. Então as equações sazonais podem ser escritas da seguinte forma

$$s_t^{(1)} + y_t = (s_{t-m_1}^{(1)} + r_t) + \gamma_1 d_t,$$

$$s_t^{(2)} - y_t = (s_{t-m_2}^{(2)} + r_t) + \gamma_2 d_t.$$

O efeito de y_t desaparece quando estas equações se somam. Assim, o efeito sazonal do vetor inicial m_1 para o menor ciclo sazonal pode ser considerado nulo sem constrangimentos.

Apesar desta correção, um número elevado de valores sazonais iniciais permanece para ser estimado quando alguns dos padrões sazonais têm períodos longos, o que leva à sobreparameterização do modelo. Para séries temporais com sazonalidade dupla, Gould et al. (2008) tentaram reduzir este problema dividindo o valor do período sazonal mais longo em ciclos subsazonais que tenham padrões semelhantes. Contudo, para além da adaptação ser relativamente complexa, também só pode ser utilizada para padrões sazonais duplos em que nenhum comprimento sazonal é um múltiplo do outro (Hyndman, 2013). Para evitar o problema de otimização potencialmente grande, os valores iniciais são aproximados com várias heurísticas (Taylor (2003), Taylor (2010) e Gardner Jr e McKenzie (1985)), uma prática que não conduz a valores iniciais óptimos.

Os modelos usados no alisamento exponencial assumem que os erros $\{d_t\}$ são não correlacionados. Contudo, por vezes pode não se verificar. Num estudo empírico, usando o método multiplicativo Holt-Winters, Chatfield (1978) mostrou que quando os erros são correlacionados estes podem ser descritos por um processo AR(1). Outros como Taylor (2003), Gardner Jr (1985) e Gilchrist (1976) também mencionaram este problema dos erros correlacionados e da possibilidade de melhorar a precisão das previsões procedendo à modelação dos erros.

As versões não lineares dos modelos de espaço de estados com base no alisamento exponencial, apesar de muito utilizadas, apresentam algumas fraquezas. Algumas das versões não lineares podem ser instáveis e ter variâncias de previsão infinitas, quando o horizonte de previsão ultrapassa determinado limite, Akram et al. (2009). Outra das lacunas destes modelos é o facto dos resultados analíticos das distribuições de previsão não serem conhecidos. O problema destes modelos é que nenhuma das aproximações pode ser usada para lidar com padrões sazonais complexos, como sazonalidade não inteira, efeitos de calendário ou séries temporais com padrões sazonais não aninhados. Uma das soluções para estes problemas é proposta por De Livera et al. (2011).

4.7 Modelos Modificados

Padrões sazonais complexos, tais como períodos sazonais múltiplos, sazonalidade com período não inteiro, efeitos de duplo calendário e sazonalidade de alta frequência, são cada vez mais frequentes nas séries temporais. A necessidade de modelos com capacidade para lidar com estes problemas levou à introdução de modificações aos modelos de espaço de estados de alisamento exponencial. Surgiram dois novos métodos propostos por De Livera et al. (2011): BATS (*Box-Cox Transformation, ARMA errors, Trend, and Seasonal Components*) e TBATS (*Trigonometric, Box-Cox Transformation, ARMA errors, Trend, and Seasonal Components*). Tendo em consideração os problemas da não linearidade, vão-se considerar apenas os modelos lineares homocedásticos, mas permitindo alguns tipos de não linearidade usando as transformações Box-Cox, Box e Cox (1964). Desta forma, ao modelo 4.5 foi incluída a transformação Box-Cox, erros ARMA e T padrões sazonais da seguinte forma:

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega}, & \omega \neq 0, \\ \log\ y_t, & \omega = 0 \end{cases} \tag{4.6}$$

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t, \tag{4.7}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t, \tag{4.8}$$

$$b_t = (1-\phi) b_{t-1} + \phi b_{t-1} + \beta d_t, \tag{4.9}$$

$$s_t^{(i)} = s_{t-m_i}^{(i)} + \gamma_i d_t, \tag{4.10}$$

$$d_t = \sum_{i=1}^{p} \varphi d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t, \tag{4.11}$$

onde y_t é a observação original da série temporal observada no instante t, $t = 1, \ldots, n$, $y_t^{(w)}$ representa a observação transformada (w é o parâmetro de transformação Box-Cox), $m_1, \ldots, m_T$ representam os períodos sazonais, l_t é o nível estocástico local, b é a tendência de longo prazo, b_t é a tendência de curto prazo no período t, $s_t^{(i)}$ representa a i-ésima componente sazonal no tempo t e d_t é um processo $ARMA(p, q)$ e ε_t é um ruído branco com média zero e variância σ^2 constante. Os parâmetros de alisamento são dados por α, β e γ_i, $i = 1,\ldots, T$. O número de padrões sazonal

da série temporal é representadoor T. Foi adotada a tendência de amortecimento, Gardner Jr e McKenzie (1985), com parâmetro de amortecimento ϕ, mas seguindo a sugestão de Arca et al. (2006) de completar com uma tendência de longo prazo b. Esta mudança assegura que as previsões dos valores futuros da tendência a curto prazo b_t convergem para a tendência a longo prazo b, em vez de convergirem para zero. Os valores iniciais dos estados são l_0, b_0, $\{s_0^1, s_0^2, \ldots, s_0^{m_1}\}$, $\{s_0^1, s_0^2, \ldots, s_0^{m_2}\}$, ... , $\{s_0^1, s_0^2, \ldots, s_0^{m_k}\}$, (De Livera et al., 2011).

4.7.1 Modelo BATS

O modelo BATS é acrónimo para as principais funcionalidades do modelo: *Box-Cox Transformation*, *ARMA errors*, *Trend*, *and Seasonal Components* e representa-se por

$$BATS(\omega, p, q, \phi, m_1, m_2, \ldots, m_T).$$

Os parâmetros do modelo são $\omega, p, q, \phi, m_1, m_2, \ldots, m_T$. O parâmetro de transformação Box-Cox é representado por ω, quando $\omega = 1$ não existe tranformação, Box & Cox (1964). O parâmetro de amortecimento é representado por ϕ, quando $\phi = 1$ ou próximo de 1 o parâmetro não tem efeito, Gardner Jr e McKenzie (1985). Os erros são modelados por um processo $ARMA(p, q)$ (Anderson (1976), Chen et al. (1996)) e $m_1, \ldots, m_T$ são os períodos sazonais usados no modelo. O modelo BATS é a generalização mais próxima dos modelos modificados sazonais tradicionais que permitem períodos sazonais múltiplos. Contudo, tem limitações, é incapaz de lidar com sazonalidade de período não inteiro e com sazonalidade de alta frequência e pode envolver um número grande de estados. A componente sazonal inicial sozinha contém m_T estados não nulos, isto leva a um grande número de valores para padrões sazonais de períodos longos. Não consegue incorporar variáves explicativas o que pode ser uma desvantagem, uma vez que em problemas de previsão, sabe-se que a informação adicional pode estar disponível na forma de variáveis de influência externa. A suposição de $\varepsilon_i \sim N(0, \sigma^2)$ pode não ser válida.

Assim, a transformação Box-Cox permite lidar com os problemas de não linearidade dos dados. O processo ARMA sobre os resíduos permite resolver o problema da autocorrelação. Estes têm a capacidade de obter não só a previsão pontual, como também os intervalos de previsão. Não é necessário definir valores iniciais e o desempenho do modelo é significativamente melhor, quando comparada com um modelo de espaço de estados tradicional.

Por exemplo, o modelo $BATS(1, 1, 0, 0, m_1)$ representa o bem conhecido modelo

de Holt-Winters aditivo sazonal, De Livera et al. (2011).

O modelo BATS é totalmente automático, encontra-se implementado no *package forecast* do *software* R, Hyndman et al. (2018). A generalização do modelo BATS é o modelo TBATS, em que o modelo BATS difere do TBATS nos T padrões sazonais.

4.7.2 Modelos Sazonais Trigonométricos (TBATS)

A existência de séries temporais com padrões sazonais complexos e as limitações do modelo BATS, como a incapacidade de lidar com séries temporais de sazonalidade de período não inteiro e a fraca capacidade de lidar com modelos de sazonalidade de alta frequência, conduziu à formulação do modelo TBATS. O modelo TBATS é introduzido através da representação trignométrica de componentes sazonais baseadas em séries de Fourier, Harvey e Fernandes (1989); West e Harrison (1989), representadas da seguinte forma:

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)}, \tag{4.12}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t, \tag{4.13}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t, \tag{4.14}$$

onde $\gamma_1^{(i)}$, $\gamma_2^{(i)}$ são os parâmetros de alisamento e $\lambda_j^{(i)} = \frac{2\pi j}{m_i}$ $(j = 1, \ldots, k_i)$, sendo k_i o número de termos de Fourier necessário para os termos trigonométricos na i-ésima componente sazonal, $i = 1, \ldots, T$. O nível estocástico da i-ésima componente sazonal é descrito por $s_{j,t}^{(i)}$ e o crescimento estocástico no nível da i-ésima componente sazonal que é necessário para descrever a mudança na componente sazonal ao longo do tempo é dada por $s_{j,t}^{*(i)}$.

O modelo TBATS representa-se por

$$TBATS(\omega, p, q, \phi, \{m_1, k_1\}, \{m_2, k_1\}, \ldots, \{m_T, k_T\}).$$

TBATS é o acrónimo para as principais funcionalidades do modelo: *Trigonometric, Box-Cox Transformation, ARMA errors, Trend, and Seasonal Components.*

O parâmetro de transformação Box-Cox é representado por ω, quando $\omega = 1$ não existe tranformação. O parâmetro de amortecimento é ϕ, quando $\phi = 1$ ou muito próximo de 1 não tem efeito. Os erros são modelados por um processo $ARMA(p, q)$, $m_1, \ldots, m_T$ são os períodos sazonais usados no modelo e $k_1, \ldots, k_T$ são os números

correspondentes aos termos de Fourier usados em cada período sazonal.

O modelo apresenta algumas fraquezas, nomeadamente, a suposição de $\varepsilon_i \sim N(0, \sigma^2)$ pode não ser válida. Pois, quando $p = q = 0$ o modelo TBATS é equivalente ao modelo de alisamento exponencial aditivo (TETS - *Trigonometric Exponential Smoothing*). É, ainda, incapaz de incorporar variáveis explicativas. O desempenho do modelo para obter previsões de longo prazo não é muito robusto e o custo em termos computacionais é elevado, se a dimensão dos dados for grande. Os intervalos de confiança das previsões podem também ser bastante grandes. O modelo TBATS tem todas as potencialidades do modelo BATS com o acréscimo da já referida capacidade de lidar com sazonalidade com período não inteiro e sazonalidade de alta frequência. Consegue, ainda, usar sazonalidade múltipla sem aumentar muito o número de parâmetros do modelo.

4.8 Formulação em Espaço de Estados

Os modelos descritos nas Secções anteriores são casos particulares de modelos em espaços de estados (Anderson e Moore (1979)), adaptados, neste caso, para incorporar a transformação de Box-Cox para resolver os problemas de não linearidade. Assim, o modelo tem a seguinte forma, com a equação de observação e a equação de estado seguintes

$$y_t^{(\omega)} = \boldsymbol{w'} \boldsymbol{x_{t-1}} + \boldsymbol{\varepsilon_t}, \tag{4.15}$$

$$\boldsymbol{x_t} = \boldsymbol{F} \boldsymbol{x_{t-1}} + \boldsymbol{g} \boldsymbol{\varepsilon_t}, \quad t = 1, \ldots, n \tag{4.16}$$

onde $\boldsymbol{w'}$ é um vetor linha conhecido, $\boldsymbol{g}$ é um vetor coluna, $\boldsymbol{F}$ é a matriz de transição e $\boldsymbol{x_t}$ é o vetor dos estados não observáveis no tempo t. De forma mais pormenorizada, veja-se Apêndice A, De Livera et al. (2011).

4.9 Estimação do Modelo em Espaço de Estados

O processo de estimação implica a formulação dos modelos em espaço de estados, que requer a estimação das matrizes dos seus sistemas.

A estimação dos parâmetros desconhecidos do modelo linear em espaço de estados, como os parâmetros de alisamento e o parâmetro de amortecimento é feita usando a soma dos quadrados dos erros (mínimos quadrados) ou a função de máxima verosimilhança Gaussiana. No contexto dos modelos BATS e TBATS ainda

é necessário estimar o parâmetro desconhecido de transformação Box-Cox (ω) e os coeficientes do processo ARMA.

Geralmente, os valores iniciais dos modelos em espaço de estados são tratados como vetores aleatórios. Assim, para determinados valores de ϕ e σ^2, a taxa de crescimento de curto prazo dos valores iniciais assumem-se com uma distribuição $N(0, \sigma^2/(1-\phi^2))$. Contudo, como a maioria dos estados são não estacionários, assume-se que têm distribuições Gaussianas com variâncias arbitrariamente elevadas, Ansley e Kohn (1985). O filtro de Kalman geralmente é utilizado para obter os erros de previsão a um passo e as variâncias associadas necessárias para avaliar os critérios de ajustamento, em particular a significância dos parâmetros. Apesar disto, o filtro de Kalman necessita de equações adicionais para lidar com estados não estacionários.

De uma forma mais simples, a alternativa encontra-se no contexto das inovações dos modelos em espaço de estados. Condicionando todos os valores iniciais e assumindo que se tratam de parâmetros fixos desconhecidos, o alisamento exponencial pode ser utilizado em detrimento do filtro de Kalman para gerar os erros de predição a um passo, necessários para a avaliação da verosimilhança. Assim, os parâmetros e os valores inciais são selecionados para maximizar a função de verosimilhança condicional resultante.

A função de verosimilhança condicional dos dados observados $\boldsymbol{y} = (y_1, \ldots, y_n)$ é definida assumindo que $\varepsilon_t \sim N(0, \sigma^2)$. Assim, a densidade da série transformada $y_t^{(\omega)} \sim N(\boldsymbol{w'x}_{t-1}, \sigma^2)$ e a densidade dos dados transformados é dada por

$$p(y^{(\omega)} \mid \boldsymbol{x_0}, \boldsymbol{\vartheta}, \sigma^2) = \prod_{t=1}^{n} p(y_t^{(\omega)} \mid \boldsymbol{x_{t-1}}, \boldsymbol{\vartheta}, \sigma^2) = \prod_{t=1}^{n} p(\varepsilon_t) = \frac{1}{(2\pi\sigma^2)}^{\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{t=1}^{n}\varepsilon_t^2\right),$$

onde ϑ é um vetor contendo o parâmetro Box-Cox, os parâmetros de alisamento e os coeficientes do processo ARMA. Assim, a densidade da série original, usando o Jacobiano da tranformação Box-Cox é dada por

$$p(y \mid \boldsymbol{x_0}, \boldsymbol{\vartheta}, \sigma^2) = p(y_t^{(\omega)} \mid \boldsymbol{x_0}, \boldsymbol{\vartheta}, \sigma^2) \left| det\left(\frac{\partial y_t^{(\omega)}}{\partial y}\right)\right| = p(y_t^{(\omega)} \mid \boldsymbol{x_0}, \boldsymbol{\vartheta}, \sigma^2) \prod_{t=1}^{n} y_t^{\omega-1} =$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(\frac{-1}{2\sigma^2}\prod_{t=1}^{n}\varepsilon_t^2\right)\prod_{t=1}^{n} y_t^{\omega-1}.$$

Considerando a variância σ^2 com a sua estimativa de máxima verosimilhança resulta

$$\hat{\sigma}^2 = n^{-1}\sum_{t=1}^{n}\varepsilon_t^2. \tag{4.17}$$

Desta forma, a log-verosimilhança é dada por

$$\mathcal{L}(\boldsymbol{x_0}, \boldsymbol{\vartheta}, \sigma^2) = \frac{-n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\varepsilon_t^2 + (\omega - 1)\sum_{t=1}^{n}\log y_t. \tag{4.18}$$

Substituindo 4.17 em 4.18, multiplicando por -2 e omitindo os termos constantes, obtém-se

$$\mathcal{L}*(\boldsymbol{x_0}, \boldsymbol{\vartheta}) = n\log\sum_{t=1}^{n}\varepsilon_t^2 - 2(\omega - 1)\sum_{t=1}^{n}\log y_t. \tag{4.19}$$

O objetivo é minimizar a quantidade 4.18 para obter as estimativas de máxima verosimilhança. Contudo a dimensão do vetor $\boldsymbol{x_0}$ dos valores iniciais torna o processo computacionalmente desafiador. A abordagem a este problema é baseada na observação de que ε_t é uma função linear do vetor de valores iniciais $\boldsymbol{x_0}$. Desta forma, é possível concentrar os valores iniciais fora da verosmilhança, reduzindo de forma substancial a dimensão do problema de optimização numérica. A partir da equação de transição 4.15 a inovação ε_t pode ser eliminada, obtendo-se $\boldsymbol{x_t} = \boldsymbol{D x_{t-1}} + \boldsymbol{g y_t}$, onde $\boldsymbol{D} = \boldsymbol{F} - \boldsymbol{g w'}$. A equação para o estado, obtida por resolução recursiva da equação para o período 0, pode ser utilizada em conjugação com a equação de medida para obter

$$\begin{aligned}
\varepsilon_t &= y_t^{(\omega)} - \boldsymbol{w'}\sum_{\boldsymbol{j=1}}^{\boldsymbol{t-1}}\boldsymbol{D^{j-1}}\mathbf{g}\, y_{t-j}^{(\omega)} - \boldsymbol{w' D^{t-1} x_0} \\
&= y_t^{(\omega)} - \boldsymbol{w'\tilde{x}_{t-1}} - \boldsymbol{w'_{t-1}x_0} \\
&= \tilde{y}_t - \boldsymbol{w'_{t-1}x_0},
\end{aligned} \tag{4.20}$$

onde $\tilde{y}_t = y_t^{(\omega)} - \boldsymbol{w'\tilde{x}_{t-1}}$, $\boldsymbol{\tilde{x}_t = D\tilde{x}_{t-1} + \mathbf{g}y_t}$, $\boldsymbol{w'_t = D w'_{t-1}}$, $\boldsymbol{\tilde{x}_0 = 0}$ e $w'_0 = w'$. Assim, a relação entre a relação entre cada erro e o vetor de estado inicial $\boldsymbol{x_0}$ é linear. Pela equação 4.20, também se percebe que o vetor de estados inicial $\boldsymbol{x_0}$ corresponde a um vetor de coeficientes de regressão e que, desta forma, também pode ser estimado usando o método de estimação linear de mínimos quadrados. Assim, o problema reduz-se a minimizar a seguinte função em relação a $\boldsymbol{\vartheta}$

$$\mathcal{L}^*(\boldsymbol{\vartheta}) = n\log(SSE^*) - 2(\omega - 1)\sum_{t=1}^{n}\log y_t, \tag{4.21}$$

SSE^* é o valor otimizado da soma do quadrado dos erros para os valores dos parâmetros. Esta abordagem concentra-se nos valores iniciais da função de verosimilhança, deixando apenas para otimização o menor dos vetores de parâmetros, o que permite

a obtenção de melhores previsões. Outra das vantagens é a redução dos tempos computacionais, uma vez que se estima diretamente o vetor de estados iniciais não sendo necessário recorrer a um otimizador numérico.

4.10 Predição

A distribuição de predição no espaço transformado para um período futuro $n+h$, dado o vetor de estados final x_n e dados os parâmetros $\boldsymbol{\vartheta}$, $\boldsymbol{\sigma^2}$, é Gaussiana. A variável aleatória associada é designada por $y^{(w)}_{n+h|n}$. A sua média $E(y^{(w)}_{n+h|n})$ e variância $V(y^{(w)}_{n+h|n})$ são dadas, após a transformação Box-Cox pelas seguintes equações, Shenstone e Hyndman (2005):

$$E(y^{(w)}_{n+h|n}) = \boldsymbol{w'F^{h-1}x_n}, \tag{4.22}$$

$$V(y^{(w)}_{n+h|n}) = \begin{cases} \sigma^2, & \text{se } h = 1 \\ \sigma^2 \left[1 + \sum_{j=1}^{h-1} c_j^2\right], & \text{se } h \geq 2 \end{cases} \tag{4.23}$$

onde $c_j = \boldsymbol{w'F^{j-1}g}$.

A distribuição de predição $y_{n+h|n}$ não é Normal. Contudo, as previsões pontuais e os intervalos de previsão devem ser obtidos usando a transformação inversa de Box-Cox dos quantis apropriados da distribuição $y^{(w)}_{n+h|n}$. O ponto de previsão obtido desta maneira é a mediana, um preditor do erro absoluto médio mínimo. Os intervalos de predição retêm a cobertura de probabilidade necessária sob transformação inversa, uma vez que, a transformação de Box-Cox aumenta de forma monótona. Os parâmetros e o estado final são substituídos pelas suas estimativas na fórmula descrita. Desta forma, o impacto do erro de estimação é ignorado, mas o último é um efeito de segunda ordem na maioria dos contextos práticos.

4.10.1 Seleção do Número de Harmónicos nos Modelos Trigonométricos

As previsões do modelo TBATS dependem do número de harmónicos, k_i, usados na componente sazonal i. Na procura de encontrar a melhor das combinações é impraticável considerar todas as combinações possíveis. Hyndman et al. (2008) sugerem a utilização de regressão linear dada por

$$\sum_{i=1}^{T} \sum_{j=1}^{k_i} a_j^{(i)} \cos(\lambda_j^{(i)} t) + b_j^{(i)} \sin(\lambda_j^{(i)} t).$$

Para tal, estes autores descrevem os seguintes passos

1. iniciar o processo com um único harmónico. De forma gradual, vão adicionando-se mais, testando a significância de cada um usando testes F. Seja k_i^* é o número de harmónicos significativos (com valor de prova $< 0,001$) para a i-ésima componente sazonal;

2. então ajusta-se o modelo aos dados com $k_i = k_i^*$ e calcula-se o respectivo AIC;

3. considerando apenas uma componente sazonal de cada vez, o modelo é ajudado repetidamente à amostra de estimação, aumentando gradualmente o k_i e mantendo todos os outros harmónicos constantes para cada i, até que o menor valor de AIC seja alcançado.

Esta abordagem baseada na regressão linear múltipla, é preferível à abordagem em que se assume $k_i^* = 1$ para cada componente, que é pouco prática e consome muito tempo.

4.10.2 Seleção das Ordens p e q do Processo ARMA

No processo de seleção do modelo é necessária a determinação dos valores adequados para as ordens p e q do processo ARMA. Esta seleção utiliza um procedimento em duas etapas:

1. seleciona-se um modelo adequado sem a componente ARMA. Depois aplica-se o algoritmo ARIMA automático de Khandakar e Hyndman (2008) aos resíduos do modelo selecionado a fim de escolher as ordens adequadas p e q, assumindo os resíduos estacionários;

2. o modelo selecionado é ajustado novamente, mas com a componente de erro $ARMA(p, q)$, em que os coeficientes ARMA são estimados em conjunto com os restantes parâmetros do modelo. A componente $ARMA$ é utilizada apenas se o modelo resultante do ajustamento apresentar um menor AIC, quando comparado com o modelo sem esta componente.

De Livera et al. (2011) mostraram que este procedimento de duas etapas apresenta melhores previsões (fora da amostra), em comparação com outras abordagens alternativas.

4.11 Seleção de Modelos

Na modelação de uma série temporal, podem existir mais do que um modelo a verificar os diferentes critérios de validação, o que dificulta o processo de seleção do melhor modelo. Desta forma, é necessário recorrer a critérios de seleção que considerem as estatísticas baseadas nos resíduos do modelo ajustado.

São vários os critérios, baseados na função de verosimilhança, existentes na literatura, sendo os mais utilizados o critério de informação de Akaike (*Akaike's information criterion*,AIC), o critério de informação de Akaike corrigido (*Corrected Akaike's information criterion*, AICc) e o critério de informação Bayesiano (*Bayesian information criterion*, BIC).

Critério de informação de Akaike

Akaike (1974) introduziu um critério baseado na quantidade de informação, para avaliar a qualidade de um ajustamento. Considere-se que um modelo com p parâmetros foi ajustado a uma série com n observações, o critério é definido por

$$\mathrm{AIC} = -2\log L + 2p, \tag{4.24}$$

onde L é a verosimilhança.

Bozdogan (1987) propôs uma correção ao critério AIC, definindo AICc por

$$\mathrm{AICc} = -2\log L + 2p + 2\frac{p(p+1)}{n-p-1}, \tag{4.25}$$

O modelo escolhido, deverá ser o que apresente menor AIC. Nesses casos, deve optar-se pelo modelo mais simples, de forma que seja parcimonioso ou para obter um melhor ajustamento do modelo.

Critério de informação Bayesiano

Schwarz (1978) propõe o critério de informação Bayesiano, definido como

$$\mathrm{BIC} = -2\log L + p\log(n), \tag{4.26}$$

onde L é a verosimilhança, p é o número de parâmetros do modelo e n é a dimensão da amostra.

O BIC depende da dimensão da amostra, n. Para uma amostra de dimensão superior a 7, a penalização do BIC é superior à penalização do AIC. Consequente-

mente, a minimização do BIC leva, em geral, à seleção de modelos com um menor número de parâmetros do que os obtidos pela minimização do critério AIC, evitando, a sobrestimação do número de componentes.

4.12 Medidas de Avaliação

Uma série temporal, geralmente, é dividida em dois períodos amostrais: uma amostra de treino (para ajustar o modelo) e uma amostra de teste (para avaliar a qualidade das previsões). A dimensão da amostra de teste é usualmente cerca de 20% da dimensão da amostra total. No entanto, este valor depende da dimensão da amostra e do horizonte temporal para a previsão. A dimensão da amostra de teste deve ser, pelo menos, igual ou superior ao horizonte temporal de previsão pretendido.

Considerando uma série temporal de dimensão n, $\{Y_1, \ldots, Y_n\}$ e um horizonte temporal de previsão, h, divide-se a série em série de treino $\{Y_1, \ldots, Y_{n-h}\}$ e em série de teste $\{Y_{n-h+1}, \ldots, Y_n\}$.

Para a avaliação da qualidade preditiva usualmente consideram-se previsão a 1-passo e previsão multi-passos à frente. A previsão a 1-passo, tal como o nome indica, prevê uma unidade temporal à frente ($h = 1$) da última observação. Isto é, se Y_t é a observação no instante t e $\hat{Y}_t$ a sua estimativa obtida usando as observações $Y_1, Y_2, \ldots, Y_{t-1}$, então $\hat{Y}_t$ é a previsão a 1-passo de Y_t. Da mesma forma, utilizando todas as observações até ao tempo t (inclusive), é possível obter a previsão h-passos à frente, $\hat{Y}_{t+h|t}$.

Um erro de previsão é a diferença entre o valor observado e a sua previsão. De forma geral, o erro de previsão (a h-passos) pode ser escrito como $e_{t+h} = Y_{t+h} - \hat{Y}_{t+h|t}$. Se se considerar $h = 1$, ou seja, a previsão a 1-passo, o erro de previsão a 1-passo é dado simplesmente por $e_t = Y_t - \hat{Y}_t$.

Hyndman e Koehler (2006) classificam as medidas de avaliação em três tipos: medidas dependentes da escala, medidas de erros percentuais e medidas de erros escalados.

Medidas dependentes da escala

Estas medidas de precisão baseadas possuem uma escala que depende da escala dos dados e, por isso, não devem ser utilizadas para comparar dados com diferentes escalas. As medidas mais utilizadas, dependentes da escala, são baseadas em erros absolutos ou erros quadráticos:

$$\text{Erro Médio: EM} = \frac{1}{n}\sum_{t=1}^{n} e_t = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t),$$

$$\text{Erro Quadrático Médio: EQM} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2.$$

Por vezes, a raiz do erro quadrático médio, REQM $= \sqrt{\text{EQM}}$, é preferida ao EQM porque permite reduzir a grandeza dos valores para a mesma escala dos dados.

Na comparação entre métodos de previsão, considera-se que o método mais preciso é o que apresenta menor EQM e, consequentemente, menor REQM. A REQM e o EQM são as medidas mais utilizadas, no entanto, estas medidas são mais sensíveis a *outliers* quando comparadas com outras do mesmo tipo, como por exemplo, o erro absoluto médio.

$$\text{Erro Absoluto Médio: EAM} = \frac{1}{n}\sum_{t=1}^{n} |e_t| = \frac{1}{n}\sum_{t=1}^{n} \left|Y_t - \hat{Y}_t\right|.$$

Medidas de erros percentuais

As medidas de erros percentuais são independentes da escala e são frequentemente usadas para comparar o desempenho da previsão entre diferentes séries. A medida mais utilizada é o

$$\text{Erro Percentual Absoluto Médio: EPAM} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100 \quad (\%)\,,$$

que representa a percentagem média do erro de previsão em relação à grandeza das observações, sendo assim, uma medida de erros percentuais.

Tal como acontece no EQM, quanto menor for o EPAM mais preciso é o método de previsão. Apesar de possibilitar a comparação da precisão entre diversas séries, esta medida não pode ser calculada quando existem zeros na série. Também quando as observações se aproximam de zero, o EPAM apresenta valores extremos.

Medidas de erros escalados

As medidas de erros escalados foram propostas por Hyndman e Koehler (2006), como alternativa à medida dos erros percentuais. Comparam a qualidade preditiva de séries com diferentes escalas, uma vez que, o erro escalado é independente da escala dos dados. Um erro escalado é dado por

$$q_t = e_t/Q,$$

onde Q é o erro absoluto médio da previsão *naïve* sazonal e é dado por $Q = \frac{1}{n-s}\sum_{t=s+1}^{n} |Y_t - Y_{t-s}|$, onde s é o período sazonal.

O erro escalado absoluto médio (EEAM) define-se por

$$\text{Erro Escalado Absoluto Médio: EEAM} = \frac{1}{n}\sum_{t=1}^{n}|q_t| = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|}\right|.$$

Quanto à interpretação do EEAM, valores superiores a 1 indicam que as previsões para o método são menos precisas, em média, do que as *naïve* e, portanto, quanto mais próximo de zero maior é a precisão do método. Desta forma, na comparação entre diferentes métodos de previsão, considera-se que o mais preciso é o que apresenta o menor EEAM.

4.13 Valores em Falta

Bases de dados reais, geralmente, apresentam valores em falta. Trabalhar com séries temporais com dados em falta pode ser muito problemático. Estes surgem por diversos motivos, aleatórios ou não, como falhas na recolha dos dados, erros nas medições ou amostragens insuficientes. A ausência destes valores conduzem a enviesamentos no modelo de estimação e/ou previsão. Algumas técnicas permitem o estabelecimento do modelo que se pretende implementar com a existência de valores em falta, sem causar problemas, como por exemplo, os métodos "*naïve*", os modelos ARIMA, os modelos de regressão dinâmica, entre outros. Contudo, outros processos de modelação não conseguem lidar com dados em falta em particular, funções no *software* R como *ets()* , *stlf()* e *tbats()* .

De forma alternativa, pode substituir-se os valores em falta por estimativas. A função *na.interp()*,do *package forecast* do *software* R é desenhada com este propósito e funciona de forma bastante eficaz quando comparada com outros métodos de imputação de dados automáticos (Moritz et al., 2015).

Usualmente, para dados sem sazonalidade é utilizada a interpolação linear simples para preencher os valores em falta. A interpolação linear é o método que ajusta um segmento de reta entre dois pontos, os pontos finais entre as lacunas dos dados. Esta permite que os valores ausentes sejam calculados diretamente usando a equação do segmento de reta. A equação do método de interpolação linear, Chapra e Canale (1998), é dada por

$$f(x) = f_{x_0} + \frac{f_{x_1} - f_{x_0}}{x_1 - x_0}(x_1 - x_0),$$

onde x_1 e x_0 são os valores conhecidos da variável independente, $f(x)$ é o valor da variável dependente para um valor x desconhecido da variável independente.

Na presença de sazonalidade é utilizada, usualmente, uma decomposição STL (*Seasonal-Trend Decomposition by Loess*), STL é o acrónimo para a decomposição sazonal e de tendência usando Loess. Após uma decomposição STL, segue-se com uma interpolação linear aplicada aos dados ajustados sazonalmente e, no final, a componente sazonal é adicionada novamente. A decomposição STL é um método versátil e bastante robusto para decompor séries temporais baseada em regressão local robusta. Este método consiste na ideia de que uma série temporal pode ser decomposta em três componentes: a componente sazonal, a componente de tendência e a componente do erro. Este método foi desenvolvido por Cleveland et al., (1990). Denotanto os dados, a componente tendência, a componente sazonal e a componente do erro por Y_ν, T_ν, s_ν e R_ν, respetivamente, para $\nu = 1, \ldots, n$ equação do modelo é representada por

$$Y_\nu = T_\nu + S_\nu + R_\nu.$$

Este método é robusto na presença de *outliers*, impedindo que estas observações afetem as estimativas das componentes da tendência e da sazonalidade. No entanto, a componente do erro é sempre afetada pelos *outliers*. Computacionalmente é de fácil utilização, mesmo para séries temporais longas.

Capítulo 5

Análise Exploratória de Dados

A análise descritiva é o primeiro e muito importante passo em qualquer estudo estatístico. Esta análise permite identificar as características intrínsecas dos dados em estudo. O presente estudo tem como foco a análise de quatro variáveis meteorológicas a temperatura máxima (ºC), a temperatura mínima (ºC), a velocidade média do vento (m/s) e a precipitação (mm) registadas numa estação meteorológica localizada numa quinta designada por Senhora da Ribeira em Carrazeda de Ansiães, no distrito de Bragança, no Norte de Portugal. São observações diárias, recolhidas no período de 1 de janeiro de 2010 a 23 de abril de 2019 (o que corresponde a 3400 dias). Neste período de tempo apenas dois anos foram bissextos: 2012 e 2016. Assim, apesar das metodologias implementadas neste estudo, nomeadamente o modelo TBATS, conseguirem incorporar na componente sazonalidade de período de 365 dias, os anos bissextos (com 366 dias), seguiu-se a sugestão de Benth e Benth (2010) de remover as observações registadas a 29 de fevereiro. Na Figura 5.1 estão representadas as séries temporais para as 4 variáveis em estudo, no período observado. É clara a existência de um padrão sazonal forte, principalmente nas séries da temperatura máxima do ar e da temperatura mínima do ar. As representações gráficas da precipitação e da velocidade média do vento mostram um comportamento volátil e com uma enorme variabilidade.

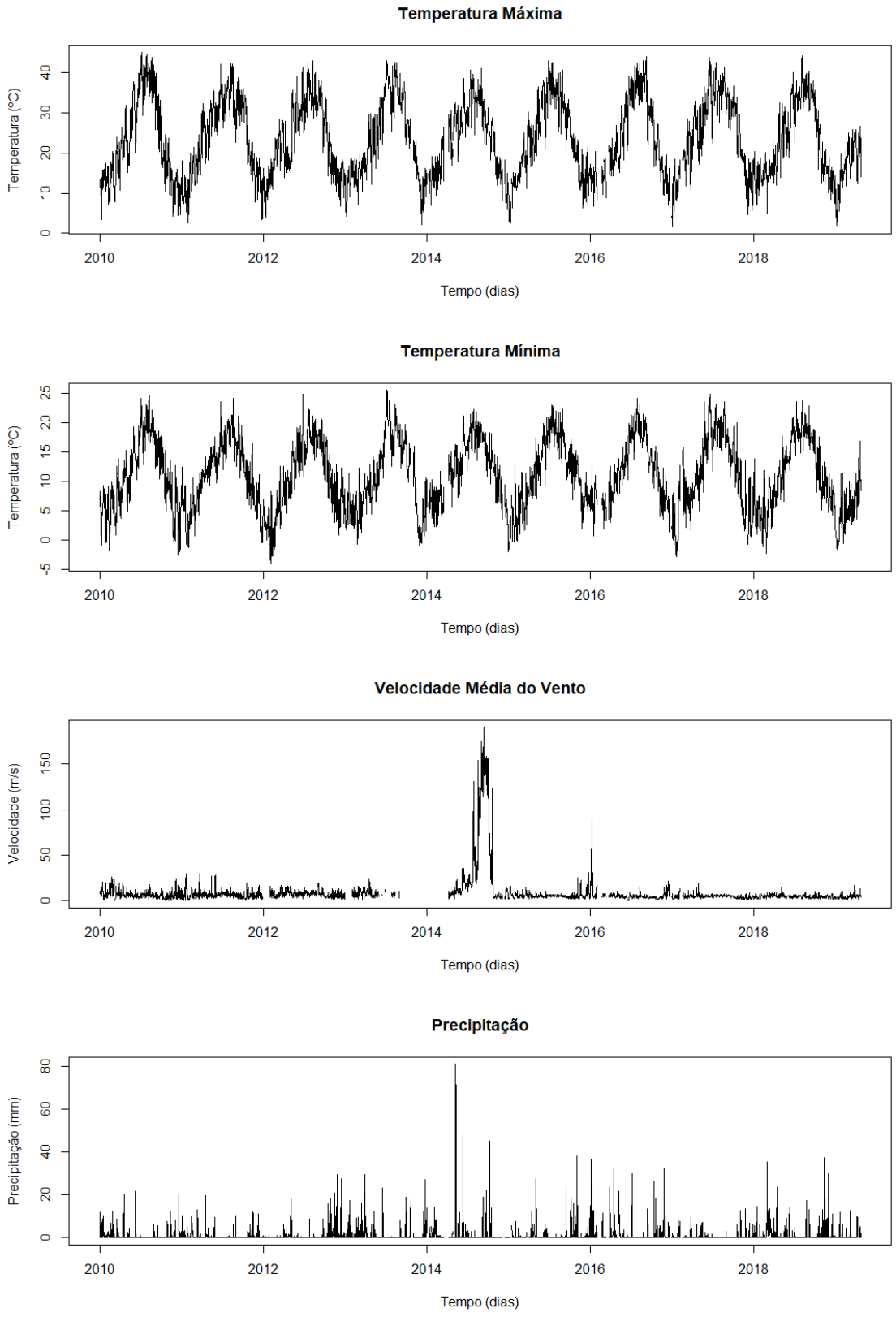


Figura 5.1: Representação gráfica das séries temporais, das variáveis em estudo, para o período observado.

A Tabela 5.1 apresenta as estatísticas descritivas das séries temporais em estudo para o período observado. As variáveis temperatura máxima, mínima e precipitação têm 114 valores em falta. A velocidade média do vento é a variável que mais observações em falta tem, como é possível ver graficamente, com um total de 422 dias sem observações.

No período observado houve 2404 dias em que o valor da precipitação registado foi nulo (não choveu). As estatísticas descritivas apresentadas na Tabela 5.1, relativamente à precipitação, foram calculadas para os valores observados diferentes de zero, i.e., dias com precipitação (no total 882 dias). A velocidade média do vento caracteriza-se por uma elevada variabilidade, o desvio padrão é de 18,05 m/s. As temperaturas máximas registadas no local em estudo apresentam uma média de 22,90 ºC e, pelo menos, 75% das temperaturas são superiores 15,70 ºC. Em contrapartida, a média das temperaturas mínimas ronda os 10,99 ºC e, pelo menos, 25% delas estão situadas abaixo dos 6,50 ºC.

Tabela 5.1: Estatísticas descritivas das observações diárias das variáveis em estudo.

Variável	Mín.	Máx.	1.º Quart.	Med.	3.º Quart.	Média	Desvio Padrão	Assim.	Curt.	Total (em dias)	NA's
Temp. Mín.	-4,10	25,50	6,50	10,90	15,80	10,99	5,91	-0,03	-0,83	3400	114
Temp. Máx.	1,70	45,00	15,70	22,80	31,80	23,66	9,52	0,12	-1,03	3400	114
Precip.	0,20	80,80	0,60	2,20	5,80	4,70	7,20	4,07	26,45	3400	114
Vel. Méd. Vento	0,00	190,40	3,80	5,40	7,80	9,01	18,05	6,65	47,36	3400	422

Na Figura 5.2 estão representados os diagramas em caixas de bigodes de cada uma das variáveis em estudo. Por observação gráfica, é possível perceber que a variável temperatura máxima e temperatura mínima não apresentam valores *outliers*. Por sua vez, as variáveis precipitação e velocidade média do vento apresentam muitos valores *outliers*. No caso da precipitação, destaca-se o valor registado de precipitação de 80,80 mm no dia 10 de março de 2014 e de 62,40 mm no dia 11 de maio de 2014. No caso da velocidade média do vento, o maior valor registado foi de 190,40 m/s no dia 12 de setembro de 2014 e um valor de 175,30 m/s no dia 2 de setembro de 2014.

Os histogramas da Figura 5.3 mostram que a distribuição das variáveis não se comportam como uma distribuição Normal. As distribuições da velocidade média do vento e da precipitação são assimétricas positivas com uma tendência de um acumular de frequências para valores próximos de zero (75% dos dados são inferiores ou iguais a 5,8 mm e 7,8 m/s para a precipitação e para a velocidade média do vento, respectivamente).

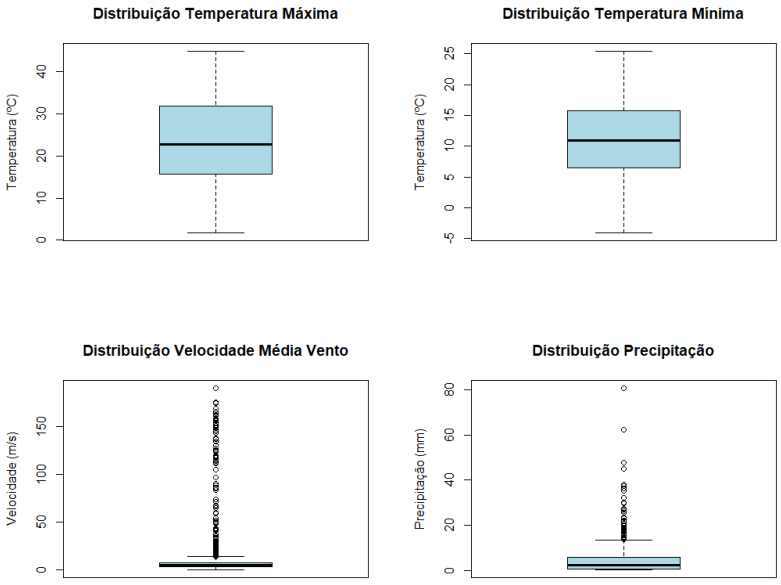


Figura 5.2: Distribuição das variáveis em estudo.

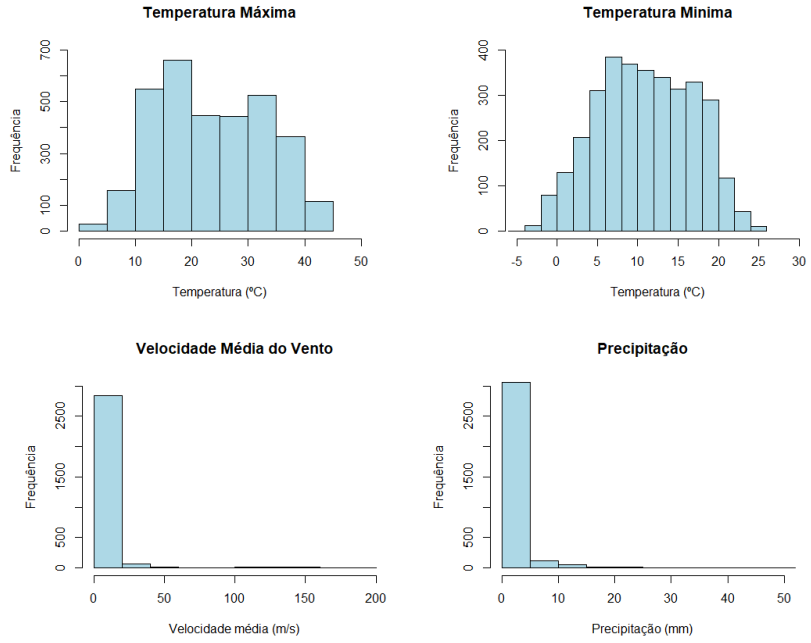


Figura 5.3: Histograma das variáveis em estudo.

Como já referido anteriormente, as variáveis não seguem uma distribuição Normal. Desta forma, o método utilizado para testar a associação entre as variáveis em estudo, é o teste de correlação ordinal de *Spearman*. Na Tabela 5.2 são apresentadas as correlações de *Spearman* para as variáveis em estudo. Como resultado verifica-se que, as temperaturas máximas estão correlacionadas de forma estatisticamente significativa com as temperaturas mínimas e com a precipitação. A precipitação apresenta, ainda, uma correlação significativa com a temperatura mínima. As únicas variáveis que não apresentam uma correlação ordinal estatisticamente significativa são a temperatura máxima e a velocidade média do vento. De referir, ainda, que seria expectável que todas as correlações calculadas fossem estatisticamente significativas, dada a elevada dimensão da amostra em estudo e a natureza das variáveis. A relação ordinal negativa entre a temperatura máxima e a precipitação indica que à medida que as temperaturas máximas aumentam a precipitação diminui.

Tabela 5.2: Correlação ordinal de *Spearman*.

Variável	1	2	3	4
1.Temp. Máxima	1	-	-	-
2.Temp. Mínima	0,87***	1	-	-
3.Precipitação	-0,33***	-0,13***	1	-
4.Velocidade Média do Vento	0,03	0,15***	0,14***	1

** $p <, 01$; ***$p <, 001$

5.1 Análise das Subséries Mensais e Anuais

5.1.1 Temperatura Máxima

Na Figura 5.4 está representada a série temporal da temperatura máxima no período total observado: entre 1 de janeiro de 2010 a 23 de abril de 2019. A representação gráfica sugere que a série temporal exibe um forte comportamento sazonal, como expectável dada a natureza dos dados. Os dados diários exibem uma forte sazonalidade anual (período de 365 dias, já que os dias 29 de fevereiro foram excluídos do período de observação) com picos de temperatura nas estações quentes.

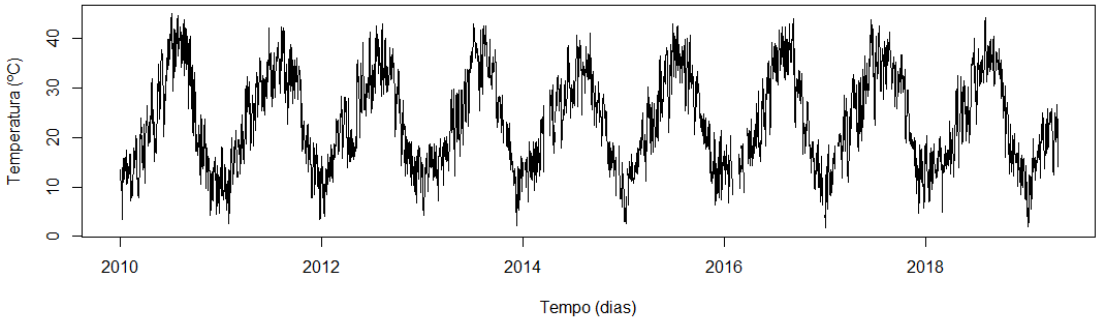


Figura 5.4: Série temporal da distribuição diária da temperatura máxima, para o período observado.

A Tabela 5.3 apresenta as estatísticas descritivas da série temporal da temperatura máxima diária para o período observado, por mês. Como esperado, a média das temperaturas é superior no meses de Verão e apresenta valores inferiores nos meses de Inverno. O desvio padrão mensal indica uma variabilidade superior durante o mês de maio, junho, julho e outubro. Realça-se ainda, que nos meses de Verão não se verificam valores em falta. Observando as amplitudes, destaca-se o mês de outubro que apresenta uma amplitude de 26,20 ºC.

Tabela 5.3: Estatísticas descritivas das temperaturas máximas diárias das subséries mensais.

Mês	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assim.	Curt.	Total (em dias)	NA's (114)
janeiro	1,70	20,50	18,80	9,80	12,70	14,70	12,07	3,80	-0,54	-0,06	408	16
fevereiro	2,50	24,90	22,40	12,60	15,20	17,70	15,19	3,60	-0,21	0,45	363	33
março	7,90	30,30	22,40	15,70	18,40	22,23	18,85	4,30	0,18	-0,52	394	30
abril	10,80	32,80	22,00	18,12	22,00	25,98	22,18	4,89	0,09	-0,83	400	7
maio	15,10	38,00	22,90	22,40	26,80	30,30	26,80	5,18	-0,16	-0,83	393	0
junho	19,00	43,80	24,80	28,02	31,60	35,77	31,78	5,32	-0,08	-0,62	384	0
julho	22,80	45,00	22,20	32,75	35,80	38,60	35,60	4,29	-0,24	-0,38	393	0
agosto	24,60	44,30	19,70	33,20	36,20	39,00	35,93	3,95	-0,24	-0,40	393	0
setembro	20,20	44,10	23,90	26,50	32,45	35,08	32,03	4,62	-0,30	-0,18	384	0
outubro	10,60	36,80	26,20	20,90	24,40	27,90	24,55	5,30	0,03	-0,44	383	10
novembro	4,20	23,80	19,60	13,90	16,50	18,88	16,33	3,63	-0,34	0,05	384	0
dezembro	2,20	21,40	19,20	9,70	12,60	14,90	12,28	3,66	-0,24	-0,48	375	18

Pela observação da Figura 5.5 as subséries mensais apresentam um padrão sazonal evidente. As temperaturas medianas, como expectável, vão aumentanto de forma crescente à medida que se passa dos meses de Inverno para os de Verão. De facto, pela representação gráfica é possível observar que as temperaturas mais altas se verificam nos meses de julho e agosto. Consegue-se, ainda, identificar *outliers* em alguns meses (janeiro, fevereiro, julho, setembro e novembro).

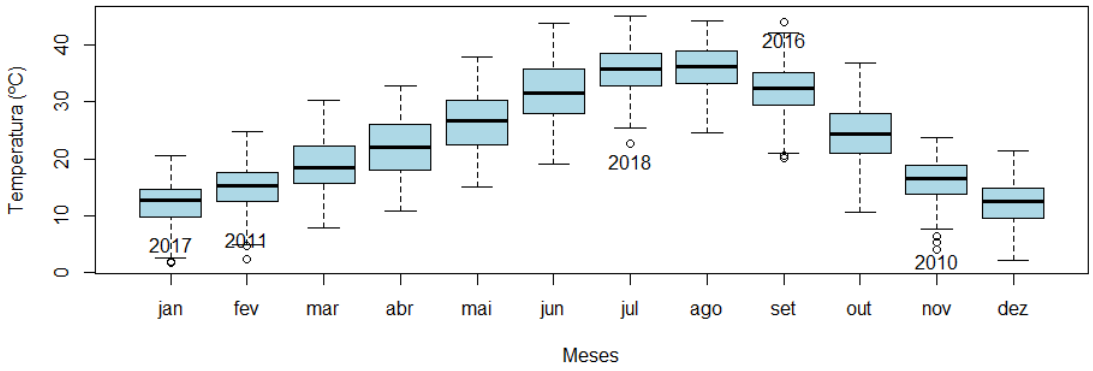


Figura 5.5: Diagramas em caixa de bigodes para as subséries mensais da temperatura máxima, no período observado.

A Tabela 5.4 apresenta as estatísticas descritivas das subséries temporais das temperaturas máximas diárias, no período observado, por ano. O ano 2019 apenas tem observações registadas até dia 23 de abril, o que leva a que os valores das estatísticas nesse ano não correspondam a um ano completo, o que deve ser tido em consideração na comparação dos valores dos restantes anos. De uma forma geral, as temperaturas médias ao longo do ano não apresentam grandes diferenças. Variam entre os 23 ºC e 25 ºC. Contudo, é possível afirmar que o ano 2017 foi o mais quente, com uma média de 25,34 ºC e com uma amplitude térmica elevada. De facto, e "segundo a NASA e a Organização Mundial de Meteorologia (OMM), 2017 foi o segundo ano mais quente desde 1880, ano quando começaram a ser feitos os registos" (João Dias, (2018)).

Tabela 5.4: Estatísticas descritivas das temperaturas máximas diárias por ano.

Ano	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assi.	Curt.	Total (em dias)	NA's (114)
2010	3,40	45,00	41,60	14,15	20,65	33,48	23,32	10,79	0,35	-1,16	360	5
2011	2,50	42,30	39,80	16,00	25,60	32,50	24,19	9,51	-0,17	-1,09	361	4
2012	4,10	43,00	38,90	16,30	22,80	31,20	23,54	9,12	0,14	-1,02	364	2
2013	2,20	43,10	40,90	14,88	21,90	32,15	23,20	9,86	0,22	-1,09	356	9
2014	5,90	41,10	35,20	16,75	25,30	30,35	24,09	8,33	-0,12	-0,96	327	38
2015	2,70	43,10	40,40	15,75	23,00	31,75	23,52	9,56	0,05	-0,99	363	2
2016	3,20	44,10	41,40	16,65	23,10	32,75	24,52	9,70	0,17	-1,12*	322	44
2017	1,70	43,80	42,10	18,15	26,20	33,10	25,34	9,36	-0,15	-1,00	355	10
2018	5,00	44,30	39,30	15,20	21,10	32,30	23,37	9,33	0,26	-1,23	365	0
2019	1,90	26,60	24,70	13,90	17,50	21,70	17,17	5,84	-0,70	0,001	113	0

Ter em consideração que no ano 2019, apenas existem observações até dia 23 de abril.

5.1.2 Temperatura Mínima

Na Figura 5.6 está representada a série temporal da temperatura mínima no período total observado (1 de janeiro de 2010 a 23 de abril de 2019). Novamente, a representação gráfica sugere um comportamento fortemente sazonal da temperatura: uma sazonalidade de 365 dias. A Tabela 5.5 apresenta as estatísticas descritivas da

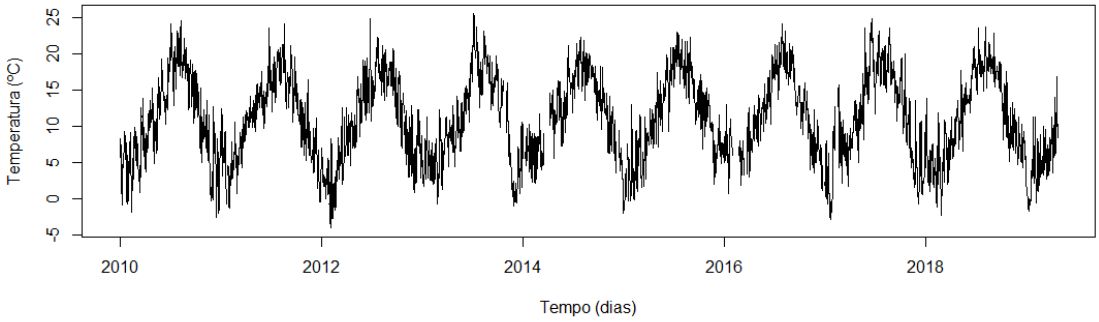


Figura 5.6: Série temporal da distribuição diária da temperatura mínima, para o período observado.

série temporal das temperaturas mínimas diárias, no período observado, por mês. O desvio padrão mensal indica uma variabilidade superior durante o mês de novembro, dezembro, janeiro e fevereiro. Assim, a variabilidade do comportamento da série parece ser superior nas estações frias relativamente à variabilidade das estações quentes. O maior valor do coeficiente de variação para a temperatura mínima é de 79,00% no mês de fevereiro e o menor de 11,90 ºC%, no mês de agosto. A temperatura mínima apresenta uma distribuição quase simétrica (apresenta valores perto de zero do coeficiente de assimetria amostral), por mês. Novamente, não há valores em falta nos meses de Verão.

Tabela 5.5: Estatísticas descritivas das temperaturas mínimas diárias, das subséries mensais.

Mês	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assi.	Curt.	Total (em dias)	NA's (114)
janeiro	-2,90	13,80	16,70	1,70	4,40	6,90	4,41	3,47	0,20	-0,60	408	16
fevereiro	-4,10	15,70	19,80	2,10	4,40	6,60	4,38	3,46	0,21	0,14	363	33
março	0,70	13,90	13,20	4,70	6,55	8,43	6,60	2,67	0,02	-0,50	394	30
abril	3,00	16,80	13,80	7,90	9,50	11,38	9,53	2,36	-0,07	-0,26	400	7
maio	4,80	23,50	18,70	10,40	12,40	14,30	12,27	2,93	0,12	0,29	393	0
junho	9,10	24,90	15,80	13,50	15,60	17,80	15,81	3,08	0,40	-0,20	384	0
julho	10,70	25,50	14,80	16,85	18,60	20,30	18,56	2,65	-0,05	-0,07	393	0
agosto	12,40	24,50	12,10	17,05	18,50	20,00	18,49	2,20	-0,04	-0,10	393	0
setembro	10,20	22,90	12,70	14,43	16,30	17,70	16,00	2,39	-0,27	-0,39	384	0
outubro	5,00	20,00	15,00	10,40	12,40	14,50	12,35	2,90	-0,07	-0,31	383	10
novembro	-1,10	16,40	17,50	5,45	8,00	10,38	7,83	3,45	-0,15	-0,45	384	0
dezembro	-2,60	12,70	15,30	2,80	5,50	7,40	5,23	3,30	-0,15	-0,64	375	18

A Figura 5.7 apresenta os diagramas em caixa de bigodes das temperaturas mínimas diárias, por mês. São identificados alguns *outliers* nos meses de fevereiro, abril, maio, junho, julho, agosto e setembro. As temperaturas mínimas alcançam os valores mais baixos nos meses de janeiro, fevereiro e dezembro, como expectável.

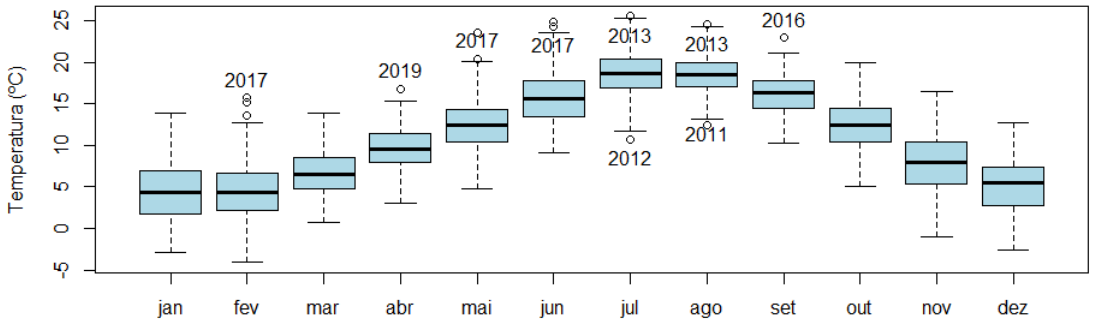


Figura 5.7: Diagramas em caixas de bigodes para as subséries mensais da temperatura mínima, no período observado.

A Tabela 5.6 apresenta as estatísticas descritivas da série temporal das temperaturas mínimas diárias, no período observado, por ano. No ano de 2010 foi registada a temperatura mínima de $-2,90^{o}C$, a temperatura mais baixa observada. Cerca de 75% das temperaturas mínimas no ano 2017 foram superiores a 6,45 ºC e o desvio padrão anual indica uma variabilidade superior nos anos de 2013 e 2017.

Tabela 5.6: Estatísticas descritivas das temperaturas mínimas diárias por ano.

Ano	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assim.	Curt.	Total (em dias)	NA's (114)
2010	-2,60	24,50	27,10	6,68	10,65	16,10	10,99	6,08	-0,01	-0,79	360	5
2011	-1,30	24,20	25,50	6,90	11,40	15,00	11,10	5,36	-0,12	-0,71	361	4
2012	-4,10	24,90	29,00	5,70	10,50	15,70	10,30	6,18	-0,11	-0,93	364	2
2013	-1,00	25,50	26,50	5,80	10,10	15,80	10,56	6,28	0,14	-0,93	356	9
2014	-2,10	22,30	24,40	7,95	12,10	16,10	12,01	5,06	-0,16	-0,81	327	38
2015	-1,50	23,00	24,50	7,15	11,20	15,90	11,29	5,84	-0,11	-0,86	363	2
2016	0,70	24,20	23,50	7,90	11,80	16,45	11,96	5,29	0,14	-0,85	322	44
2017	-2,90	24,90	27,80	6,45	11,40	16,90	11,40	6,52	-0,09	-0,96	355	10
2018	-2,40	23,70	26,10	6,80	10,80	16,80	11,27	5,78	0,02	-0,98	365	0
2019	-1,70	16,80	18,50	62,80	5,40	7,70	5,18	3,66	0,27	-0,16	113	0

Ter em consideração que no ano 2019, apenas existem observações até dia 23 de abril.

5.1.3 Precipitação

Na Figura 5.8 está representada a série temporal da precipitação para o período total em estudo. Esta série apresenta uma enorme variabilidade ao longo do período

observado, com vários valores extremos.

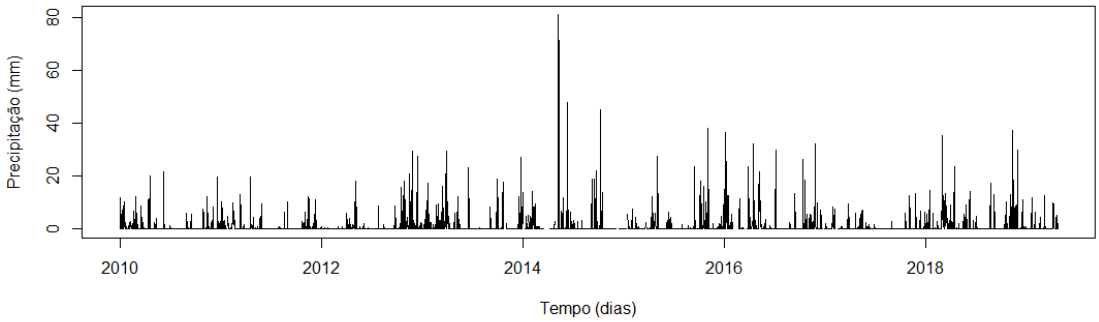


Figura 5.8: Série temporal da distribuição diária da precipitação, para o período observado.

A Tabela 5.7 apresenta as estatísticas descritivas da série temporal da precipitação diária durante o período observado por mês. Para a análise descritiva dos dados foram considerados apenas os dias em que ocorreu precipitação, no total 882 dias (25,94% do período observado). O desvio padrão mensal indica uma variabilidade superior durante o mês de fevereiro e março. Os maiores valores médios de precipitação são observados no mês de fevereiro e outubro. Os meses que apresentam valores em falta são os meses de janeiro, fevereiro, março, outubro e dezembro. O mês fevereiro apresenta o maior coeficiente de variação amostral (CV=191,89) e o mês de janeiro o menor (CV=102,17)

Tabela 5.7: Características amostrais da precipitação das subséries mensais.

Mês	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assim.	Curt.	Total (em dias)	NA's (114)
janeiro	0,20	14,60	14,40	0,80	2,20	5,25	3,69	3,77	1,32	0,82	230	16
fevereiro	0,20	80,80	80,60	1,00	2,60	6,40	5,92	11,36	4,99	28,78	198	33
março	0,20	62,40	62,20	0,80	2,80	6,60	6,31	10,37	3,30	12,34	203	30
abril	0,20	37,20	37,00	0,25	2,30	6,20	4,35	6,18	3,10	12,16	206	7
maio	0,20	27,40	27,20	0,60	1,60	4,60	3,45	4,46	2,59	8,88	179	0
junho	0,20	38,00	37,80	0,40	1,90	6,10	4,45	6,76	2,68	8,83	152	0
julho	0,20	36,40	36,20	0,40	1,10	3,80	3,86	6,60	2,99	9,95	128	0
agosto	0,20	20,80	20,60	0,60	2,80	7,60	4,77	5,61	1,33	0,74	130	0
setembro	0,20	32,00	31,80	0,55	1,80	7,30	5,79	8,32	1,72	1,96	143	0
outubro	0,20	29,80	29,60	0,60	2,60	5,75	5,09	7,01	2,13	4,08	187	10
novembro	0,20	32,20	32,00	0,80	2,40	5,80	4,99	6,29	2,21	5,62	196	0
dezembro	0,20	27,00	26,80	0,80	2,60	6,90	4,87	5,75	1,88	3,74	193	18

Por observação da Figura 5.9 identifica-se a presença de muitos *outliers* em todos os meses do ano. As precipitações medianas mantêm-se aproximadamente constantes ao longo dos meses (variam entre 1,10 mm e 2,80 mm).

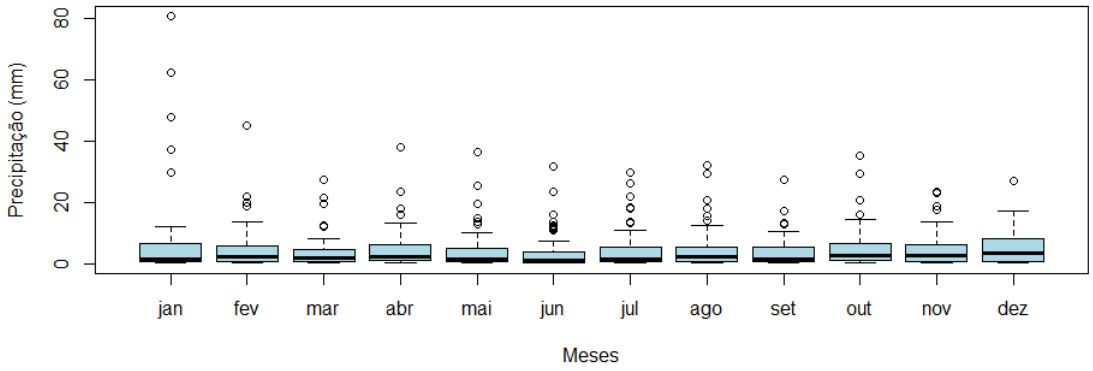


Figura 5.9: Diagramas em caixa de bigodes para as subséries mensais da precipitação, no período observado.

A Tabela 5.8 apresenta as estatísticas descritivas da série temporal da precipitação diária durante o período observado, por ano. Como referido anteriormente, para o cálculo das estatísticas descritivas foram considerados apenas os dias em que houve precipitação. O valor mínimo de precipitação é de 0,20 mm (observado em vários anos), sendo o valor máximo de precipitação registado no ano de 2014 (80,80 mm).

Tabela 5.8: Características amostrais da precipitação por ano.

Ano	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assim.	Curt.	Total (em dias)	NA's (114)
2010	0,20	21,60	21,40	0,80	2,10	4,90	3,71	4,57	2,04	4,18	360	5
2011	0,20	19,60	19,40	0,40	1,60	4,70	3,18	3,79	1,81	3,48	361	4
2012	0,20	29,60	29,40	0,40	1,20	3,50	3,65	6,02	2,61	6,63	364	2
2013	0,20	29,60	29,40	0,60	2,60	7,60	5,10	6,36	1,81	3,02	356	9
2014	0,20	80,80	80,60	0,80	2,60	6,75	7,17	13,77	3,55	13,40	327	38
2015	0,20	38,00	37,80	0,40	1,90	4,50	4,05	6,43	3,03	10,54	363	2
2016	0,20	36,40	36,20	0,80	2,80	8,20	6,41	8,24	1,83	2,84	322	44
2017	0,20	13,40	13,20	0,80	2,20	5,15	3,39	3,28	1,27	1,01	355	10
2018	0,20	37,20	37,00	1,00	2,90	7,90	5,31	6,72	2,57	8,17	365	0
2019	0,20	12,60	12,40	0,40	2,20	5,00	3,62	3,86	0,97	-0,26	113	0

Ter em consideração que no ano 2019, apenas existem observações até dia 23 de abril.

5.1.4 Velocidade Média do Vento

Na Figura 5.10 está representada a série temporal da velocidade média do vento para o período total observado. A representação gráfica mostra claramente a existência de muitos valores em falta. A velocidade média do vento apresenta um comportamento bastante particular. É visível no ano de 2014, um período com uma enorme variabilidade e apresentando valores muito elevados de velocidade média do

vento.

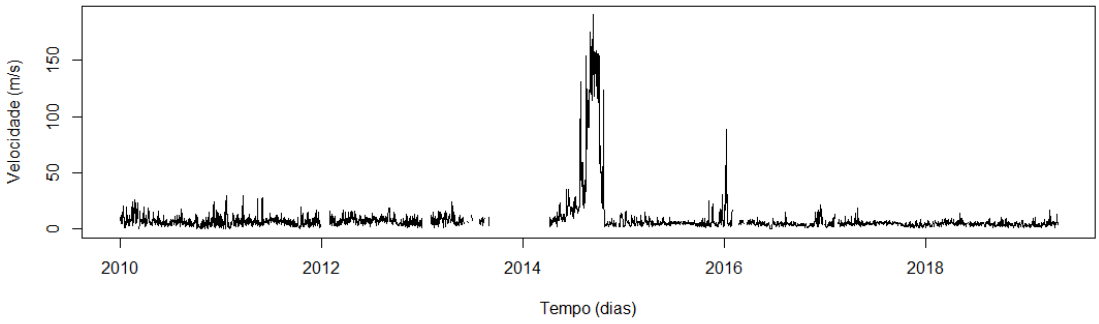


Figura 5.10: Série temporal da distribuição diária da velocidade média do vento, para o período observado.

Na Tabela 5.9 são apresentadas as estatísticas descritivas da série temporal da velocidade média do vento diária durante o período observado, por cada mês do ano. O desvio padrão do mês de setembro destaca-se dos restantes meses, uma vez que tem um valor muito superior. Contudo, em termos de variação relativa é o mês de outubro que tem um coeficiente de variação superior (CV=234%).

Tabela 5.9: Características amostrais da velocidade média do vento das subséries mensais.

Mês	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assi.	Curt.	Total (em dias)	NA's (114)
janeiro	0,20	88,40	88,20	3,00	4,80	7,80	7,20	9,13	5,23	36,90	310	100
fevereiro	0,40	25,90	25,50	3,60	5,40	7,90	6,51	4,31	1,79	4,14	282	61
março	0,50	29,70	29,20	4,50	6,10	8,48	7,01	3,91	1,84	5,34	310	48
abril	1,30	24,00	22,70	4,70	6,15	8,60	6,92	3,28	1,45	3,12	293	7
maio	1,90	27,10	25,20	4,30	5,70	7,90	6,55	3,40	2,22	7,72	393	-
junho	0,00	35,70	35,70	4,30	5,40	8,10	7,22	5,76	2,48	7,28	384	25
julho	1,90	130,70	128,80	4,90	6,20	8,20	8,68	10,76	7,18	68,47	279	26
agosto	1,80	154,40	152,60	4,70	6,00	8,20	13,77	24,93	3,53	12,10	279	14
setembro	1,30	190,40	189,10	3,78	5,10	6,93	22,90	47,49	2,34	3,68	384	30
outubro	0,00	155,90	155,90	2,60	3,80	5,90	9,66	22,56	4,64	22,41	383	32
novembro	0,60	25,10	24,50	2,98	4,45	6,50	5,34	3,70	2,27	7,08	384	30
dezembro	0,80	30,30	19,50	2,90	4,65	8,10	4,65	4,89	1,60	2,81	375	49

Na Figura 5.11 estão representados os diagramas em caixa de bigodes da precipitação diária por mês, identificando-se um grande número de *outliers*, em todos os meses do ano.

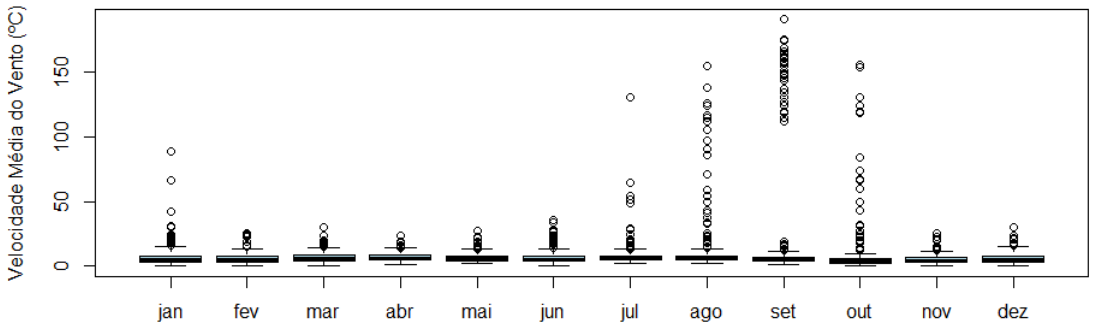


Figura 5.11: Diagramas em caixa de bigodes para as subséries mensais da velocidade média do vento, no período observado.

Na Tabela 5.10 estão descritos os valores das estatísticas descritivas da série temporal da velocidade média do vento diária, no período observado, por ano. Em 2014 destaca-se o valor médio muito superior aos restantes anos, bem como um desvio padrão de 50,49 m/s e uma amplitude 180,50 m/s, o que representa uma enorme variabilidade de velocidade média do vento neste ano.

Tabela 5.10: Características amostrais da velocidade média do vento por ano.

Ano	Mín.	Máx.	Ampli.	1.ºQuart.	Med.	3.ºQuart.	Média	Desvio Padrão	Assim.	Curt.	Total (em dias)	NA's (422)
2010	0,00	25,90	25,90	3,88	6,00	8,60	6,99	4,67	1,47	2,42	360	5
2011	0,20	29,70	29,50	4,00	6,10	8,20	6,65	4,27	2,31	8,72	361	4
2012	0,90	18,90	18,00	4,83	7,00	9,40	7,38	3,32	0,77	0,62	334	31
2013	1,20	24,00	22,80	5,30	7,80	10,35	8,19	3,71	0,80	1,26	147	218
2014	1,90	190,40	188,50	6,80	16,20	49,15	39,75	50,49	1,44	0,60	258	107
2015	1,40	30,30	28,90	3,95	5,20	6,30	5,80	3,46	2,98	12,79	363	2
2016	0,00	88,40	88,40	3,50	4,60	6,08	6,03	7,30	6,99	64,76	322	43
2017	0,80	18,50	17,70	3,50	4,70	5,70	4,78	2,11	1,64	6,73	355	10
2018	1,10	14,10	13,00	3,40	4,60	5,60	4,66	1,74	0,87	2,10	365	-
2019	1,30	17,10	15,80	3,10	4,20	5,50	4,66	2,42	1,96	6,18	113	-

Ter em consideração que no ano 2019, apenas existem observações até dia 23 de abril.

Capítulo 6

Aplicação dos Modelos TBATS e de Regressão com Erros Correlacionados

Após realizada a análise preliminar das séries temporais e conhecendo melhor as suas características, procede-se à aplicação aos dados das metodologias descritas no Capítulo 4.

As séries temporais analisadas apresentam correlação temporal forte e sazonalidade complexa. Assim, é necessário aplicar modelos flexíveis que permitam integrar as várias características dos dados e que contemplem a incorporação destas componentes.

O modelo TBATS e o modelo de regressão com erros correlacionados possuem esta flexibilidade pela sua estrutura versátil, que permite integrar as várias características dos dados e podem ser aplicados a séries não estacionárias (como as deste estudo).

Assim, neste Capítulo são apresentados os resultados da aplicação destes modelos para as séries temporais da temperatura máxima, da temperatura mínima, da precipitação e da velocidade média do vento. As séries temporais foram divididas em série de treino e série de teste com o objetivo de testar a precisão dos modelos de previsão propostos. O período da série de treino foi de 1 de janeiro de 2010 a 31 de dezembro de 2017 (2923 dias) e o período da série de teste foi de 1 de janeiro de 2018 a 23 de abril de 2019 (os últimos 477 dias). Esta abordagem possibilita a comparação das medidas de avaliação da qualidade das predições e previsões dos diferentes métodos. Todo o processo de estimação e previsão dos modelos foi efetuado com recurso ao *software* R, *package forecast* Hyndman et al. (2019). Além da utilização

de funções já existentes, foi necessária a implementação de algumas novas.

As observações em falta condicionam a utilização de alguns métodos de modelação e previsão. Neste caso, a função do *software* R construída para o modelo TBATS não suporta séries que contenham dados omissos. Por predefinição, a função *tbats* (para implementar o modelo TBATS) e de forma automática seleciona parte da série que não contenha quaisquer valores em falta e modela apenas para essa subsérie. Mas não é isso que se pretende neste estudo, assim foi necessário proceder a um método de imputação de dados utilizando a função do *software R*, *na.interp*, já abordada no Capítulo 4. Trata-se de um método completamente automático que estima os dados em falta através de uma decomposição STL, Hyndman et al. (2019).

Neste Capítulo serão apresentados os resultados da modelação TBATS e da regressão com erros correlacionados para cada uma das séries temporais. O desempenho destes processos de modelação, para cada uma das séries temporais em estudo (da temperatura mínima, da temperatura máxima, da precipitação e da velocidade média do vento) foi avaliado através de medidas de avaliação e pela análise dos resíduos.

Neste estudo considera-se, em todas as decisões, um nível de significância de 5%.

6.1 Temperatura Mínima

Para completar a série temporal da temperatura mínima, com 114 valores em falta, procedeu-se à imputação de dados pelo método já descrito anteriormente. Na Figura 6.1 está representada a série temporal da temperatura mínima já completa (a vermelho encontram-se os valores calculados dos dados em falta).

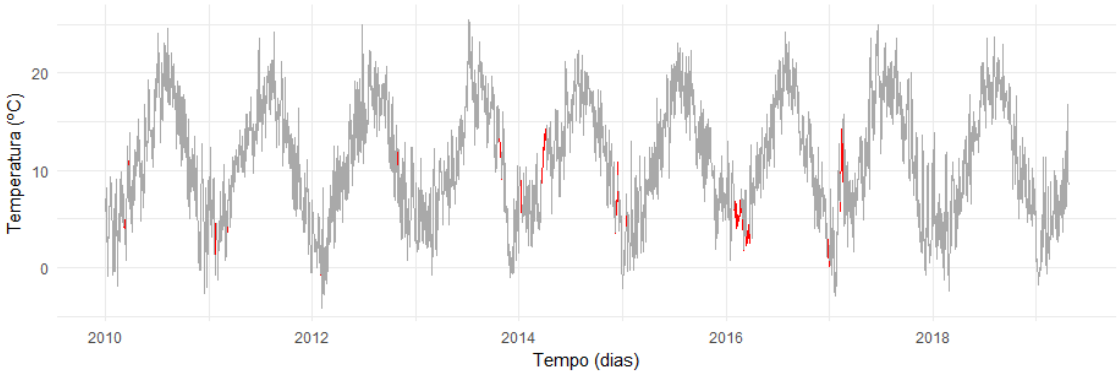


Figura 6.1: Série valores originais e valores imputados (a vermelho) da temperatura mínima.

Modelo TBATS

Os dados da temperatura mínima são observados de forma diária e apresentam um forte padrão sazonal anual, ou seja, um período sazonal de 365 dias ($m_1 = 365$). A série vai ser definida por um modelo de decomposição aditivo, pois a amplitude das oscilações sazonais não varia com o nível da série. Na Figura 6.2 está representada a divisão da série temporal (2932 dias para treino e 477 dias para teste). Para analisar a dependência temporal das séries foram calculadas as funções de autocorrelação (FAC) e autocorrelação parcial(FACP) amostrais da série de treino da temperatura mínima.

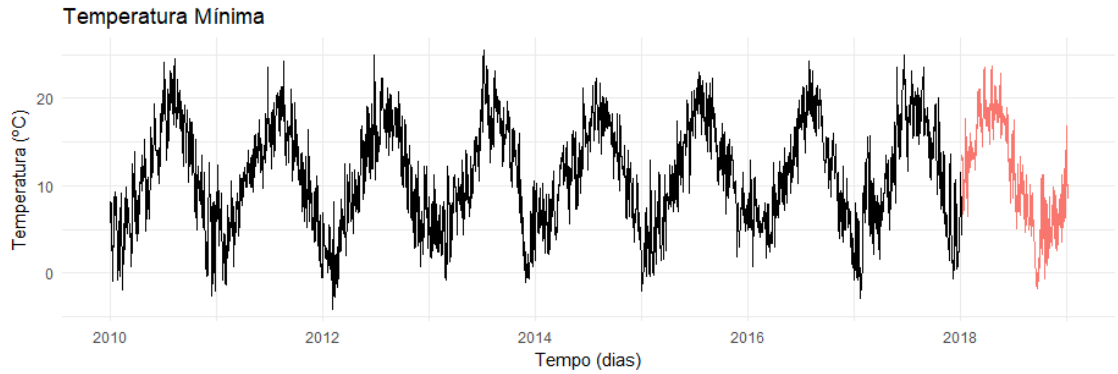


Figura 6.2: Série de treino (a preto) e série de teste (a vermelho) da temperatura mínima.

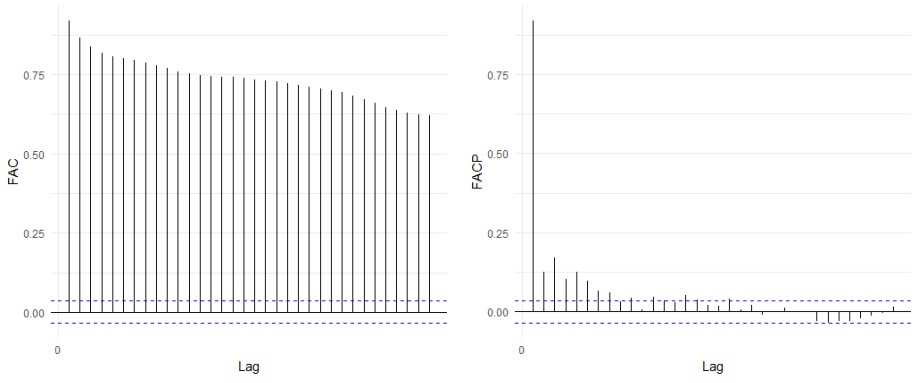


Figura 6.3: FAC e FACP da série de treino da temperatura mínima

O modelo TBATS é automático, desta forma, não exige a pré-especificação dos valores iniciais, uma vez que o modelo os calcula de forma automática. A função *tbats* aplicada à série de treino calcula as estimativas de máxima verosimilhança dos valores de estados iniciais, as estimativas dos parâmetros de alisamento, seleciona o número de harmónicos necessário para modelar a componente sazonal e escolhe a ordem (p, q) do processo ARMA.

Assim, o modelo obtido que minimiza o critério de informação (AIC) é o modelo $TBATS(1, \{0,5\}, 1, \{365,1\})$. O parâmetro $\omega = 1$ indica que não foi aplicada uma transformação Box-Cox aos dados da série temporal, ou seja, não foi necessária uma tranformação para lidar com possíveis não linearidades da série. Os valores estimados $\beta = 0$ e $\phi = 1$ traduzem uma taxa de crescimento puramente determinística, sem efeito de amortecimento. A componente irregular da série é correlacionada e modelada por um processo ARMA(0,5). As estimativas dos parâmetros do modelo TBATS para a temperatura mínima são apresentadas na Tabela 6.1.

Tabela 6.1: Parâmetros do modelo TBATS selecionado.

Parâmetros	α	γ_1	γ_2	θ_1	θ_2	θ_3	θ_4	θ_5
Estimativas	0,04794	0,00011	-0,00014	0,64795	0,36503	0,23238	0,11614	0,03265

A estimativa do desvio padrão do ruído branco associado ao modelo é $\hat{\sigma} = 2,1273$ e o valor de AIC$= 27771,49$.

Desta forma, as equações do modelo TBATS para a temperatura mínima são as seguintes:

$$y_t^{(1)} = l_{t-1} + b_{t-1} + s_{t-365}^{(1)} + d_t,$$

$$l_t = l_{t-1} + b_{t-1} + 0,04794 d_t,$$

$$b_t = b_{t-1},$$

$$s_{j,t}^{(1)} = s_{j,t-1}^{(1)} \cos \lambda j^{(1)} + s_{j,t}^{*(1)} \sin \lambda_j^{(1)} + 0,00011 d_t,$$

$$s_{j,t}^{*(1)} = -s_{j,t-1} \sin \lambda_j^{(1)} + s_{j,t-1}^{*(1)} \cos \lambda_j^{(1)} + (-0,0014) d_t,$$

onde $\lambda_j^{(1)} = \frac{2\pi j}{365}$, d_t é um processo ARMA(0,5) e α, β, γ_1 e γ_2 são os parâmetros de alisamento. No processo de modelação da sazonalidade foram utilizados 19 parâmetros (17 valores iniciais para $s_{j,0}^{(1)}$ e $s_{j,0}^{*(1)}$ e dois parâmetros de alisamento, $\gamma_1^{(1)}$ e $\gamma_2^{(1)}$).

A equação do processo ARMA(0,5), ou seja, do processo de média móveis MA(5) é dada por

$$d_t = 0,64795\, \varepsilon_{t-1} + 0,36503\, \varepsilon_{t-2} + 0,23238\, \varepsilon_{t-3} + 0,11614\, \varepsilon_{t-4} + 0,03265\, \varepsilon_{t-5} + \varepsilon_t,$$

onde $\varepsilon_t \sim N(0, (2,1273)^2)$.

Na Figura 6.4 estão representados os valores observados da série temporal (a preto) e os valores estimados pelo modelo TBATS (a vermelho), no período de treino. Por observação gráfica, o modelo ajusta-se de forma bastante satisfatória

aos dados.

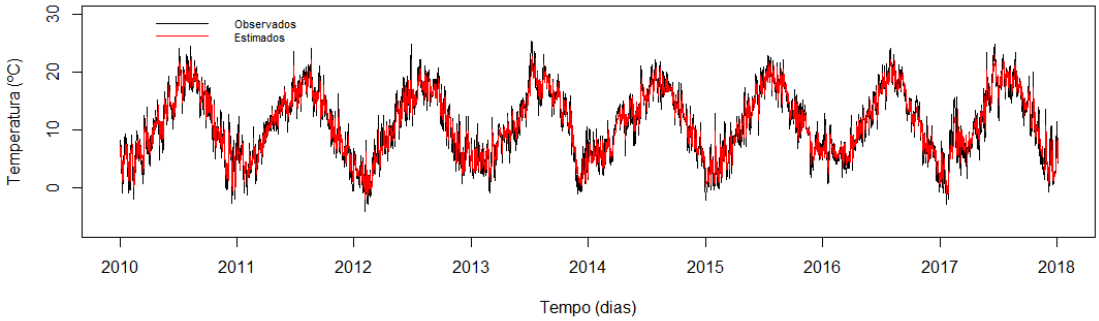


Figura 6.4: Valores observados e valores estimados pelo modelo TBATS.

Após encontrado o modelo TBATS para a série de treino foram calculadas as previsões para o período de teste. Ou seja, foram calculadas as previsões para um passo de $h = 477$ dias. Na Figura 6.5 estão representados os valores observados da temperatura mínima, as estimativas no período de modelação (período de treino), as previsões no período de previsão (período de teste) e os intervalos de previsão para um nível de confiança de 80% e 95%, resultantes da aplicação do modelo TBATS.

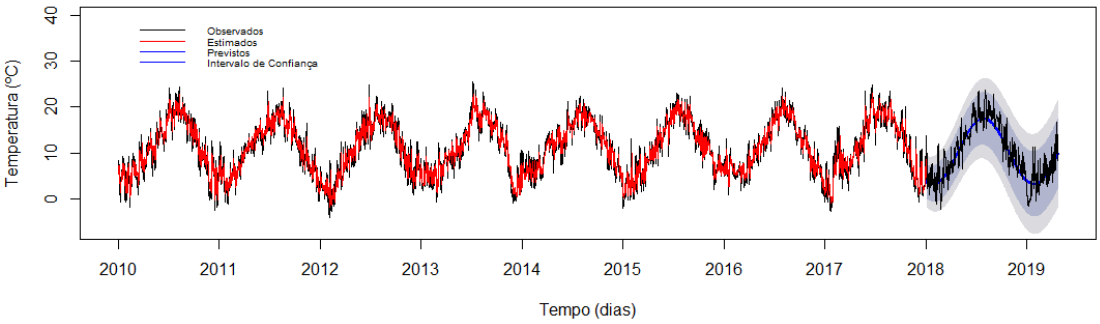


Figura 6.5: Valores observados, estimados e previstos (com intervalos de confiança de 80% e 95%) para a temperatura mínima resultante do modelo TBATS.

Para facilitar a visualização gráfica, será apresentada uma janela de visualização das 250 últimas observações. A Figura 6.6 mostra os valores observados e previsões (com limites de confiança de 80% e 95%) para a temperatura mínima resultantes do modelo TBATS. A maioria das observações encontra-se dentro dos intervalos de previsão, com exceção do primeiro trimestre onde há alguns valores que estão fora dos limites do intervalo de previsão. Na Tabela 6.2 são apresentadas as previsões e os respetivos intervalos de previsão para os primeiros 7 dias do período de teste. Percebe-se que os primeiros três valores observados não pertencem aos intervalos de

previsão. Os intervalos de previsão apresentam uma amplitude elevada. Wang e Cai (2009) referem que tal não é de estranhar, uma vez que, os intervalos de previsão tendem a ser conservadores, i.e., com margens de erro grandes.

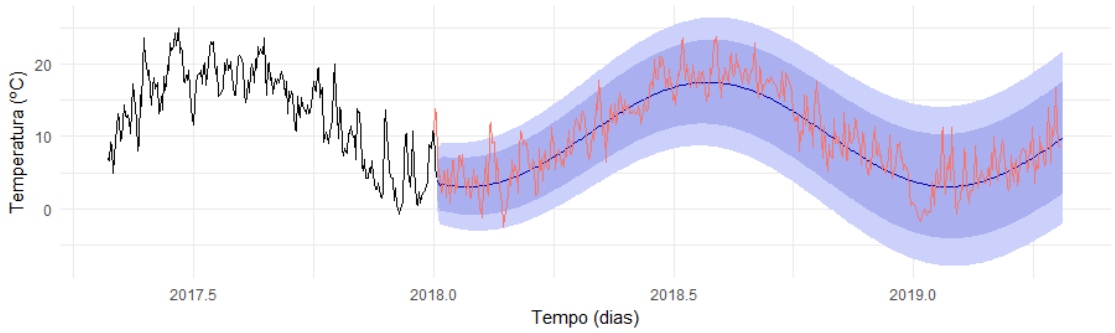


Figura 6.6: Valores observados e previsões (com limites de confiança de 80% e 95%) para a temperatura mínima resultantes do modelo TBATS.

Tabela 6.2: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos às temperaturas mínimas.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
3,872	(1,145; 6,598)	(-0,298; 8,041)	10,40
3,477	(0,155; 6,798)	(-1,603; 8,557)	13,80
3,324	(-0,212; 6,859)	(-2,083; 8,731)	12,90
3,399	(-0,248; 7,047)	(-2,179; 8,978)	7,20
3,479	(-0,228; 7,168)	(-2,186; 9,126)	4,50
3,446	(-0,274; 7,165)	(-2,243; 9,135)	4,60
3,409	(-0,324; 7,142)	(-2,300; 9,118)	2,20

Para validar o modelo escolhido é necessário realizar uma análise de resíduos. Idealmente, estes devem apresentar um comportamento próximo de uma distribuição Normal, de média nula e variância constante e não apresentar correlação temporal. O histograma da Figura 6.7 sugere que os resíduos têm uma distribuição ligeiramente assimétrica à esquerda, apesar disso, o teste de Kolmogorov-Smirnov não rejeita a hipótese da normalidade (valor de prova = 0,08776), para um nível de significância de 5%. Além disso, de acordo com a representação gráfica da série dos resíduos, estes assumem valores em torno de zero, com variância constante e sem correlação temporal. Quanto à independência, o teste de Ljung-Box é aplicado à série de resíduos e os resultados do teste mostram que a independência não é rejeitada (valor de prova = 0,06373).

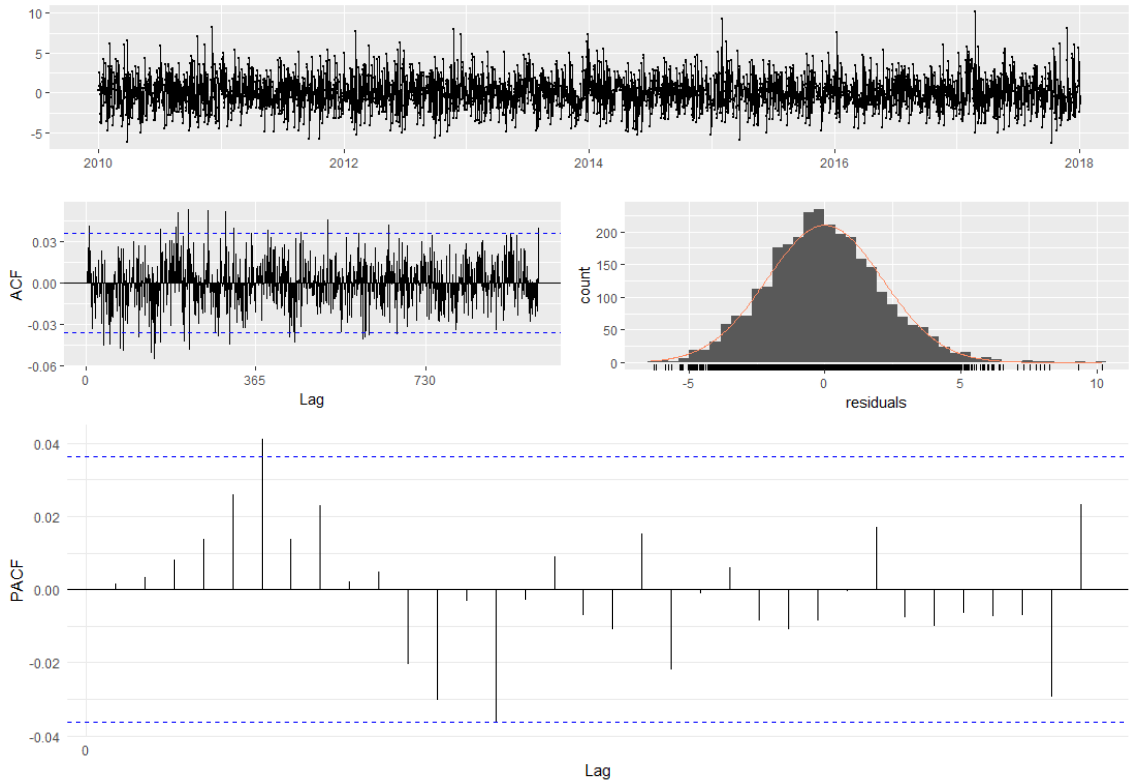


Figura 6.7: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo TBATS da temperatura mínima.

A decomposição da série da temperatura mínima, obtida dos valores ajustados pelo modelo TBATS, é apresentada na Figura 6.8. A presença de uma componente de tendência não é considerada (a série é estacionária na média). No entanto, na decomposição da série é nítida a necessidade de incorporação no modelo de uma componente sazonal.

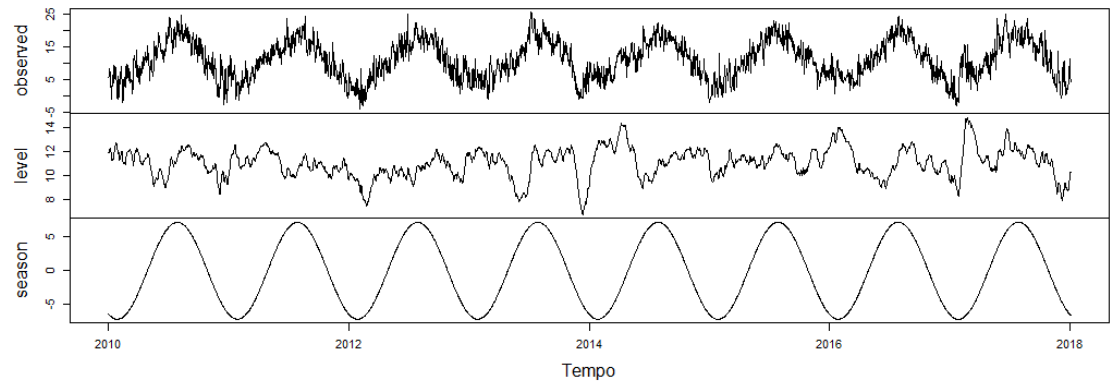


Figura 6.8: Decomposição série temporal temperatura mínima, obtida pela estimação do modelo TBATS.

Modelo de Regressão com erros Correlacionados

Para a modelação TBATS as séries não podem apresentar valores em falta, sendo necessária a utilização de métodos de imputação de dados. No caso dos modelos de regressão esta condição não precisa de ser verificada, pois estes modelos conseguem ser aplicados a séries com valores em falta. Assim, a modelação com modelos de regressão com erros correlacionados vai ser realizada com as séries com lacunas.

Os modelos de regressão com erros correlacionados conseguem modelar dados com padrões sazonais complexos, através da introdução de regressores externos na forma de termos de Fourier. Assim, o modelo de regressão considerado é o modelo na sua forma básica $y_t = b_t + s_t + \varepsilon_t$, onde b_t e s_t representam as componentes tendência e sazonalidade da série temporal no tempo t. Para o modelo de regressão com erros correlacionados é preciso determinar o termo K que representa o número de pares de senos e cosenos de Fourier e a ordem do processo ARIMA. Este termo é determinado fazendo variar $i = 1, \ldots, 5$, através do ajustamento de um modelo ARIMA com regressores externos.

Assim, ajustaram-se cinco modelos ARIMA à série de treino, verificando-se que o modelo com $K = 4$ pares de senos e cosenos apresenta o menor valor do critério de informação AICc. Considerando a série residual com $K = 4$, obtém-se a ordem do modelo ARIMA. Este processo é feito de forma automática com recurso à função *ARIMA* do *package forecast* (Tabela 6.3).

Tabela 6.3: Ordem do processo ARIMA, valor de K, valor AICc e valor BIC.

Modelo	K	AICc	BIC
Regressão com erros ARIMA(0,0,5)	1	12324,47	12378,23
Regressão com erros ARIMA(0,0,5)	2	14603,35	14663,08
Regressão com erros ARIMA(0,0,5)	3	12300,66	12378,28
Regressão com erros ARIMA(0,0,5)	4	**12297, 86**	**12387, 40**
Regressão com erros ARIMA(0,0,5)	5	14614,40	14709,90

O valor de AICc para o qual ε_t é um processo $ARIMA(p, d, q)$ é de 12297,86 e trata-se de um processo ARIMA(0,0,5), ou seja, um processo de médias móveis MA(5).

A Tabela 6.4 apresenta as estimativas dos parâmetros e os respetivos erros padrão, para o modelo de regressão com erros MA(5). No modelo final, a estimativa do desvio padrão do ruído branco associado ao modelo foi de $\hat{\sigma} = 4, 44$ e o valor AIC=12297, 70.

Tabela 6.4: Estimativas dos parâmetros resultantes do modelo de regressão com erros correlacionados e respetivos erro padrão.

Parâmetros	θ_1	θ_2	θ_3	θ_4	θ_5	β_0	α_1
Estimativas	0,6735	0,3983	0,2676	0,1464	0,0504	11,0147	-3,2423
Erro padrão	0,0190	0,0227	0,0222	0,0216	0,0186	0,1020	0,1446
Parâmetros	β_1	α_2	β_2	α_3	β_3	α_4	β_4
Estimativas	-6,4428	0,8064	-0,1582	-0,0926	-0,0185	0,3223	0,1919
Erro padrão	0,1437	0,1440	0,1441	0,1443	0,1434	0,1432	0,1440

Desta forma, o modelo obtido é

$$\begin{aligned} y_t = 11,0147 &+ (-3,2423)\sin\frac{2\pi t}{365} + (-6,4428)\cos\frac{2\pi t}{365} + 0,8064\sin\frac{4\pi t}{365} \\ &+ (-0,1582)\cos\frac{4\pi t}{365} + (-0,0926)\sin\frac{6\pi t}{365} + (-0,0185)\cos\frac{6\pi t}{365} \\ &+ 0,3223\sin\frac{8\pi t}{365} + 0,1919\cos\frac{8\pi t}{365} + d_t, \end{aligned}$$

onde d_t é um processo MA(5) descrito pela seguinte equação

$$d_t = 0,6735\varepsilon_{t-1} + 0,3983\varepsilon_{t-2} + 0,2676\varepsilon_{t-3} + 0,1465\varepsilon_{t-4} + 0.0504\varepsilon_{t-5},$$

onde $\varepsilon_t \sim N(0,(4,44)^2)$.

Na Figura 6.9 estão representados os valores observados da série temporal (a preto) e os valores estimados pelo modelo (a vermelho). Por observação gráfica o modelo parece ajustar-se de forma bastante satisfatória aos dados.

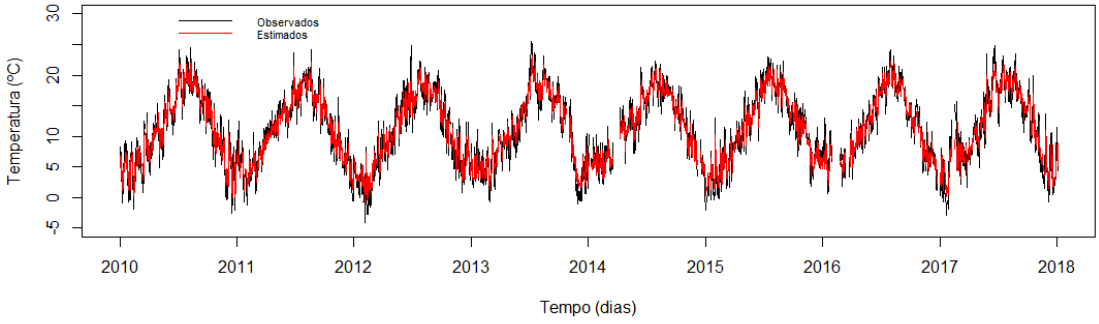


Figura 6.9: Valores observados e valores estimados pelo modelo de regressão com erros correlacionados.

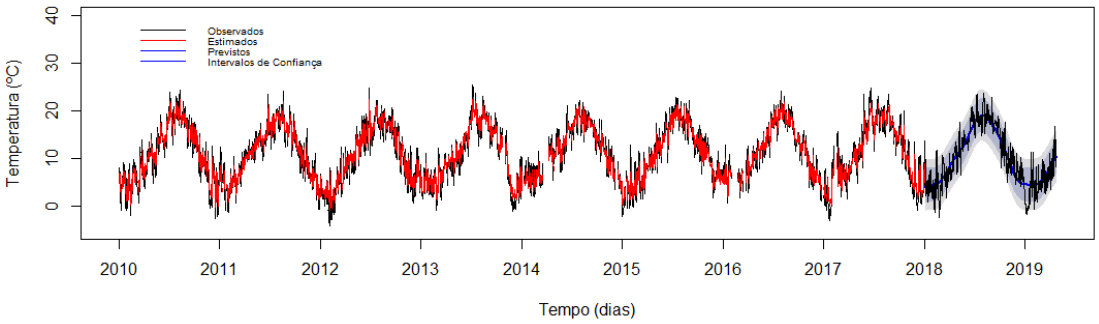


Figura 6.10: Valores observados e previsões (com intervalos de confiança de 80% e 95%) para a temperatura mínima resultante do modelo de regressão com erros correlacionados.

Na Figura 6.10 estão representados os valores observados da temperatura mínima, as estimativas no período de modelação (período de treino), as previsões no período de previsão (período de teste) e os intervalos de previsão para um nível de confiança de 80% e 95%, resultantes da aplicação do modelo de regressão com erros correlacionados. A previsão foi calculada a h=477 passos (dias). A Figura 6.11 mostra as previsões (no período de teste), pontuais e intervalares (80% e 95%) e estimativas pontuais (no período de treino) obtidas através do modelo de regressão com erros correlacionados e de forma mais ampliada (janela com as últimas 250 observações). No primeiro trimestre existem alguns valores que estão fora dos limites do intervalo de previsão, i.e., a taxa de cobertura dos intervalos de previsão não é de 100%. Na Tabela 6.5 são apresentadas as primeiras sete previsões resultantes do modelo obtido de regressão com erros correlacionados. Paralelamente ao que aconteceu com o modelo TBATS, os três primeiros valores observados estão fora dos limites dos intervalos de previsão.

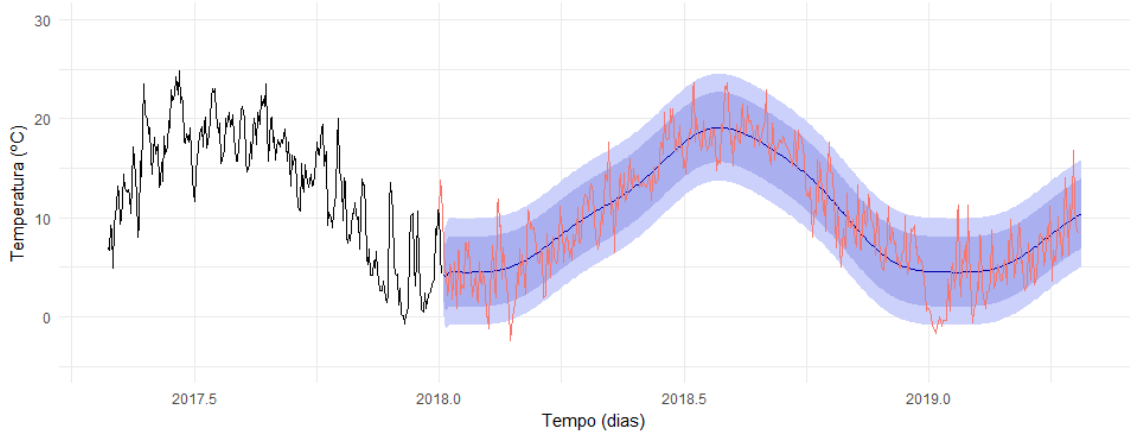


Figura 6.11: Valores observados e previsões (com limites de confiança de 80% e 95%) para a temperatura mínima resultantes do modelo de regressão com erros correlacionados.

Tabela 6.5: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos às temperaturas mínimas.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
4,227	(1,526; 6,927)	(0,097; 8,356)	10,40
4,088	(0,833; 7,344)	(-0,891; 9,067)	13,80
4,064	(0,635; 7,492)	(-1,180; 9,307)	12,90
4,270	(0,766; 7,774)	(-1,089; 9,629)	7,20
4,490	(0,963; 8,016)	(-0,903; 9,882)	4,50
4,529	(1,001; 8,058)	(-0,867; 9,926)	4,60
4,526	(0,997; 8,055)	(-0,871; 9,923)	2,20

Quanto à análise de resíduos, o histograma da Figura 6.7 sugere que os resíduos apresentam uma ligeira assimetria negativa, ou enviesamento à esquerda, suficiente para que o teste de Kolmogorov-Smirnov rejeite a hipótese da normalidade (valor de prova ≈ 0), para um nível de significância de 5%. Este comportamento pode, também, ser explicado dada a dimensão da amostra em estudo (n=2932 dias). Como a dimensão da amostra é elevada, conduz à rejeição da normalidade. Quanto à independência, o teste de Ljung-Box é aplicado à série de resíduos e conclui-se que a hipótese é rejeitada (valor de prova ≈ 0). Desta forma, do ponto de vista inferencial, os pressupostos do modelo não são verificados.

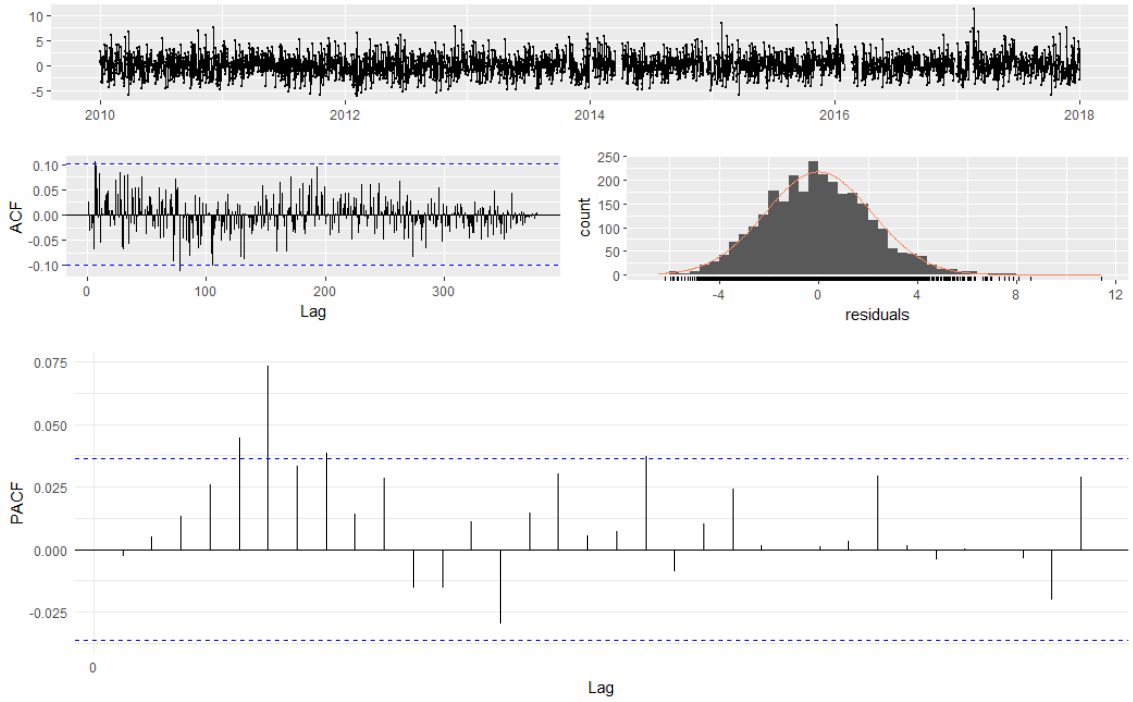


Figura 6.12: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo de regressão com erros correlacionados da temperatura mínima.

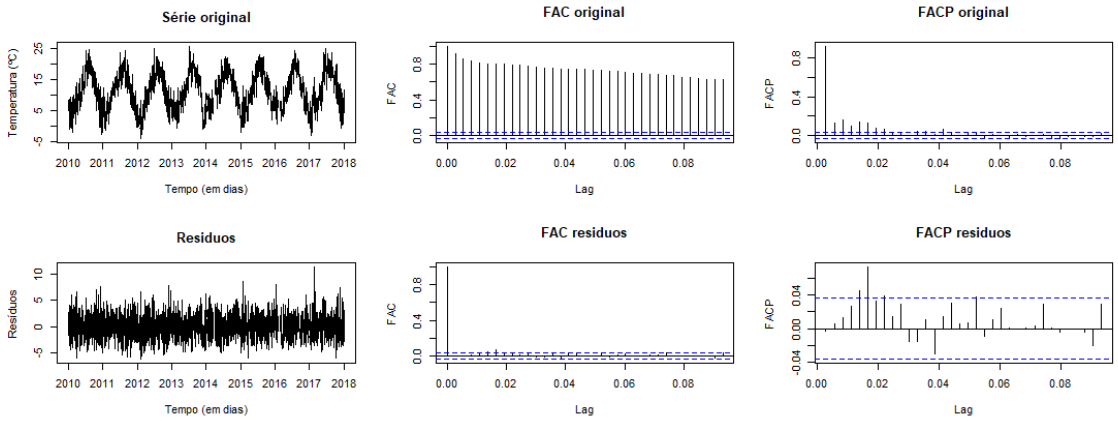


Figura 6.13: Série original e respetivas FAC e FACP, resíduos e respetivas FAC e FACP.

Avaliação da precisão dos modelos

O comportamento dos resíduos na modelação dos processos é consistente com o processo de ruído branco, apenas no caso do modelo TBATS (Figura 6.7). Do ponto de vista inferencial, o modelo de regressão não cumpre os pressupostos, como mostra a Figura 6.12.

A Tabela 6.6 mostra o resultado das medidas de precisão calculadas para o período de treino e o período de teste, dos dois métodos aplicados às séries temporais em estudo. O desempenho dos modelos concorrentes (TBATS e MR com erros correlacionados) foi avaliado utilizando o EM, REQM, EAM e EEAM. Os resultados obtidos mostraram que, no período de treino, o modelo TBATS apresenta melhor desempenho do que o modelo de regressão com erros correlacionados. Por sua vez, o modelo de regressão com erros correlacionados, que requer menos parâmetros a serem estimados (mais parcimonioso), tem um desempenho ligeiramente melhor ao do TBATS no período de teste.

Tabela 6.6: Medidas de avaliação do modelo TBATS e do modelo de regressão, para o período de teste e de treino.

Modelo	EM	REQM	EAM	EEAM	
TBATS	-0,00475	2,12730	1,66916	0,93086	Período de Treino
MR com MA(5)	-0,00533	2,14421	1,69319	0,92200	Período de Treino
TBATS	0,76651	2,74411	2,15127	1,19972	Período de Teste
MR com MA(5)	-0,05331	2,67814	2,10762	1.14767	Período de Teste

6.2 Temperatura máxima

Na Figura 6.14 está representada a série temporal da temperatura máxima já completa, com os valores em falta estimados (a vermelho).

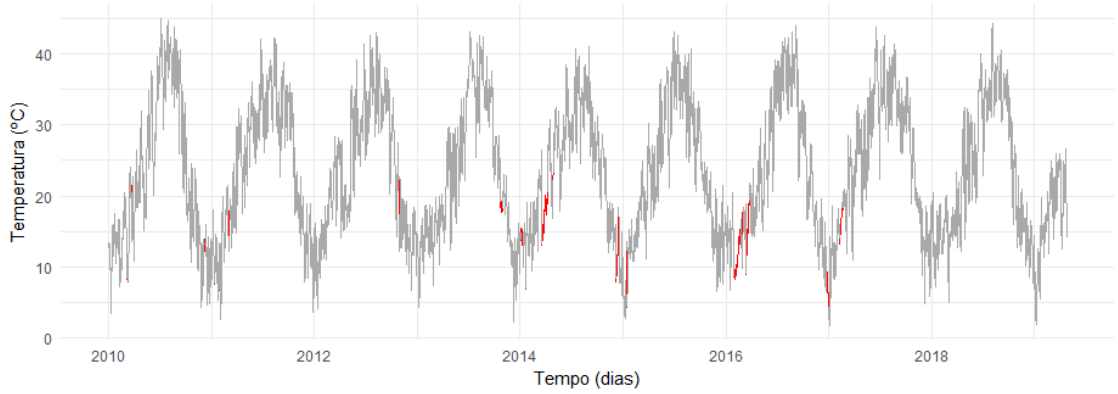


Figura 6.14: Série valores originais e valores imputados (a vermelho) da temperatura máxima.

Modelo TBATS

A Figura 6.15 representa a série de treino (de 1 de janeiro de 2010 até 31 de dezembro de 2018) e a série de teste (de 1 de janeiro de 2018 a 23 de abril de 2010), novamente, considera-se o período sazonal de 365 dias ($m_1 = 365$). Na Figura 6.16 estão representadas as respetivas FAC e FACP da série de treino.

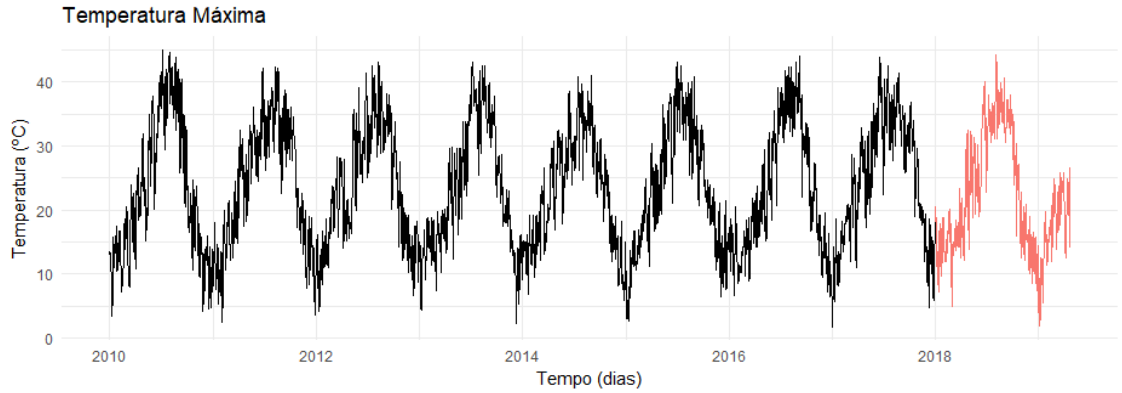


Figura 6.15: Serie de treino (a preto) e série de teste (a vermelho) da temperatura máxima.

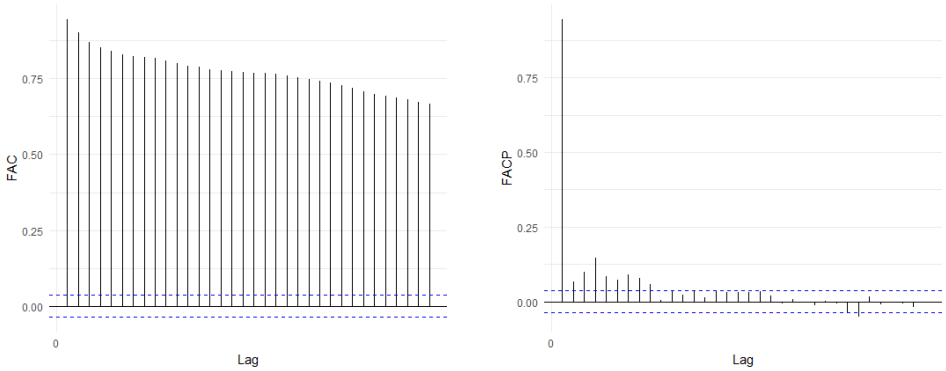


Figura 6.16: FAC e FACP da série de treino da temperatura máxima

Para a temperatura máxima, o modelo TBATS final foi $TBATS(1, 0, 4, 1, \{365, 6\})$. Não foi aplicada nenhuma tranformação Box-Cox ($\omega = 1$). Os valores de $\beta = 0$ e de $\phi = 1$ indicam a ausência de um efeito de amortecimento, a componente irregular da série é correlacionada e modelada por um processo ARMA(0,4), ou seja um processo de média móveis MA(4). As estimativas dos parâmetros obtidas para o modelo TBATS da temperatura máxima são apresentadas na Tabela 6.7.

Tabela 6.7: Parâmetros do modelo TBATS selecionado.

Parâmetros	α	γ_1	γ_2	θ_1	θ_2	θ_3	θ_4
Estimativas	1,31888	-0,00015	-0,00009	-0,55716	-0,06879	-0,22371	-0,08824

A estimativa do desvio padrão do ruído branco associado ao modelo é $\hat{\sigma} = 2,93$ e o valor de $AIC = 29664,02$.

Assim, as equações do modelo são as seguintes:

$$y_t^{(1)} = l_{t-1} + b_{t-1} + s_{t-365}^{(1)} + d_t,$$

$$l_t = l_{t-1} + b_{t-1} + 1,31888 d_t,$$

$$b_t = b_{t-1},$$

$$s_{j,t}^{(1)} = s_{j,t-1}^{(1)} \cos \lambda j^{(1)} + s_{j,t}^{*(1)} \sin \lambda_j^{(1)} + (-0,00015) d_t,$$

$$s_{j,t}^{*(1)} = -s_{j,t-1} \sin \lambda_j^{(1)} + s_{j,t-1}^{*(1)} \cos \lambda_j^{(1)} + (-0,00009) d_t,$$

onde $\lambda_j^{(1)} = \frac{2\pi j}{365}$, d_t um processo ARMA(0,4) e α, β, γ_1 e γ_2 os parâmetros de alisamento. A sazonalidade foi modelada por 19 parâmetros (17 valores iniciais para

$s_{j,0}^{(1)}$ e $s_{j,0}^{*(1)}$ e dois parâmetros de alisamento $\gamma_1^{(1)}$ e $\gamma_2^{(1)}$).

A equação do processo de médias móveis MA(4) é dada por

$$d_t = (-0,55716)\,\varepsilon_{t-1} + (-0,06879)\,\varepsilon_{t-2} + (-0,22371)\,\varepsilon_{t-3} + (-0,08824)\,\varepsilon_{t-4} + \varepsilon_t,$$

onde $\varepsilon_t \sim N(0,(2,93)^2)$.

A representação dos valores observados da série temporal e os valores estimados pelo modelo TBATS para a temperatura máxima estão representados na Figura 6.17. A representação gráfica sugere um ajustamento bastante satisfatório do modelo TBATS aos dados da série temporal da temperatura máxima.

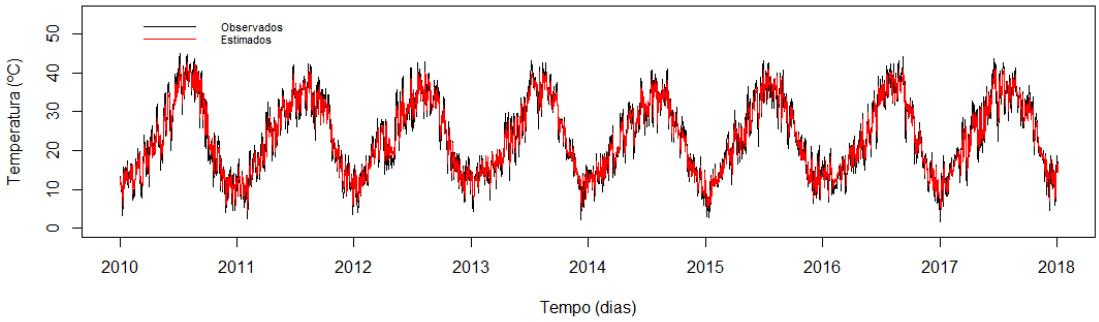


Figura 6.17: Valores observados e valores estimados pelo modelo TBATS para a temperatura máxima.

Os valores observados da temperatura máxima, as estimativas no período de modelação (período de treino) e os intervalos de previsão para um nível de confiança de 80% e 95%, resultantes da aplicação do modelo TBATS estão representados na Figura 6.18. No primeiro trimestre do ano de 2018 verifica-se que algumas previsões ultrapassam os limites dos intervalos de previsão, tal é visível na Figura 6.19.

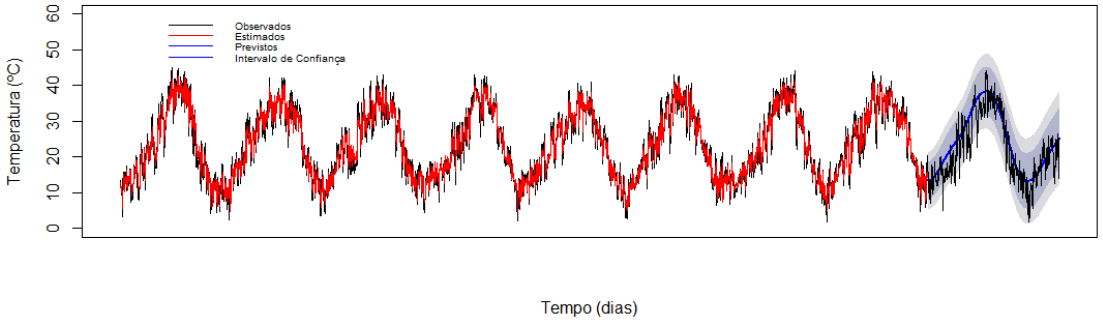


Figura 6.18: Valores observados, estimados e previstos pelo modelo TBATS para a temperatura máxima.

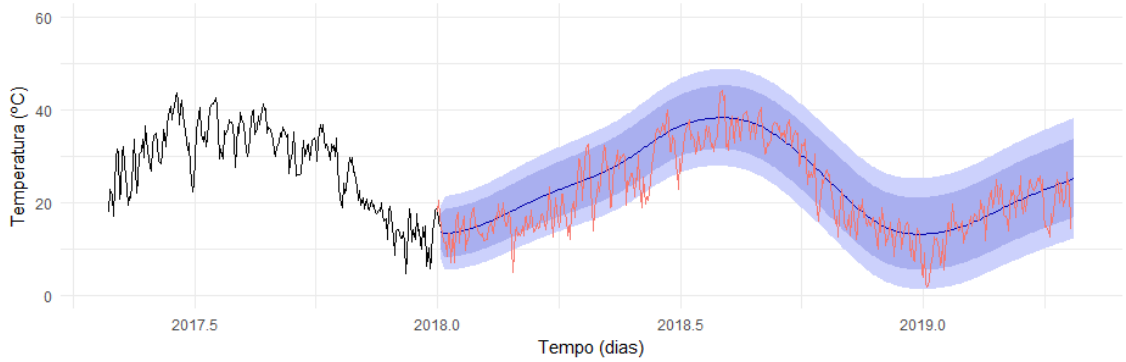


Figura 6.19: Valores observados, estimados e previstos pelo modelo TBATS para a temperatura máxima.

Na Tabela 6.8 mostra-se os valores previstos, intervalos de confiança resultantes do processo de previsão e, ainda, os valores observados para os primeiros 7 dias do período de teste. Os primeiros três valores observados não pertencem ao intervalo de previsões.

Tabela 6.8: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos às temperaturas máximas.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
4,227	(1,526; 6,927)	(0,097; 8,356)	10,40
4,088	(0,833; 7,344)	(-0,891; 9,067)	13,80
4,064	(0,635; 7,492)	(-1,180; 9,307)	12,90
4,270	(0,766; 7,774)	(-1,089; 9,629)	7,20
4,490	(0,963; 8,016)	(-0,903; 9,882)	4,50
4,529	(1,001; 8,058)	(-0,867; 9,926)	4,60
4,526	(0,997; 8,055)	(-0,871; 9,923)	2,20

Pela análise de resíduos do modelo, verifica-se que o pressuposto da normalidade é rejeitado (valor de prova = 0,00379 do teste de Kolmogorv-Smirnov). O pressuposto da independência foi avaliado estimando a função de autocorrelação e a função de autocorrelação parcial dos resíduos. Pelo teste Ljung-Box, aplicado à série dos resíduos, existe evidência estatisticamente significativa, a um nível de significância de 5%, para rejeitar a independência dos erros (valor de prova = 0,01313). O histograma dos resíduos sugere, ainda, que os resíduos apresentam uma assimetria negativa (ou enviesamento à direita), com uma cauda pesada à esquerda.

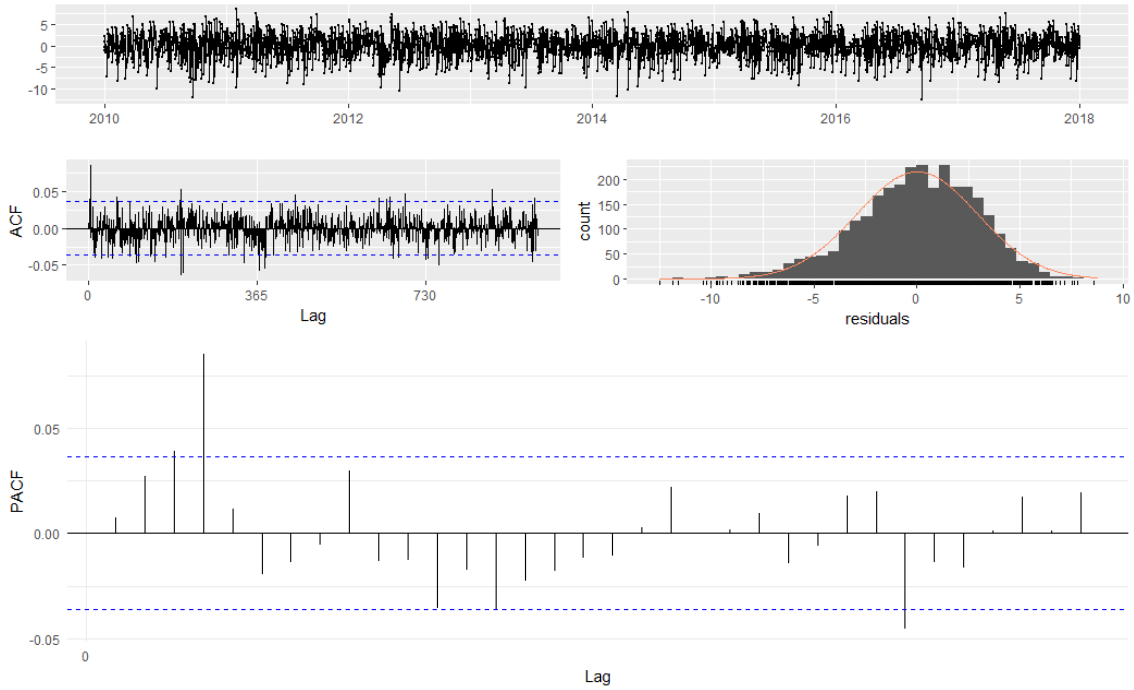


Figura 6.20: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo TBATS da temperatura máxima.

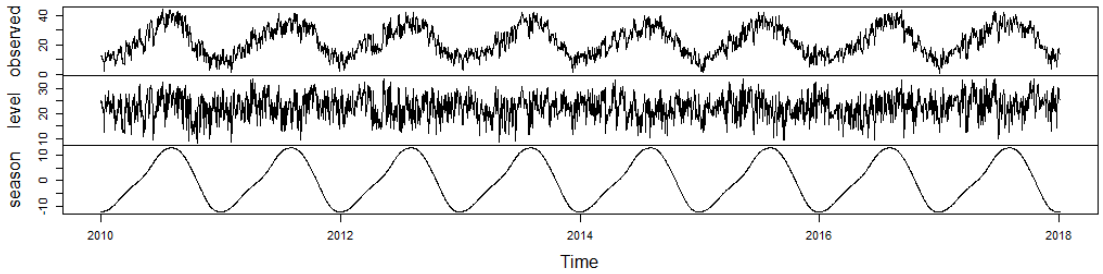


Figura 6.21: Decomposição série temporal temperatura máxima, obtida pela estimação do modelo TBATS.

A decomposição da série da temperatura máxima, obtida dos valores ajustados do modelo TBATS, é apresentada na Figura 6.21. A presença de uma componente de tendência é negligenciada e um padrão sazonal cíclico é evidente.

Modelo de Regressão com erros Correlacionados

Para a temperatura máxima, o número K de pares de senos e cosenos de Fourier, para o qual o valor de AIC do modelo ARIMA se torna mínimo é de $K = 2$ (Tabela 6.9).

Tabela 6.9: Ordem do processo ARIMA, valor de K, valor AIC corrigido e BIC.

Modelo	K	AICc	BIC
Regressão com erros ARIMA(1,0,1)	1	14161,06	14196,91
Regressão com erros ARIMA(1,0,1)	2	**14100, 96**	**14148, 75**
Regressão com erros ARIMA(1,0,1)	3	14103,36	14163,40
Regressão com erros ARIMA(0,0,4)	4	17869,90	17947,52
Regressão com erros ARIMA(0,0,4)	5	17873,93	17963,47

O valor de AICc para o qual ε_t é um processo ARIMA(1,0,1) é de AICc = 14103,36. Na Tabela 6.10 são apresentadas as estimativas dos parâmetros e os respetivos erros padrão, para o modelo de regressão com erros ARIMA(1,0,1). No modelo final, a estimativa do desvio padrão do ruído branco associado ao modelo é de $\hat{\sigma} = 9,26$ e o valor de $AIC = 14802,64$.

Tabela 6.10: Estimativas dos parâmetros resultantes do modelo de regressão com erros correlacionados e os erros padrão correspondentes.

Parâmetros	ϕ_1	θ_1	β_0	α_1	α_2	β_1	β_2
Estimativas	0,6734	0,0747	23,6349	-3,9918	2,0099	-11,2822	-0,9514
Erros padrão	0,0192	0,0256	0,1822	0,2583	0,2568	0,2565	0,2565

Desta forma, o modelo obtido é dado pela seguinte equação

$$y_t = 23,6349 + (-3,9918)\sin\frac{2\pi t}{365} + (-11,2822)\cos\frac{2\pi t}{365} + 2,0099\sin\frac{4\pi t}{365} + (-0,9514)\cos\frac{4\pi t}{365} + d_t,$$

onde d_t é um processo $ARMA(1,0,1)$ descrito pela seguinte equação

$$d_t = 0,6734 d_{t-1} + 0,0747\varepsilon_{t-1},$$

onde $\varepsilon_t \sim N(0,(9,26)^2)$.

Na Figura 6.22 é possível ver a representação dos valores observados da série temporal (a preto) e dos valores estimados pelo modelo (a vermelho). Percebe-se que o modelo se ajusta de forma satisfatória aos dados, conseguindo acompanhar o comportamento fortemente sazonal da série da temperatura máxima.

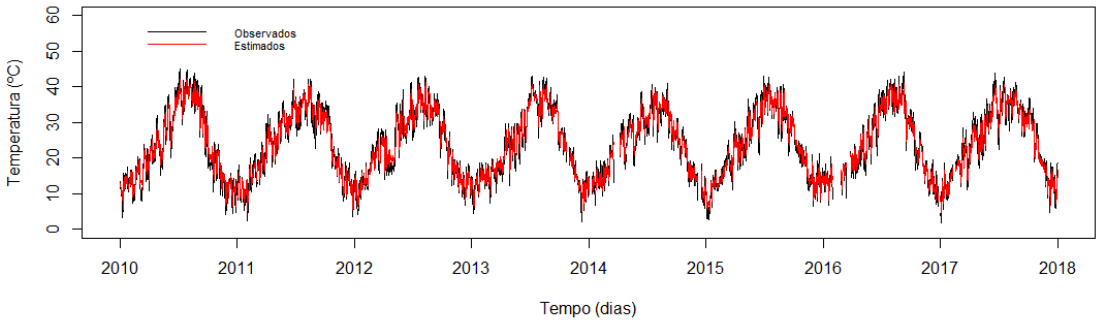


Figura 6.22: Valores observados e valores estimados pelo modelo de regressão com erros correlacionados.

Na Figura 6.23 apresentam-se os valores observados da temperatura máxima, as estimativas no período de modelação (período de treino), as previsões no período de previsão (período de teste) e os intervalos de previsão para um nível de confiança de 80% e 95%, resultantes da aplicação do modelo de regressão com erros correlacionados. Na Figura 6.24 é possível verificar que alguns valores estão fora dos limites do intervalo de previsão.

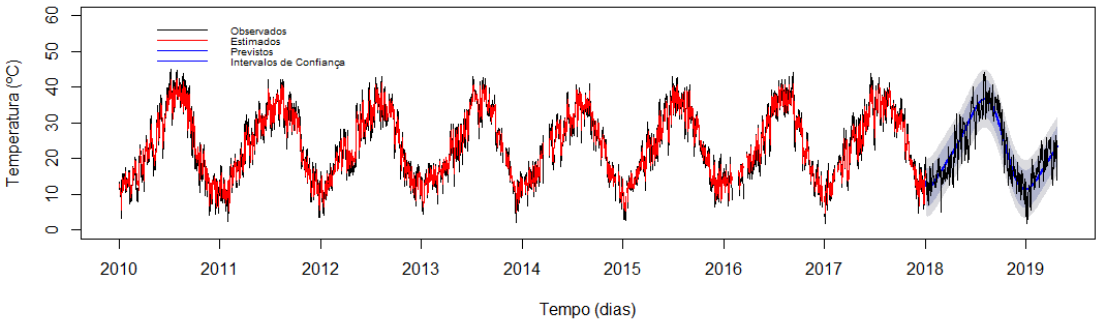


Figura 6.23: Valores observados e previsões (com intervalos de confiança de 80% e 95%) para a temperatura máxima resultante do modelo de regressão com erros correlacionados.

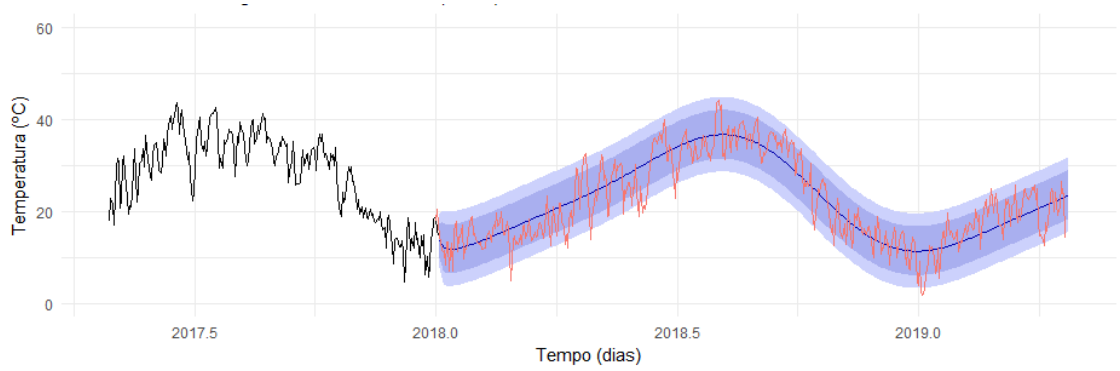


Figura 6.24: Valores observados e previsões (com limites de confiança de 80% e 95%) para a temperatura máxima resultantes do modelo de regressão com erros correlacionados.

Na Tabela 6.11 são apresentados os valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão para a temperatura máxima, resultante do modelo de regressão com erros correlacionados.

Tabela 6.11: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos às temperaturas máximas.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
13,662	(9,937; 17,387)	(7,965; 19,359)	18,00
12,956	(8,303; 17,608)	(5,840; 20,071)	20,40
12,491	(7,474; 17,508)	(4,818; 20,163)	17,40
12,190	(7,016; 17,363)	(4,277; 20,102)	14,60
11,999	(6,757; 17,243)	(3,981; 20,019)	11,80
11,887	(6,612; 17,161)	(3,820; 19,953)	11,20
11,826	(6,537; 17,114)	(3,738; 19,914)	11,60

Após a escolha do modelo final, é importante verificar se estes cumprem os pressupostos associados aos modelos de regressão. Na Figura 6.25 está representada a análise de resíduos. Pelo teste Ljung-Box, aplicado à série dos resíduos, existe evidência estatisticamente significativa, a um nível de significância de 5%, para se rejeitar a independência dos erros (valor de prova = 0,03739). Para o pressuposto da normalidade dos resíduos foi aplicado o teste Kolmogorov-Smirnov e como resultado conclui-se que os resíduos não seguem uma distribuição Normal (valor de prova ≈ 0). Assim, do ponto de vista inferencial não se verificam os pressupostos do modelo.

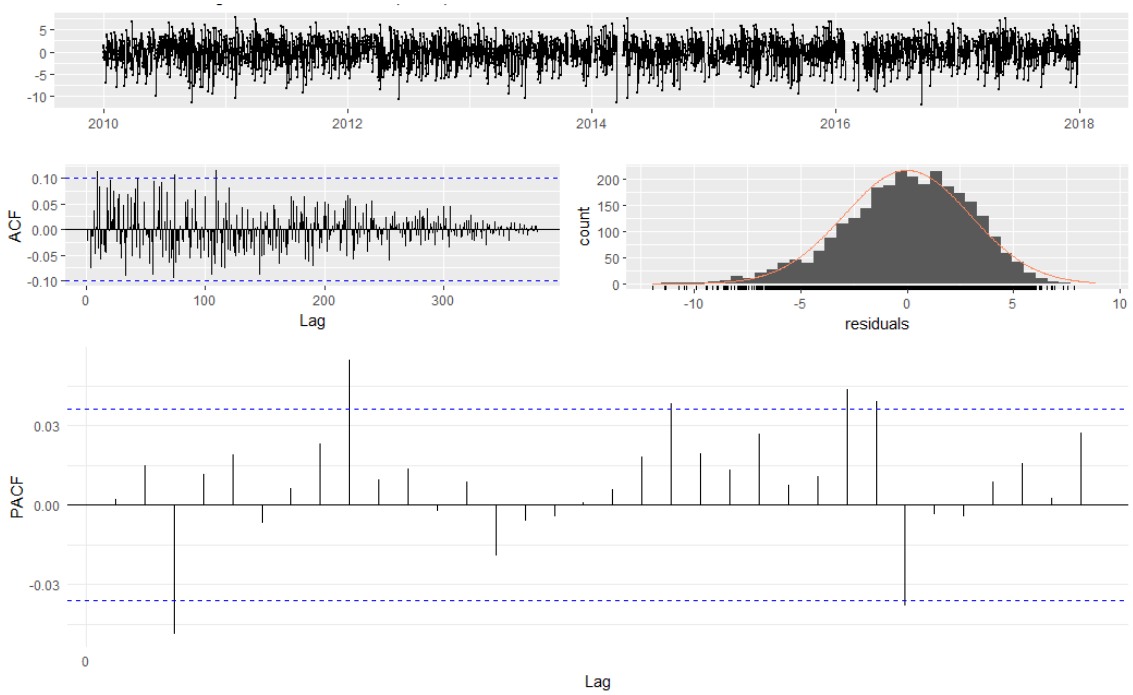


Figura 6.25: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo de regressão com erros correlacionados da temperatura máxima.

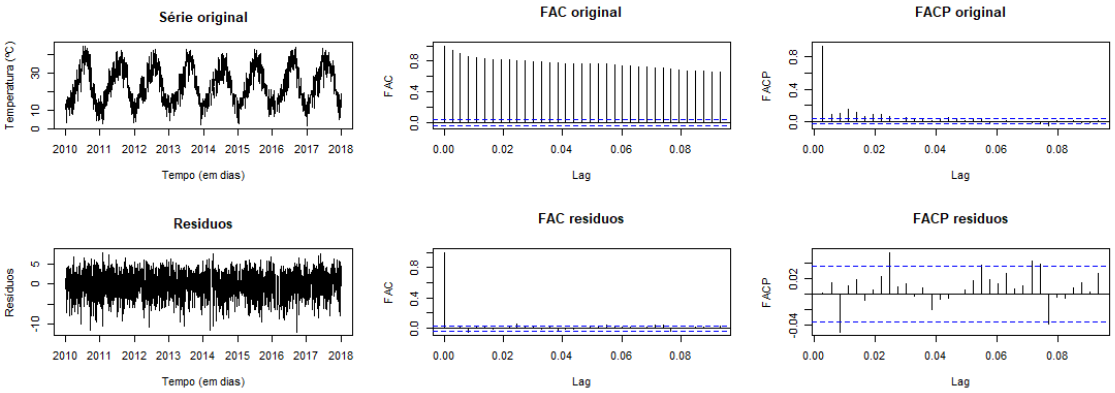


Figura 6.26: Série original e respetivas FAC e FACP, resíduos e respetivas FAC e FACP.

Avaliação da precisão dos modelos

A Tabela 6.12 mostra o resultado das medidas de avaliação calculadas para os períodos de treino e teste dos dois métodos aplicados às séries temporais em estudo. Para o período de teste o MR com ARIMA (1,0,1) apresenta melhor desempenho, por sua vez no período de treino os dois modelos comportam-se de forma semelhante. O modelo TBATS apresenta melhor valor de REQM e EAM e o modelo RM com ARIMA (1,0,1), o melhor valor de EM e EEAM.

Tabela 6.12: Medidas de avaliação do modelo TBATS e do modelo de regressão, para o período de teste e de treino.

Modelo	EM	REQM	EAM	EEAM	
TBATS	0,00793	2,93148	2,32264	0,94896	Período de Treino
MR com ARIMA(1,0,1)	-0,00045	2,96170	2,35666	0,93767	Período de Treino
TBATS	-1,96582	4,54560	3,54778	1,44952	Período de Teste
MR com ARIMA (1,0,1)	-0,12670	4,04458	3,28463	1,30689	Período de Teste

6.3 Precipitação

Na Figura 6.27 está representada a série temporal da precipitação, com os valores calculados dos dados omissos representados a vermelho.

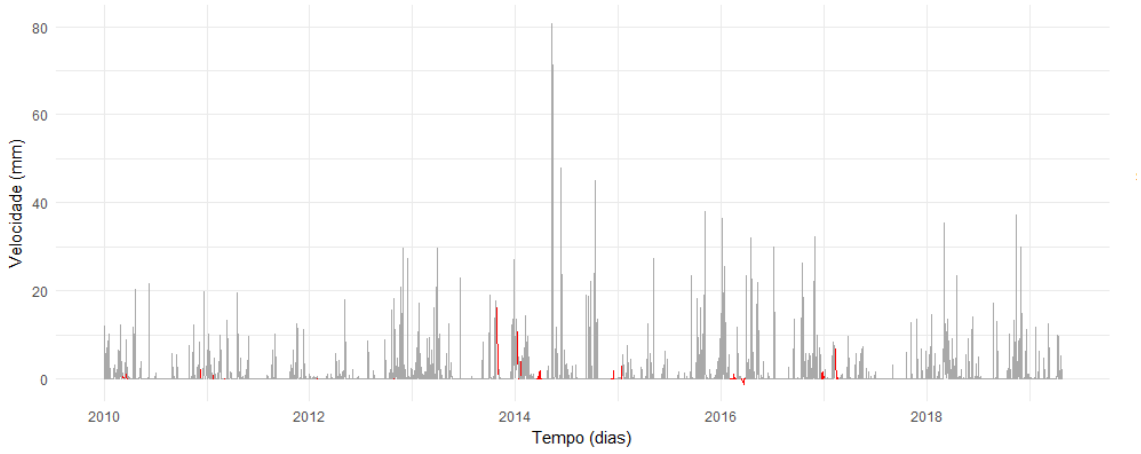


Figura 6.27: Série valores originais e valores imputados (a vermelho) da precipitação.

Modelo TBATS

Os dados da precipitação são observados de forma diária e apresentam forte período sazonal ($m_1 = 365$). Na Figura 6.28 está representada a série de treino (de 1 de janeiro de 2010 até 31 de dezembro de 2018) e a série de teste (de 1 de janeiro de 2018 a 23 de abril de 2010). Na Figura 6.29 estão representadas as respetivas FAC e FACP da série de treino original.

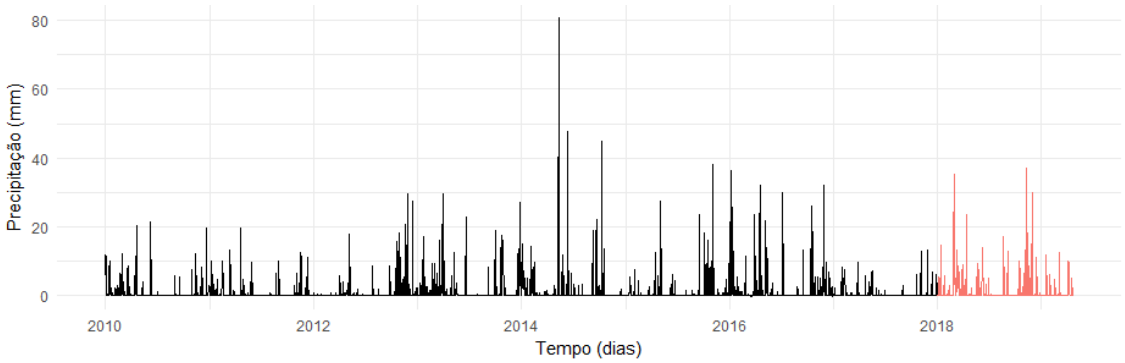


Figura 6.28: Série de treino (a preto) e série de teste (a vermelho) da precipitação.

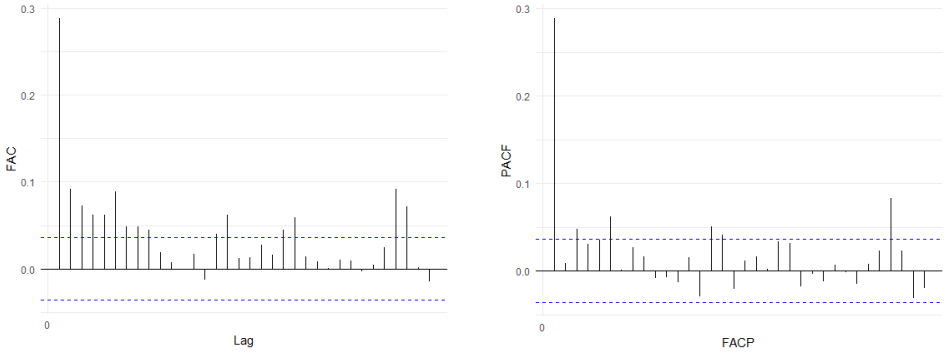


Figura 6.29: FAC e FACP da série de treino da precipitação.

Para a precipitação, o modelo TBATS obtido foi $TBATS(1, 2, 0, 1, \{365, 6\})$. O parâmetro ω do modelo indica que não foi necessário aplicar nenhuma transformação Box-Cox à série temporal. Os valores estimados para $\beta = 0$ e $\phi = 1$ indicam que o modelo não sofreu efeito de amortecimento e a componente irregular da série é correlacionada e modelada por um processo ARMA(2,0), ou seja, um autorregressivo de ordem 2. As estimativas dos parâmetros obtidas do modelo TBATS para a precipitação são apresentadas na Tabela 6.13.

Tabela 6.13: Parâmetros do modelo TBATS selecionado.

Parâmetros	α	γ_1	γ_2	φ_1	φ_2
Estimativas	0,00595	-0,00068	0,00050	0,27043	-0,00951

A estimativa do desvio padrão do ruído branco associado ao modelo é $\hat{\sigma}$=3,798 e o valor de AIC= $31167,78$.

Assim, as equações do modelo são as seguintes:

$$y_t^{(1)} = l_{t-1} + b_{t-1} + s_{t-365}^{(1)} + d_t,$$

$$l_t = l_{t-1} + b_{t-1} + 0,00595 d_t,$$

$$b_t = b_{t-1},$$

$$s_{j,t}^{(1)} = s_{j,t-1}^{(1)} \cos \lambda j^{(1)} + s_{j,t}^{*(1)} \sin \lambda_j^{(1)} + -0,00068 d_t,$$

$$s_{j,t}^{*(1)} = -s_{j,t-1} \sin \lambda_j^{(1)} + s_{j,t-1}^{*(1)} \cos \lambda_j^{(1)} + 0,00050 d_t,$$

onde $\lambda_j^{(1)} = \frac{2\pi j}{365}$, d_t um processo ARMA(2,0), ou seja, um AR(2) e α, β, γ_1 e γ_2 os parâmetros de alisamento. No processo de modelação da sazonalidade foram utilizados 17 parâmetros (15 valores iniciais para $s_{j,0}^{(1)}$ e $s_{j,0}^{*(1)}$ e dois parâmetros de alisamento $\gamma_1^{(1)}$ e $\gamma_2^{(1)}$).

A equação do processo autorregressivo AR(2) é dada por

$$d_t = 0,27043 d_{t-1} + (-0,00951) d_{t-2} + \varepsilon_t,$$

onde $\varepsilon_t \sim N(0, (3,80)^2)$.

Na Figura 6.30 estão representados os valores observados da série temporal (a preto) e os valores estimados pelo modelo (a vermelho). A observaçao gráfica mostra, claramente, a dificuldade de uma tentativa de estabelecimento de um processo de modelação desta variável (precipitação). Apenas modelos estatísticos não conseguem traduzir de forma eficiente o comportamento, muito particular, deste tipo de variável.

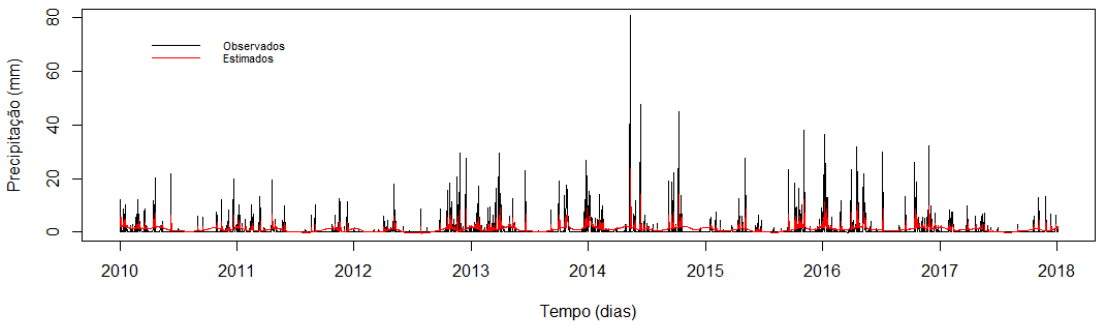


Figura 6.30: Valores observados e valores estimados pelo modelo TBATS para a precipitação.

Na Figura 6.31 são apresentados os valores observados da precipitação, as estimativas no período de modelação (período de treino) e os intervalos de previsão para

um nível de confiança de 80% e 95%, resultantes da aplicação do modelo TBATS. A volatilidade existente nesta série temporal acaba por influenciar os resultados obtidos, dando origem a previsões pouco precisas no domínio temporal em que esta se observa.

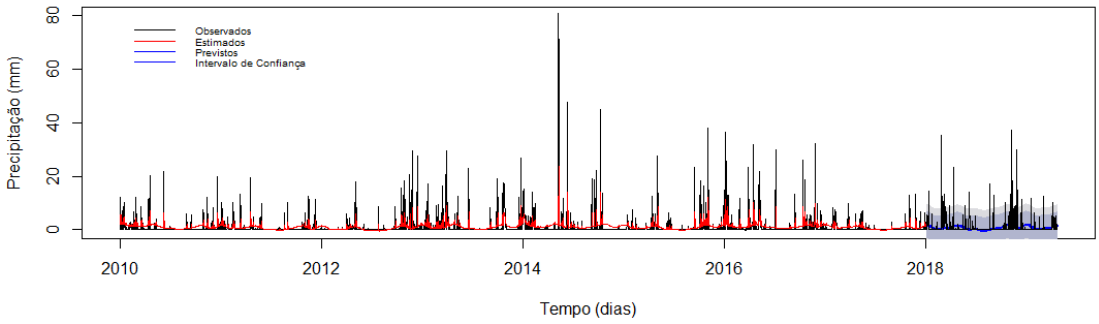


Figura 6.31: Valores observados, estimados e previstos pelo modelo TBATS para a precipitação.

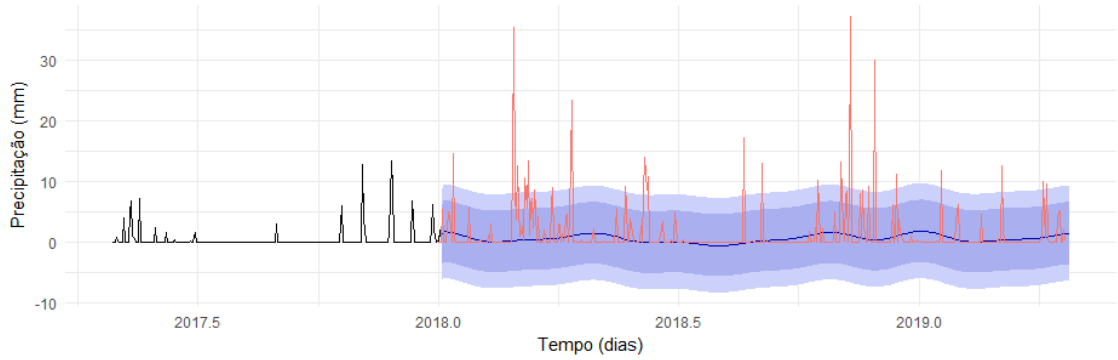


Figura 6.32: Valores observados, estimados e previstos pelo modelo TBATS para a precipitação.

Na Tabela 6.14 encontram-se as primeiras 7 previsões para o período de teste e os respetivos intervalos de previsão (a 80% e 95%) resultantes do modelo. Os valores observados da precipitação encontram-se contidos nos intervalos de previsão.

Tabela 6.14: Valores previstos e respetivos intervalos de confiança a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos à precipitação.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
1,305	(-3,562; 6,172)	(-6,139; 8,749)	0,00
1,667	(-3,377; 6,711)	(-6,048; 9,382)	0,00
1,738	(-3,316; 6,793)	(-5,992; 9,469)	0,00
1,736	(-3,320; 6,791)	(-5,996; 9,467)	5,40
1,712	(-3,344; 6,768)	(-6,020; 9,444)	0,00
1,679	(-3,376; 6,735)	(-6,053; 9,411)	0,20
1,641	(-3,415; 6,696)	(-6,091; 9,372)	0,00

A validação do modelo foi avaliada pela análise dos resíduos, representada na Figura 6.33. O pressuposto da independência foi avaliado estimando a função de autocorrelação e a função de autocorrelação parcial dos resíduos. Pelo teste Ljung-Box, aplicado à série dos resíduos, conclui-se que existe evidência estatística, a um nível de significância de 5%, para rejeitar a independência dos erros (valor de prova $\approx$ 0). Para o pressuposto da normalidade dos resíduos foi aplicado o teste *Kolmogorov-Smirnov*, a um nível de significância de 5%, verificando-se que os resíduos não seguem uma distribuição Normal (valor de prova $\approx$ 0).

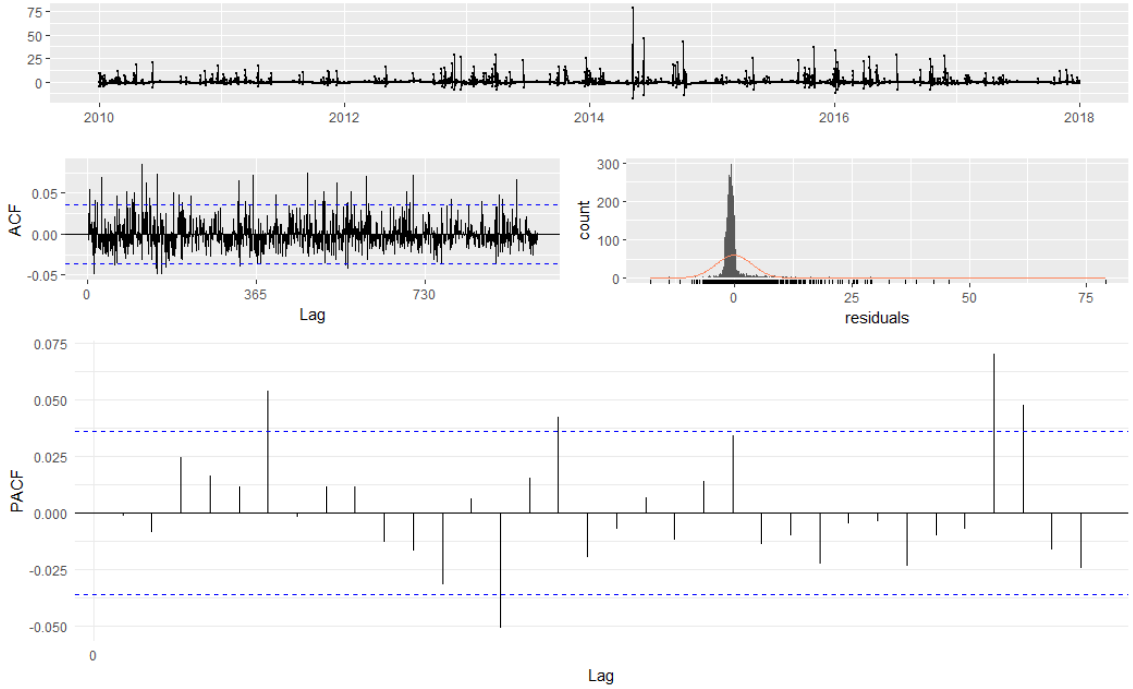


Figura 6.33: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo TBATS da precipitação.

A decomposição da série da precipitação, obtida dos valores ajustados do modelo TBATS, é apresentada na Figura 6.34.

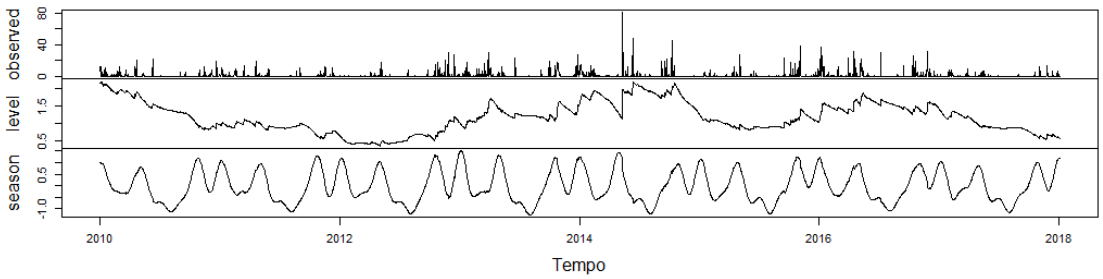


Figura 6.34: Decomposição série temporal da precipitação, obtida pela estimação do modelo TBATS.

Modelo de Regressão com Erros Correlacionados

Na Tabela 6.15 estão apresentadas as ordens dos processos ARIMA estabelecidos, os valores para K, bem como, o valor AICc e o valor BIC para a seleção do modelo para a precipitação. O valor de K para o qual o valor de AICc do modelo se torna mínimo é o de $K = 2$.

Tabela 6.15: Ordem do processo ARIMA, valor de K, valor AIC corrigido e BIC

Modelo	K	AICc	BIC
Regressão com erros ARIMA(2,0,3)	1	15567,22	15620,98
Regressão com erros ARIMA(2,0,2)	2	**15561, 89**	15621,62
Regressão com erros ARIMA(2,0,3)	3	15564,70	15642,32
Regressão com erros ARIMA(2,0,2)	4	15563,79	15647,37
Regressão com erros ARIMA(2,0,2)	5	15567,19	15662,69

O valor de AICc, para o qual ε_t é um processo ARIMA(2,0,2), é de AICc = 15561,89. Na Tabela 6.16 são apresentadas as estimativas dos parâmetros e os respetivos erros padrão, para o modelo de regressão com erros ARIMA(2,0,2).

Tabela 6.16: Estimativas dos parâmetros resultantes do modelo de regressão com erros correlacionados e os erros padrão correspondentes.

Parâmetros	φ_1	φ_2	θ_1	θ_2	β_0	α	β
Estimativas	0,9598	-0,0346	-0,7042	-0,1667	1,0685	0,1982	0,4805
Erro padrão	0,1613	0,1069	0,1608	0,0838	0,1244	0,1746	0,1734

No modelo final, a estimativa do desvio padrão do ruído branco associado ao modelo é de $\hat{\sigma} = 14,29$ e o valor de $AIC = 15565,61$.

Desta forma, as equações para o modelo obtido são dadas por

$$y_t = 1,0685 + (-0,1982)\sin\frac{2\pi t}{365} + (0,4805)\cos\frac{2\pi t}{365} + d_t,$$

onde d_t é um processo $ARMA(2,0,2)$ descrito pela seguinte equação

$$d_t = 0,9598d_{t-1} + (-0,0346)d_{t-2} + (-0,7042)\varepsilon_{t-1} + (-0,1667)\varepsilon_{t-2},$$

onde $\varepsilon_t \sim N(0,(14,29)^2)$.

A Figura 6.35 mostra os valores observados da série temporal (a preto) e os valores estimados pelo modelo de regressão com erros correlacionados para a precipitação (a vermelho), no período observado.

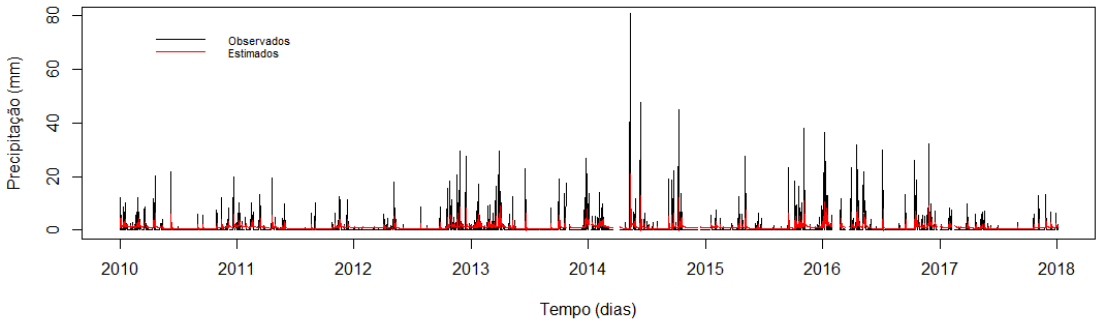


Figura 6.35: Valores observados e valores estimados pelo modelo de regressão com erros correlacionados.

Na Figura 6.36 estão representados os valores observados da precipitação, as estimativas no período de modelação (período de treino), as previsões no período de previsão (período de teste) e os intervalos de previsão para um nível de confiança de 80% e 95%, resultantes da aplicação do modelo de regressão com erros correlacionados à precipitação. Ainda, por inspeção gráfica da Figura 6.37, percebe-se que claramente o modelo não consegue captar de forma eficiente o comportamento volátil da precipitação.

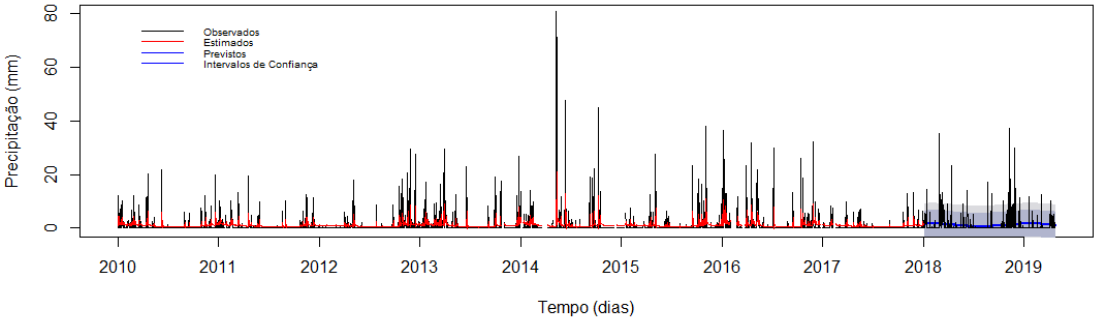


Figura 6.36: Valores observados e previsões (com intervalos de confiança de 80% e 95%) para a precipitação resultante do modelo de regressão com erros correlacionados.

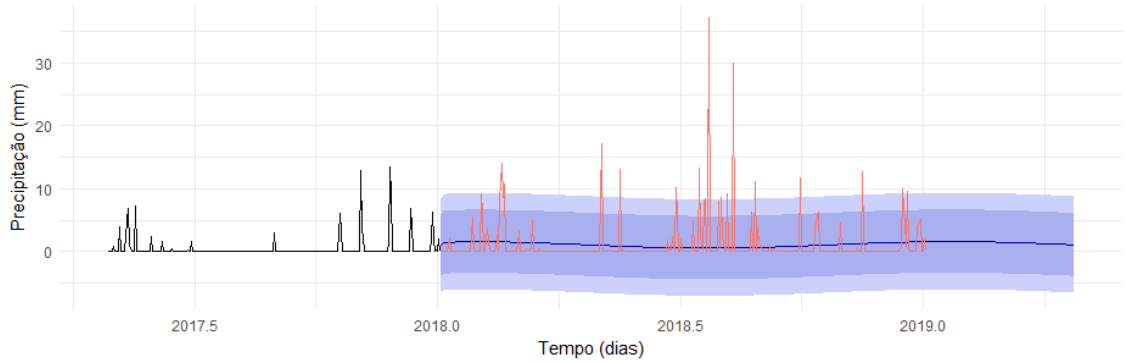


Figura 6.37: Valores observados e previsões (com limites de confiança de 80% e 95%) para a precipitação resultantes do modelo de regressão com erros correlacionados, em particular últimas 250 observações.

Na Tabela 6.17 são apresentadas as primeiras sete previsões resultantes do modelo de regressão com erros correlacionados obtido para a precipitação. Tal como no modelo TBATS, os valores observados da precipitação encontram-se contidos nos intervalos de previsão.

Tabela 6.17: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos à precipitação.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
1,030	(-3,815; 5,875)	(-6,379; 8,440)	0,00
1,376	(-3,625; 6,377)	(-6,272; 9,024)	0,00
1,405	(-3,601; 6,410)	(-6,251; 9,060)	0,00
1,420	(-3,588; 6,428)	(-6,239; 9,079)	5,40
1,434	(-3,576; 6,444)	(-6,228; 9,096)	0,00
1,447	(-3,565; 6,459)	(-6,218; 9,112)	0,20
1,459	(-3,554; 6,472)	(-6,209; 9,127)	0,00

A análise de resíduos mostrou que pelo teste Ljung-Box existe evidência estatística, a um nível de significância de 5%, para se rejeitar a independência dos erros (valor de prova ≈ 0). Pela aplicação do teste Kolmogorov-Smirnov, conclui-se que os resíduos não seguem uma distribuição Normal (valor de prova ≈ 0). Assim, percebe-se que do ponto de vista inferencial os pressupostos do modelo não são verificados.

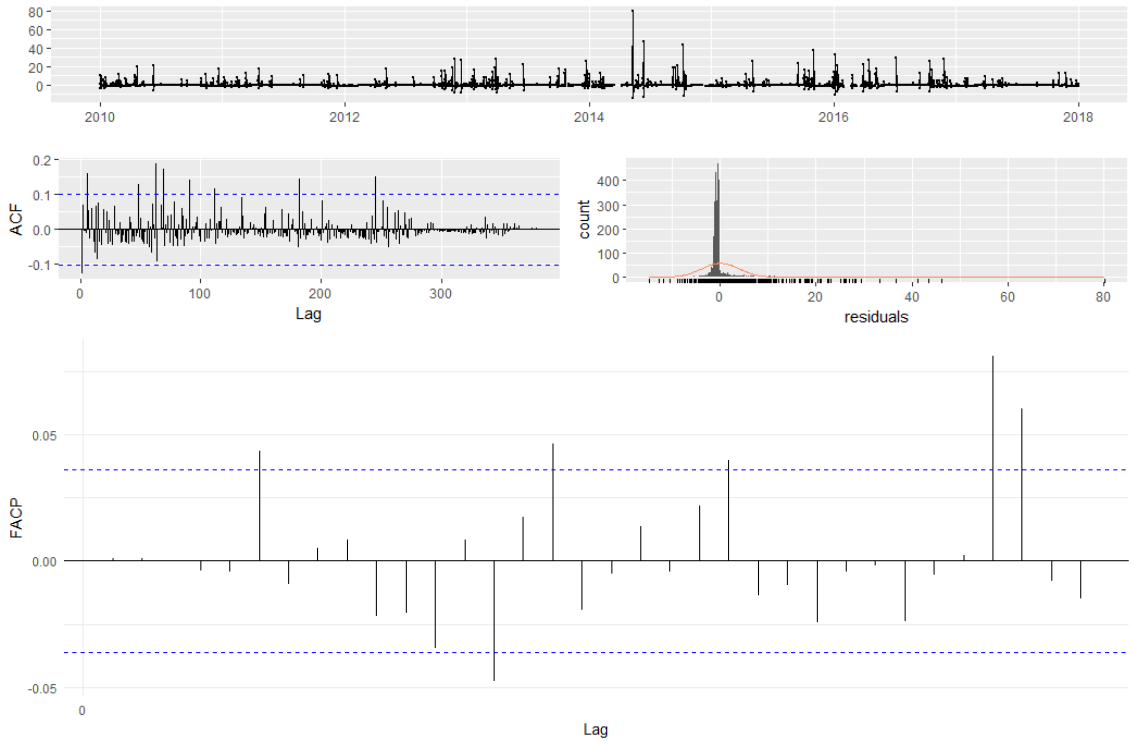


Figura 6.38: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo de regressão com erros correlacionados da precipitação.

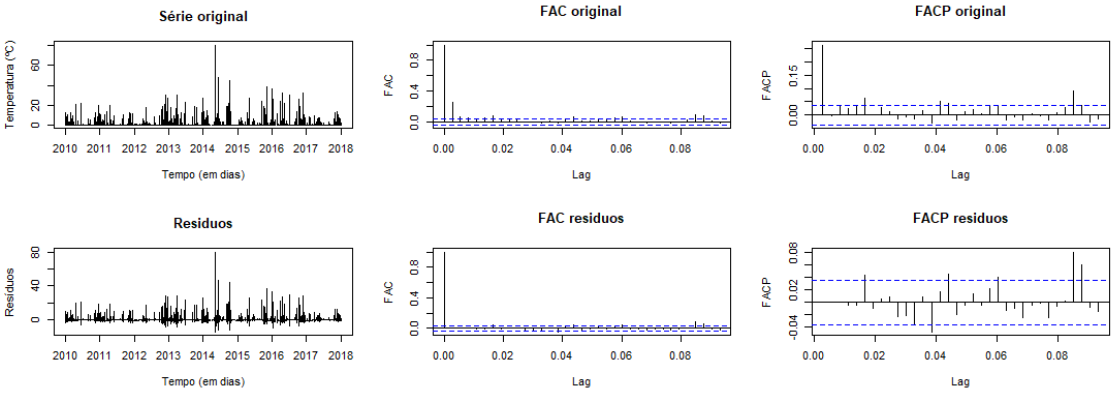


Figura 6.39: Série original e respetivas FAC e FACP, resíduos e respetivas FAC e FACP.

Avaliação da precisão dos modelos

Para se proceder à comparação das metodologias, quanto à qualidade das previsões foram calculadas as medidas de avaliação. A Tabela 6.18 mostra os valores das medidas de avaliação calculadas para o período de treino e período de teste dos dois métodos aplicados às séries temporais em estudo. No período de treino, em termos de erro quadrático médio (REQM) e erro médio (EM), o modelo TBATS

apresenta melhor desempenho. Por sua vez para o mesmo período de treino, o MR com ARIMA(2,0,2) apresenta melhor desempenho nas medidas de erro absoluto médio (EAM) e de erro escalado absoluto médio (EEAM).

Tabela 6.18: Medidas de avaliação do modelo TBATS e do modelo de regressão, para o período de teste e de treino.

Modelo	EM	REQM	EAM	EEAM	
TBATS	-0,06482	3,79789	1,61094	1,07389	Período de Treino
MR com ARIMA(2,0,2)	0,00110	3,85192	1,56417	1,01822	Período de Treino
TBATS	0,76699	4,06740	1,72774	1,15175	Período de Teste
MR com ARIMA (2,0,2)	0,22338	3,96320	1,98855	1,29448	Período de Teste

6.4 Velocidade Média do Vento

Na Figura 6.40 está representada a série temporal da velocidade média do vento completa (com os valores calculados a vermelho). De todas as variáveis em estudo, a velocidade média do vento é a que apresenta um maior número de valores em falta, como já foi dito. É também a série temporal com um comportamento mais irregular e, por isso, à partida dificilmente será obtido um bom ajustamento e previsão.

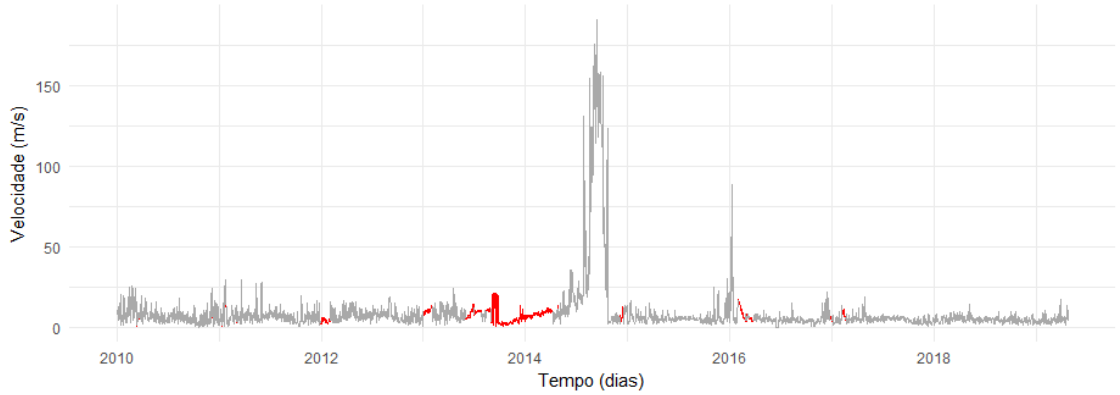


Figura 6.40: Série valores originais e valores imputados (a vermelho) da velocidade média do vento.

Modelo TBATS

Os dados da velocidade média do vento são observados de forma diária e apresentam um forte período sazonal ($m_1 = 365$). Na Figura 6.41 está representada a série de treino (de 1 de janeiro de 2010 até 31 de dezembro de 2018) e a série de teste (de 1 de janeiro de 2018 a 23 de abril de 2010). Na Figura 6.42 estão representadas as respetivas FAC e FACP da série de treino já completa.

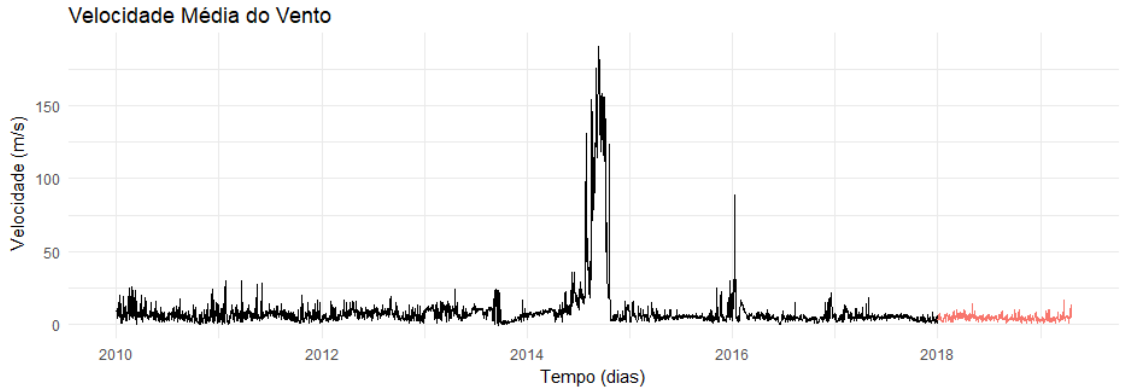


Figura 6.41: Serie de treino (a preto) e série de teste (a vermelho) da velocidade média do vento.

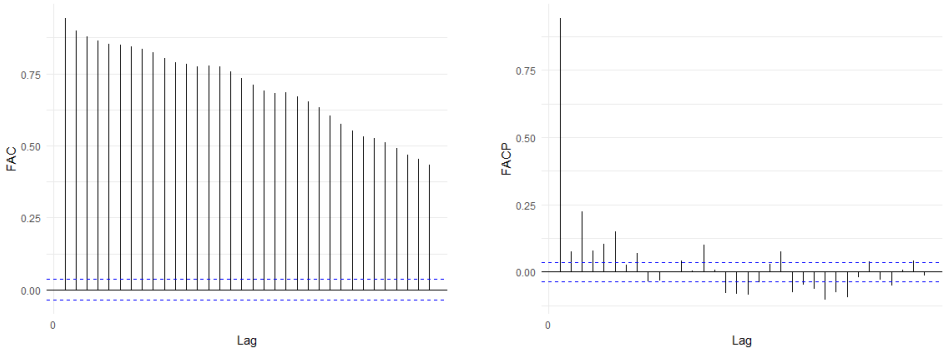


Figura 6.42: FAC e FACP da série de treino da velocidade média do vento.

Para a velocidade média do vento, o modelo TABTS obtido foi: $TBATS(1, 5, 4, 0.9, \{365, 7\})$. No modelo não foi aplicada nenhuma transformação Box-Cox ($\omega = 1$) aos dados da série temporal, foi utilizado um parâmetro de amortecimento de valor $\phi = 0,9$. O parâmetro de amortecimento é incluído no nível. A componente irregular da série temporal é correlacionada e foi modelada por um processo ARMA(5,4). O período sazonal é de 365 dias e foram utilizados 7 termos de Fourier para este período. As estimativas dos parâmetros obtidos para o modelo TBATS são apresentadas na Tabela 6.19.

Tabela 6.19: Parâmetros do modelo TBATS selecionado.

Parâmetros	α	β	ϕ	γ_1	γ_2	φ_1	φ_2
Estimativas	-0,00609	0,02319	0,89966	0,00077	-0,00055	0,93647	-0,27314
Parâmetros	φ_3	φ_4	φ_5	θ_1	θ_2	θ_3	θ_4
Estimativas	-0,02877	-0,63985	0,51606	-0,16802	-0,01101	0,14014	0,71736

A estimativa do desvio padrão do ruído branco associado ao modelo é $\hat{\sigma} = 5,62$ e o valor de AIC$= 33498,94$.

Desta forma, as equações do modelo são dadas por

$$y_t^{(1)} = l_{t-1} + 0,89966 b_{t-1} + s_{t-365}^{(1)} + d_t,$$

$$l_t = l_{t-1} + 0,89966 b_{t-1} + (-0,00609) d_t,$$

$$b_t = (1 - 0,89966) b + 0,89966 b_{t-1} + (-0,00609) d_t,$$

$$s_{j,t}^{(1)} = s_{j,t-1}^{(1)} \cos \lambda j^{(1)} + s_{j,t}^{*(1)} \sin \lambda_j^{(1)} + (0,00077) d_t,$$

$$s_{j,t}^{*(1)} = -s_{j,t-1} \sin \lambda_j^{(1)} + s_{j,t-1}^{*(1)} \cos \lambda_j^{(1)} + (-0,00055) d_t,$$

onde a componente $\lambda_j^{(1)} = \frac{2\pi j}{365}$, d_t é um processo ARMA(5,4) e α, β, γ_1 e γ_2 os parâmetros de alisamento. A sazonalidade foi modelada por 27 parâmetros (25 valores iniciais para $s_{j,0}^{(1)}$ e $s_{j,0}^{*(1)}$ e dois parâmetros de alisamento $\gamma_1^{(1)}$ e $\gamma_2^{(1)}$).

A equação do processo ARMA(5,4) é dada por

$$d_t = (0,93647)\, d_{t-1} + (-0,27314)\, d_{t-2} + (-0,02877)\, d_{t-3} + (-0,63985)\, d_{t-4}$$

$$+0,51606\, d_{t-5} + (-0,16802)\, \varepsilon_{t-1} + (-0,01101)\, \varepsilon_{t-2} + (0,14014)$$

$$+\varepsilon_{t-3} + (0,71736)\, \varepsilon_{t-4} + \varepsilon_t,$$

onde $\varepsilon_t \sim N(0, (5,62)^2)$.

Na Figura 6.43 estão representados os valores observados da série temporal (a preto) e os valores estimados pelo modelo (a vermelho), para a velocidade média do vento no período observado.

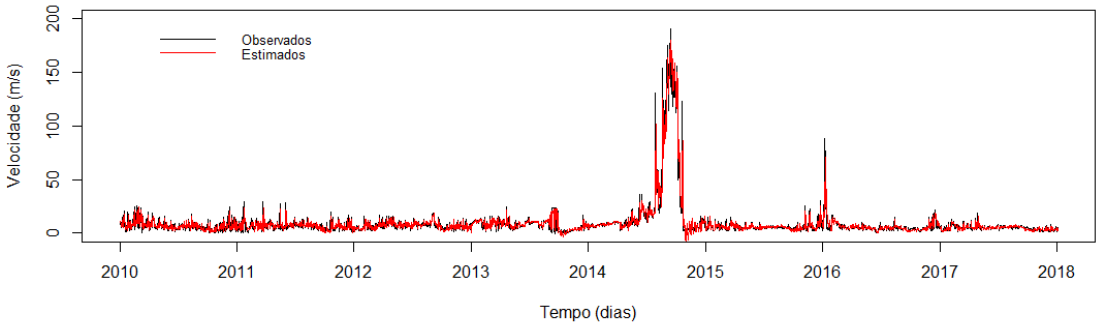


Figura 6.43: Valores observados e valores estimados pelo modelo TBATS para a velocidade média do vento.

Na Figura 6.44 são apresentados os valores observados da velocidade média do vento, as estimativas no período de modelação (período de treino) e os intervalos de previsão paa um nível de confiança de 80% e 95%, resultantes da aplicação do modelo TBATS. A Figura 6.45 mostra, em particular, as 250 últimas observações. No primeiro trimestre do ano de 2018, verifica-se que alguns valores previstos ultrapassam os intervalos de confiança.

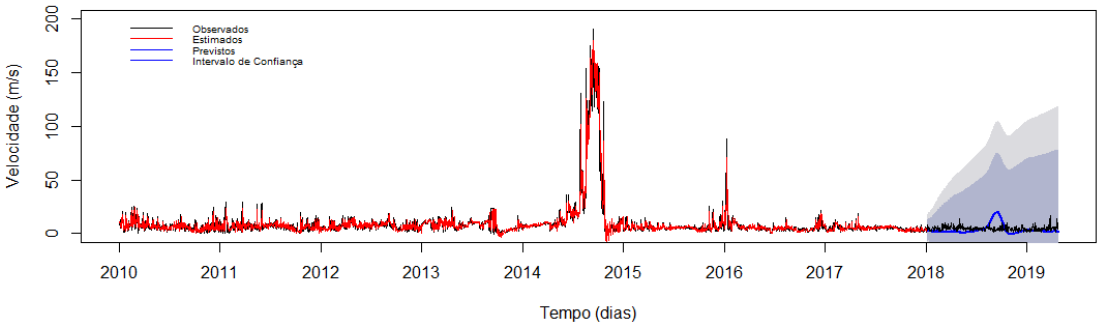


Figura 6.44: Valores observados, estimados e previstos pelo modelo TBATS para a velocidade média do vento.

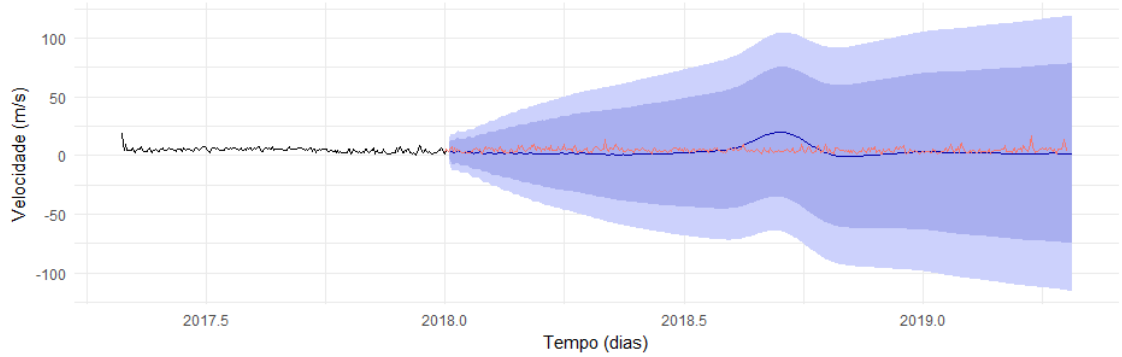


Figura 6.45: Valores observados, estimados e previstos pelo modelo TBATS para a velocidade média do vento.

Tabela 6.20: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos à velocidade média do vento.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
4,588	(-2,617; 11,793)	(-6,431; 15,607)	2,80
3,793	(-5,288; 12,873)	(-10,095; 17,680)	5,30
3,008	(-6,634; 12,650)	(-11,739; 17,754)	3,10
2,530	(-7,431; 12,492)	(-12,704; 17,765)	4,30
2,6556	(-7,525; 12,836)	(-12,914; 18,225)	6,30
3,158	(-7,161; 13,475)	(-12,623; 18,938)	7,90
3,683	(-6,860; 14,226)	(-12,442; 19,807)	3,40

O modelo foi avaliado através da análise dos resíduos representada na Figura 6.46. Para o pressuposto da independência o teste Ljung-Box, aplicado à série dos resíduos, verifica que não existe evidência estatística para assumir a independência dos erros (valor de prova ≈ 0). Para o pressuposto da normalidade dos resíduos foi aplicado o teste *Kolmogorov-Smirnov*, verificando-se que a um nível de significância de 5%, os resíduos não seguem uma distribuição Normal (valor de prova ≈ 0).

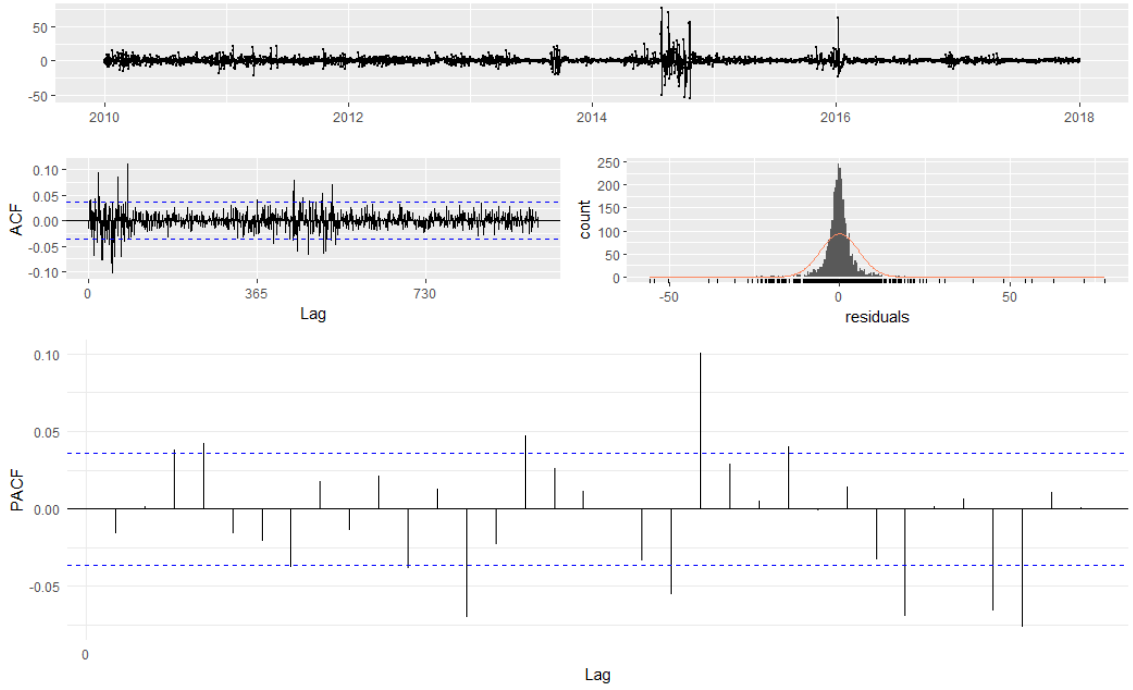


Figura 6.46: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo TBATS da velocidade média do vento.

A decomposição da série da velocidade média do vento, obtida dos valores ajustados do modelo TBATS, é apresentada na Figura 6.47.

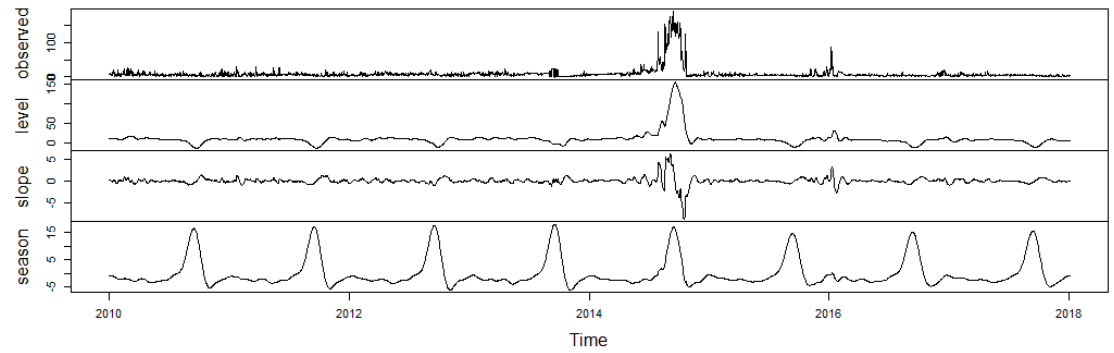


Figura 6.47: Decomposição série temporal da precipitação, obtida pela estimação do modelo TBATS.

Modelo de Regressão com Erros Correlacionados

Na Tabela 6.21 estão apresentadas as ordens dos processos ARIMA, os valores de K (para $i = 1, \ldots, 5$), os valores AICc e BIC para a seleção do modelo para a velocidade média do vento. O valor de K para o qual o valor de AIC do modelo se torna mínimo é de $K = 1$, como se pode observar na Tabela 6.21.

Tabela 6.21: Ordem do processo ARIMA, valor de K, valor AIC corrigido e BIC

Modelo	K	AICc	BIC
Regressão com erros ARIMA(5,1,0)	1	**16176,33**	**16224,12**
Regressão com erros ARIMA(5,1,0)	2	16177,51	16237,23
Regressão com erros ARIMA(5,1,0)	3	16177,63	16249,28
Regressão com erros ARIMA(5,1,0)	4	16178,10	16261,68
Regressão com erros ARIMA(5,1,0)	5	16180,10	16275,59

Para o valor de AICc = 16177,51, ε_t é um processo ARIMA(5,1,0).

A Tabela 6.22 mostra as estimativas dos parâmetros e os repetivos erros padrão, para o modelo de regressão com erros ARIMA(5,1,0). No modelo final, a estimativa do desvio padrão do ruído branco associado ao modelo é de $\hat{\sigma} = 31,78$ e o valor de AIC $= 16176,33$.

Tabela 6.22: Estimativas dos parâmetros resultantes do modelo de regressão com erros correlacionados e respetivos erro padrão.

Parâmetros	φ_1	φ_2	φ_3	φ_4	φ_5	α_1	β_1
Estimativa	-0,1734	-0,2977	-0,1477	-0,1504	-0,1629	-4,0868	-2,0551
Erro padrão	0,0198	0,0198	0,0205	0,0198	0,0197	4,8661	5,0190

Assim, as equações do modelo resultante são dadas por

$$y_t = (-4,0868)\sin\frac{2\pi t}{365} + (-2,0551)\cos\frac{2\pi t}{365} + d_t,$$

onde d_t é um processo $ARMA(5,1,0)$ descrito pela equação

$$d_t = (-0,1734)d_{t-1} + (-0,2977)d_{t-2} + (-0,1477)d_{t-3} + (-0,1504)d_{t-4} + (-0,1629))d_{t-5},$$

onde $\varepsilon_t \sim N(0,(31,78)^2)$.

A Figura 6.48 mostra os valores observados da série temporal e os valores estimados pelo modelo. A representação gráfica sugere que o modelo se ajusta de forma satisfatória aos dados.

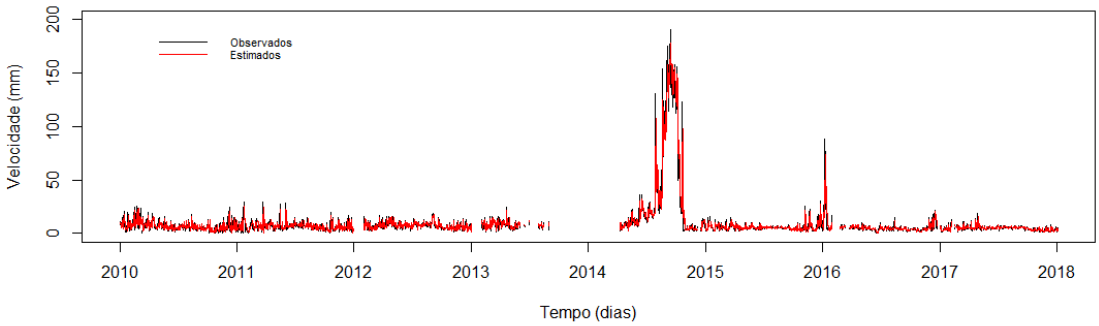


Figura 6.48: Valores observados e valores estimados pelo modelo de regressão com erros correlacionados.

Na Figura 6.49 estão representados os valores observados da velocidade média do vento, as estimativas no período de modelação (período de treino), as previsões no período de previsão (período de teste) e os intervalos de previsão para um nível de confiança de 80% e 95%, resultantes da aplicação do modelo de regressão com erros correlacionados. A Figura 6.50 mostra de forma mais visível o comportamento descrito.

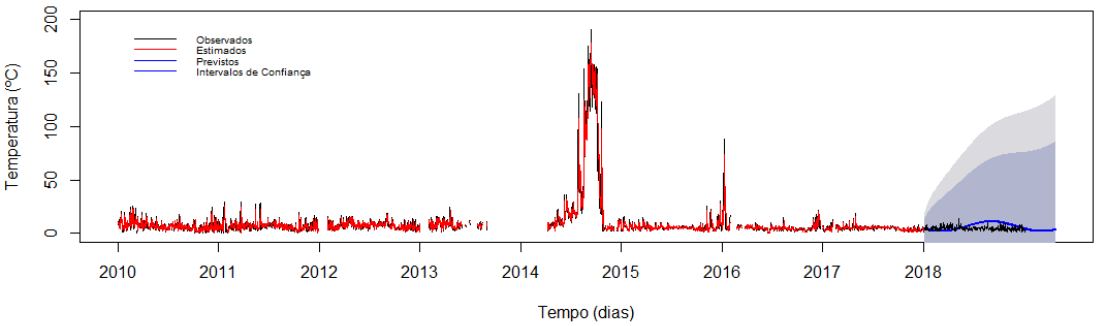


Figura 6.49: Valores observados e previsões (com intervalos de confiança de 80% e 90%) para a velocidade média do vento resultante do modelo de regressão com erros correlacionados.

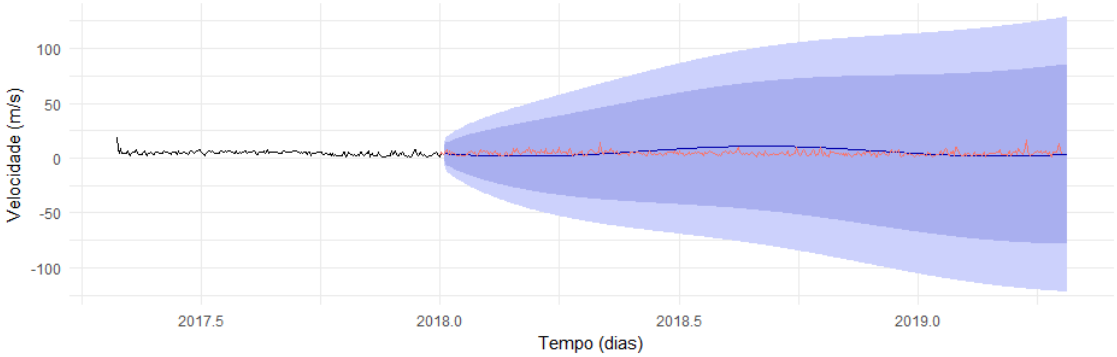


Figura 6.50: Valores observados e previsões (com limites de confiança de 80% e 95%) para a velocidade média do vento resultantes do modelo de regressão com erros correlacionados.

Na Tabela 6.23 são apresentados os resultados para as primeiras 7 observações do período de teste. Percebe-se que os intervalos de previsão apresentam amplitudes elevadas.

Tabela 6.23: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para os primeiros 7 dias de previsão, relativos à velocidade média do vento.

Valores previstos	Intervalo de Confiança a 80%	Intervalo de Confiança a 95%	Valores observados
4,657	(-2,567; 11,882)	(-6,392; 15,707)	2,80
4,457	(-4,916; 13,830)	(-9,878; 18,792)	5,30
4,026	(-6,180; 14,233)	(-11,583; 19,636)	3,10
3,444	(-7,406; 14,293)	(-13,149; 20,037)	4,30
3,446	(-7,929; 14,820)	(-13,950; 20,842)	6,30
3,658	(-8,072; 15,388)	(-14,281; 21,598)	7,90
3,679	(-8,587; 15,945)	(-15,080; 22,439)	3,40

Pelo teste Ljung-Box, aplicado à série dos resíduos existe evidência estatístisca para rejeitar a independência dos erros (valor de prova ≈ 0). Para o pressuposto da normalidade dos resíduos foi aplicado o teste Kolmogorov-Smirnov que, a um nível de significância de 5%, mostra que os resíduos não seguem uma distribuição Normal (valor de prova ≈ 0).

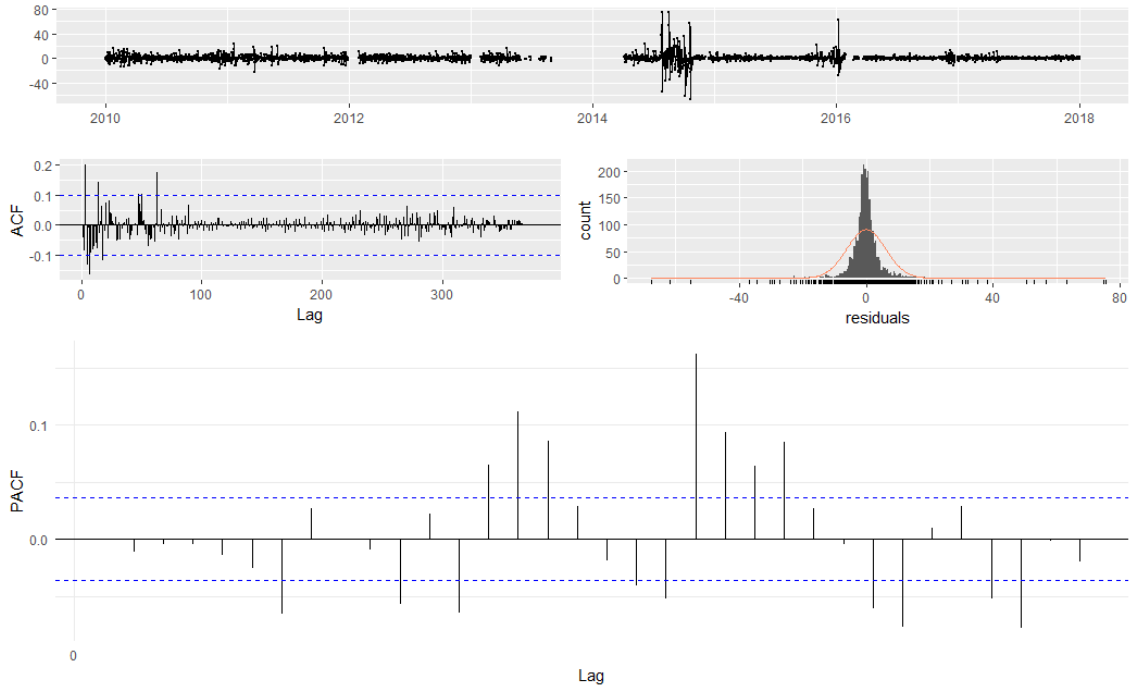


Figura 6.51: Resíduos, função de autocorrelação, histograma dos resíduos e função da autocorrelação parcial associados ao modelo de regressão com erros correlacionados da velocidade média do vento.

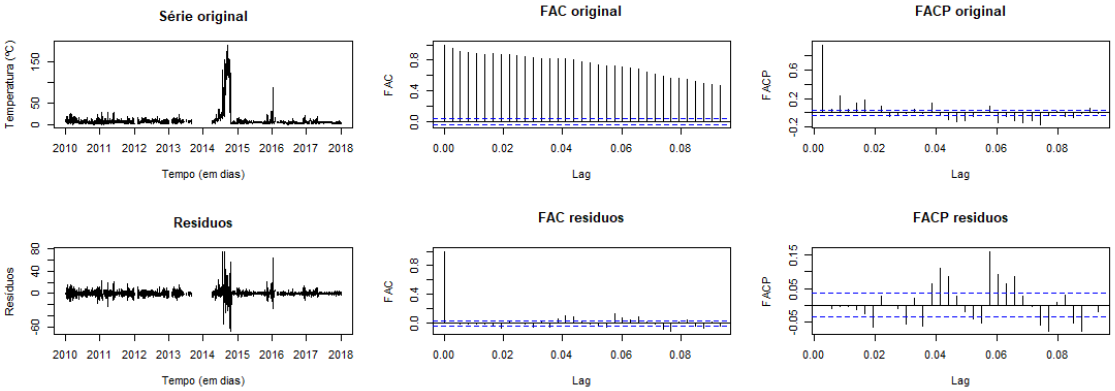


Figura 6.52: Série original e respetivas FAC e FACP, resíduos e respetivas FAC e FACP.

Avaliação da precisão dos modelos

A Tabela 6.24 mostra o resultado das medidas de precisão calculadas para os períodos de treino e teste dos dois métodos aplicados à série da velocidade média do vento. Para o período de treino o modelo TBATS apresenta melhor desempenho (com exceção da medida EEAM) e no período de teste o modelo com melhor desempenho é o MR com ARIMA(5,1,0).

Tabela 6.24: Medidas de avaliação do modelo TBATS e do modelo de regressão, para o período de teste e de treino.

Modelo	EM	REQM	EAM	EEAM	
TBATS	-0,00456	5,62211	2,90183	0,98449	Período de Treino
MR com ARIMA(5,1,0)	-0,00636	6,08616	3,07085	0,96318	Período de Treino
TBATS	1,16325	5,24639	3,85687	1,30851	Período de Teste
MR com ARIMA (5,1,0)	-0,78076	4,05068	3,32293	1,04224	Período de Teste

Capítulo 7

Considerações Finais

Este trabalho foi dedicado à análise de séries temporais de variáveis meteorológicas, nomeadamente na análise da temperatura máxima do ar, da temperatura mínima do ar, da precipitação e da velocidade média do vento, no período de 1 de janeiro de 2010 até 23 de abril de 2019 registadas numa quinta em Carrazeda de Ansiães, no Norte de Portugal.

O principal objetivo deste estudo foi identificar os modelos de previsão mais adequados, procedendo-se ao seu estabelecimento e comparação em termos da sua capacidade explicativa e preditiva. Para isso, realizou-se, então, um estudo comparativo da capacidade preditiva do modelo TBATS e do modelo de regressão linear com erros correlacionados, através da avaliação de quatro medidas distintas (ME, REQM, EAM, EEAM) para cada uma das variáveis em estudo.

Da análise efetuada foi possível concluir que o modelo de regressão com erros correlacionados, que requer menos parâmetros a serem estimados (mais parcimonioso) tem um desempenho ligeiramente melhor (no período de teste) no caso da temperatura máxima, da temperatura mínima e da velocidade média do vento. Por sua vez, no período de treino o modelo TBATS apresenta um comportamento ligeiramente melhor para a modelação da temperatura máxima e da velocidade média do vento. No caso da precipitação os dois modelos apresentam um desempenho semelhante no período de treino e de teste. De referir ainda, que apenas o modelo TBATS para a temperatura mínima cumpre os pressupostos do modelo do ponto de vista inferencial (i.e., resíduos com distribuição Normal de média nula e variância constante e independentes).

As amplitudes dos intervalos de previsão são grandes, mas apesar disso, apenas no caso da velocidade média do vento a taxa de cobertura é de 100%. Todas as restantes apresentam alguns valores fora dos limites de previsão, como é de esperar

uma vez que foram calculados intervalos de confiança a 80% e a 95%.

Em suma, o modelo TBATS e o modelo de regressão com erros correlacionados (para modelar e prever séries temporais com padrões sazonais complexos) podem de forma eficiente capturar o comportamento fortemente sazonal da temperatura máxima e da temperatura mínima do ar. Por sua vez, para a precipitação e velocidade média do vento, dificilmente as metodologias utilizadas por si só, conseguiriam lidar com características tão voláteis como as que estão presentes neste tipo de dados. Esta volatilidade acaba por influenciar os resultados obtidos, dando origem a previsões pouco precisas para o domínio temporal em que as séries se observam.

Tal mostra que apenas a modelação matemática e estatística baseada no histórico ou comportamento estocástico recente, sem a consideração de modelos físicos e geofísicos, é insuficiente para modelar e prever comportamentos meteorológicos com tanta variabilidade e tão particulares como o das séries em estudo.

Capítulo 8

Trabalho Futuro

O estudo desenvolvido fica com alguns desenvolvimentos e investigações por fazer e que não puderam ser realizados no âmbito deste trabalho. Nomeadamente, para alcançar o objetivo final do projeto "TO CHAIR - Os Desafios Óptimos na Irrigação", é necessária a obtenção de previsões mais precisas a 7 dias para o planeamento da irrigação. Assim, para investigação futura será necessário e interessante considerar-se:

- A construção de intervalos de previsão para um horizonte temporal, $h = 1, 2, \ldots, 7$, ou seja, calcular a previsão em janela a 1-passo até h-passos para cada uma das variáveis em estudo;

- Desenvolver modelos de calibração baseados no filtro de Kalman para melhorar a qualidade preditiva das previsões obtidas combinando várias fontes como as do *site* APIXU ou do IPMA, entre outros, para um horizonte temporal, $h = 1, 2, \ldots, 7$. Por exemplo, o *site* APIXU abrange dados meteorológicos de todo o planeta e as previsões são obtidas com base em dados históricos de muitos anos, que asseguram que as previsões meteorológicas sejam precisas, combinados com métodos de interpolação e de modelos matemáticos e geofísicos.

Bibliografia

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215-222. Springer.

Akram, M., Hyndman, R. J., e Ord, J. K. (2009). Exponential smoothing and non-negative data. *Australian and New Zealand Journal of Statistics*, 51(4):415-432.

Alexandratos, S. D., Barak, N., Bauer, D., Davidson, F. T., Gibney, B. R., Hubbard, S. S., Taft, H. L., e Westerhof, P. (2019). Sustaining water resources: environmental and economic impact. *ACS Sustainable Chemistry and Engineering*, 7(3):2879-2888.

Alpuim, T. e El-Shaarawi, A. (2008). On the efficiency of regression analysis with ar (p) errors. *Journal of Applied Statistics*, 35(7):717-737.

Alpuim, T. e El-Shaarawi, A. (2009). Modeling monthly temperature data in lisbon and prague. Environmetrics: *The official journal of the International Environmetrics Society*, 20(7):835-852.

Anderson, O. D. (1976). *Time series analysis and forecasting: the Box-Jenkins approach*. Butterworth.

Ansley, C. F. e Kohn, R. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *The Annals of Statistics*, pages 1286-1316.

Arca, B., Spano, D., Snyder, R. L., Fiori, M., e Duce, P. (2006). Short term forecasting of reference evapotranspiration using limited area models and time series techniques. In *17th Symposium on Boundary Layers and Turbulence, 27th Conference on Agricultural and Forest Meteorology, 17th Conference on Biometeorology and Aerobiology.*

Baranowski, P., Krzyszczak, J., Slawinski, C., Hoffmann, H., Kozyra, J., Nierobca, A., Siwek, K., e Gluza, A. (2015). Multifractal analysis of meteorological time series to assess climate impacts. *Climate Research*, 65:39-52.

Benth, J.S. e Benth, F.E. (2010). Analysis and modelling of wind speed in new york. *Journal of Applied Statistics*, 37(6):893-909.

Box, G. e Jenkins, G. (1970). Time series analysis-forecasting and control. san francisco: Holden day. 553 p.

Box, G. E. e Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211-243.

Box, G. E. e Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal os the American Statistical association*, 70(349):70-79.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345-370.

Brożyna, J., Mentel, G., Szetela, B., e Strielkowski, W. (2018). Multi-seasonality in the tbats model using demand for electric energy as a case study. *Economic Computation and Economic Cybernetics Studies and Research*, 52(1).

Chapra, S. C. e Canale, R. P. (1998). Numerical methods for engineers: With programming and software application. *WCB McGraw-Hill, Boston*.

Chatfield, C. (2000). *Time-Series Forecasting*. Chapman and Hall/CRC.

Chatfield, C. (2003). *The analysis of time series: an introduction*. Chapman and Hall/CRC.

Chen, C., Davis, R. A., e Brockwell, P. J. (1996). Order determination for multivariate autoregressive processes using resampling methods. *Journal of multivariate analysis*, 57(2):175-190.

Chen, J.-F., Wang, W.-M., e Huang, C.-M. (1995). Analysis of an adaptive time-series autoregressive moving-average (arma) model for short-term load forecasting. *Electric Power Systems Research*, 34(3):187-196.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., e Terpenning, I. (1990). Stl: a seasonal-trend decomposition. *Journal os official statistics*, 6(1):3-73.

De Livera, A. M. (2010). Exponentially weighted methods for multiple seasonal time series. *International Journal of Forecasting*. 26(4):655-657.

De Livera, A. M., Hyndman, R. J., e Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513-1527.

Dokumentov, A., Hyndman, R. J., et al. (2015). Str: A seasonal-trend decomposition procedure based on regression. *Department of Econometrics and Business Statistics, Monash University*.

Gardner Jr, Everette S e McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, 31(10):1237-1246.

Gilchrist, W. (1976). Statistical forecasting.

Gould, P. G., Koehler, A. B., Ord, J. K., Snyder, R. D., Hyndman, R. J., e Vahid-Araghi, F. (2008). Forecasting time series with multiple seasonal patterns. *European Journal of Operational Research*, 191(1):207-222.

Harvey. A. e Fernandes, C. (1989). Time series models for insurance claims. *Journal of the Institute of Actuaries*. 116(3):513-528.

Hyndman, RJ e Athanasopoulos, G. (2013). Forecasting: principles and practice.[ebook].

Hyndman, R., Koehler, A. B., Ord, J. K., e Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science and Business Media.

Hyndman, Rob J e Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., e Razbash, S. (2019). Package forecast. Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., e Razbash, S., Wang, E., e Yasmeen, F. (2018).

forecast: Forecasting functions for time series and linear models.

Hyndman, R. J. e Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679-688.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., e Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439-454.

Jain, G. (2018). Time-series analysis for wind speed forecasting. *Malaya Journal of Matematik (MJM)*, (1, 2018):55-61.

Khandakar, Y. e Hyndman, R. (2008). Automatic time series forecasting: the forecast package for rj stat. Soft.

Kostenko, A. V e Hyndman, R. J. (2008). Forecasting without significance tests? *manuscript, Monash University, Australia.*

Lamorski, K., Pastuszka, T., Krzyszczak, J., S lawinski, C., e Witkowska-Walczak, B. (2013). Soil water dynamic modeling using the physical and support vector machine methods. *Vadose Zone Journal*, 12(4).

Lee, C.-M. e Ko, C.-N. (2011). Short-term load forecasting using lifting scheme and arima models. *Expert Systems with Applications*, 38(5):5902-5911.

Lima, S. (2018). Métodos de previsão de séries temporais- uma aplicação a dados do segmento do retalho. PhD thesis, Universidade do Minho.

Makridakis, S., Wheelwright, S. C., e Hyndman, R. J. (2008). *Forecasting methods and applications*. John wiley and sons.

Metcalfe, A. V. e Cowpertwait, P. S. (2009). *Introductory time series with R*. Springer.

Monteiro, B. (2017). Modelação de Séries Temporais de Dados Oceanográficos. PhD thesis, Universidade Aberta.

Morettin, P. A. e Toloi, C. (2006). Análise de séries temporais. In *Análise de séries temporais*.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., e Stork, J. (2015). Comparison of diferent methods for univariate time series imputation in r. *arXiv preprint arXiv:1510.03924*.

Murat, Malgorzata e Malinowska, I. . H. H. . B. P. (2016). Statistical modelling of agrometeorological time series by exponential smoothing. *International agrophysics*, 30(1):57-65.

Murat, M., Malinowska, I., Gos, M., e Krzyszczak, J. (2018). Forecasting daily meteorological time series using arima and regression models. International agrophysics, 32(2):253-264.

Murteira, B., Muller, D., e Turkman, K. F. (2000). *Análise de sucessões cronológicas*.

Naim, Iram e Mahara, T. . I. A. R. (2018). Efective short-term forecasting for daily time series with complex seasonal patterns. *Procedia computer science*, 132:1832-1841.

Ord, J. K., Koehler, A. B., e Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, 92(440):1621-1629.

Pappas, S. S., Ekonomou, L., Moussas, V., Karampelas, P., e Katsikas, S. (2008). Adaptive load forecasting of the hellenic electric grid. *Journal of Zhejiang University-SCIENCE A*, 9(12):1724-1730.

Puindi, A. (2018). Contribuições para o desenho de modelos de previsão da procura: Aplicação no planeamento energético para a cidade de Cabinda. PhD thesis, Faculdade de Ciências da Universidade do Porto.

Schwarz, G. . o. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461-464.

Shenstone, L. e Hyndman, R. J. (2005). Stochastic models underlying crostons method for intermittent demand forecasting. *Journal of Forecasting*, 24(6):389-402.

Shu, M.-H., Hung, W., Nguyen, T., Hsu, B., e Lu, C. (2014). Forecasting with fourier residual modified arima model-an empirical case of inbound tourism demand

in new zealand. *WSEAS Transactions on Mathematics*, 13(1):12-21.

Shumway, R. H. e Stoffer, D. S. (2017). *Time series analysis and its applications: with R exemples*. Springer.

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799-805.

Taylor, J. W. (2010). Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting*, 26(4):627-646.

Venäläinen, A., Salo, T., e Fortelius, C. (2005). The use of numerical weather forecast model predictions as a source of data for irrigation modelling. *Meteorological Applications*, 12(4):307-318.

Wang, D. e Cai, X. (2009). Irrigation scheduling role of weather forecasting and farmers behavior. *Jounal of Water Resources Planning and Management-asce- J WATER RESOUR PLAN MAN-ASCE*, 135.

West, M. e Harrison, J.(1989). Subjective intervention in formal models. *Journal of Forecasting*, 8(1):33-53.

Wheelwright, S., Makridakis, S., e Hyndman, R. J. (1998). *Forecasting: methods and applications*. John Wiley and Sons.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324-342.

Apêndice A

Modelo TBATS (De Livera et al. (2011))

O vetor de estados do modelo TBATS com termo de crescimento não-estacionário, pode ser definido como $\boldsymbol{x_t} = (\boldsymbol{l_t}, \boldsymbol{b_t}, \boldsymbol{s_t^{(1)}}, \ldots, \boldsymbol{s_t^{(T)}}, \boldsymbol{d_t}, \boldsymbol{d_{t-1}}, \ldots, \boldsymbol{d_{t-p+1}}, \boldsymbol{\varepsilon_t}, \boldsymbol{\varepsilon_{t-1}}, \ldots, \boldsymbol{\varepsilon_{t-q+1}})'$ onde $\boldsymbol{s_t^{(i)}}$ é o vetor linha $(s_{1,t}^{(i)}, s_{2,t}^{(i)}, \ldots, s_{k_i,t}^{(i)}, s_{1,t}^{*(i)}, s_{2,t}^{*(i)}, \ldots, s_{k_i,t}^{*(i)})$.

Sejam $\mathbf{1_r} = (1, 1, ..., 1)$ e $\mathbf{0_r} = (0, 0, ..., 0)$ vetores linha de dimensão r, $\boldsymbol{\gamma_1^{(i)}} = \boldsymbol{\gamma_1^{(i)}} 1_{k_i}$, $\boldsymbol{\gamma_2^{(i)}} = \boldsymbol{\gamma_2^{(i)}} 1_{k_i}$, $\boldsymbol{\gamma^{(i)}} = (\boldsymbol{\gamma_1^{(i)}} \boldsymbol{\gamma_2^{(i)}})$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma^{(1)}}, ..., \boldsymbol{\gamma^{(T)}})$, $\boldsymbol{\varphi} = (\boldsymbol{\varphi_1}, \boldsymbol{\varphi_2}, ..., \boldsymbol{\varphi_p})$ e $\boldsymbol{\Theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ..., \boldsymbol{\theta_p})$.

Ainda, considerando $\boldsymbol{O_{u,v}}$ uma matriz de zeros de ordem $u \times v$, $\boldsymbol{I_{u,v}}$ uma matriz retangular diagonal de ordem $u \times v$ com elementos 1 na diagonal e $\boldsymbol{a^{(i)}} = (\mathbf{1_{k_i}}, \mathbf{0_{k_i}})$ e seja $\boldsymbol{a} = (\boldsymbol{a^{(1)}}, ..., \boldsymbol{a^{(T)}})$.

É também necessário definir as matrizes $\boldsymbol{B} = \boldsymbol{\gamma'\varphi}$, $\boldsymbol{C} = \boldsymbol{\gamma'\Theta}$ e

$$\boldsymbol{A_i} = \begin{bmatrix} \boldsymbol{C^{(i)}} & \boldsymbol{S^{(i)}} \\ \boldsymbol{-S^{(i)}} & \boldsymbol{C^{(i)}} \end{bmatrix}, \qquad \boldsymbol{\tilde{A}_i} = \begin{bmatrix} \boldsymbol{O_{m_i-1}} & \mathbf{1} \\ \boldsymbol{I_{m_i-1}} & \mathbf{0'_{m_i-1}} \end{bmatrix}$$

e $\boldsymbol{A} = \oplus_{\boldsymbol{i=1}}^{\boldsymbol{T}} \boldsymbol{A_i}$, onde $\boldsymbol{C^{(i)}}$ e $\boldsymbol{S^{(i)}}$ são $k_i \times k_i$ matrizes diagonais com elementos $cos(\lambda_j^{(i)})$ e $sin(\lambda_j^{(i)})$, respectivamente, para $j = 1, 2, ..., k_i$ e $i = 1, ..., T$ e onde $\oplus$ representa a soma direta das matrizes. Considere-se $\boldsymbol{\tau} = \mathbf{2} \sum_{\boldsymbol{i=1}}^{\boldsymbol{T}} \boldsymbol{k_i}$.

Assim, as matrizes do modelo TBATS podem ser escritas como

$\boldsymbol{w} = (1, \phi, \boldsymbol{a}, \boldsymbol{\varphi}, \boldsymbol{\theta})'$, $\boldsymbol{g} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, 1, \boldsymbol{0_{p-1}}, 1, \boldsymbol{0_{q-1}})'$ e

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{\phi} & \boldsymbol{0_\tau} & \boldsymbol{\alpha\varphi} & \boldsymbol{\alpha\theta} \\ \boldsymbol{0} & \boldsymbol{\phi} & \boldsymbol{0_\tau} & \boldsymbol{\beta\varphi} & \boldsymbol{\beta\theta} \\ \boldsymbol{0_\tau}' & \boldsymbol{0_\tau}' & \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0_\tau} & \boldsymbol{\varphi} & \boldsymbol{\theta} \\ \boldsymbol{0_{p-1}}' & \boldsymbol{0_{p-1}}' & \boldsymbol{0_{p-1,\tau}} & \boldsymbol{I_{p-1,p}} & \boldsymbol{0_{p-1,q}} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0_\tau} & \boldsymbol{0_p} & \boldsymbol{0_q} \\ \boldsymbol{0_{q-1}}' & \boldsymbol{0_{q-1}}' & \boldsymbol{0_{q-1,\tau}} & \boldsymbol{I_{q-1,p}} & \boldsymbol{0_{q-1,q}} \end{bmatrix}$$

.

Estas matrizes aplicam-se quando todas as componentes estão presentes no modelo. Quando uma componente é omitida, os termos correspondentes também o são.

A.0.1 Modelo BATS

A formulação em espaço de estados do modelo BATS pode ser obtida considerando $s_t^{(i)} = (s_t^{(i)}, s_{t-1}^{(i)}, ..., s_{t-(m_i-1)}^{(i)})$, $\boldsymbol{a^{(i)}} = (\boldsymbol{0_{m_{i-1}}}, \boldsymbol{1})$, $\boldsymbol{\gamma^{(i)}} = (\boldsymbol{\gamma_i}, \boldsymbol{0_{m_{i-1}}})$, $\boldsymbol{A} = \oplus_{\boldsymbol{i=1}}^{\boldsymbol{T}} \boldsymbol{\tilde{A}_i}$ e substituindo $2k_i$ por m_i nas matrizes apresentadas anteriormente para os modelos TBATS.