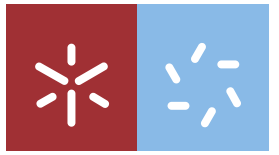
Modelos de Previsão a Curto Prazo para Variáveis Meteorológicas

Fernanda Catarina Cardoso Pereira

UMinho | 2020

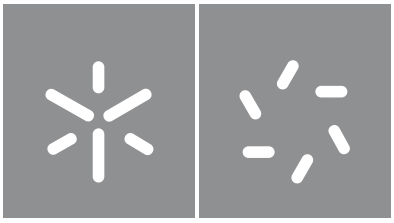


Universidade do Minho
Escola de Ciências

Fernanda Catarina Cardoso Pereira

Modelos de Previsão a Curto Prazo para Variáveis Meteorológicas

novembro de 2020



Universidade do Minho
Escola de Ciências

Fernanda Catarina Cardoso Pereira

Modelos de Previsão a Curto Prazo para Variáveis Meteorológicas

Dissertação 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

novembro de 2020

Agradecimentos

"Podes esboçar, criar e construir o lugar mais maravilhoso do mundo.
Mas são necessárias pessoas para tornar o teu sonho realidade."
(Walt Disney)

Aos meus orientadores, Professora Doutora Arminda Manuela Gonçalves e ao Professor Doutor Marco Costa, pela disponibilidade, por todo o apoio, paciência, dedicação, amizade e pela partilha de conhecimentos ao longo deste ano.

À minha querida mãe, pelo amor incondicional, pelo incentivo, conselhos, palavras carinhosas, apoio, transmissão de valores morais e por toda a base educacional. Muito obrigada.

À minha querida irmã e ao meu cunhado, por estarem presentes quando preciso, por me ouvirem, por me incentivarem a continuar e por todo o carinho.

A todos os meus amigos e colegas que me acompanharam ao longo desta jornada. Obrigada pela ajuda, incentivo e amizade.

À Professora Doutora Sofia Lopes, pela bolsa concebida.

Ao Professor Doutor Aureliano, pela cedência da base de dados.

A todos os Professores que me acompanharam e fizeram parte da minha formação.

Este trabalho foi cofinanciado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), através do SA&ICT do Programa Operacional de Competitividade e Internacionalização (POCI) - COMPETE 2020, do Portugal 2020, pela Fundação para a Ciência e a Tecnologia (FCT), através de Fundos Nacionais.

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

Nos últimos anos, devido às alterações climáticas, os períodos de seca têm sido mais frequentes e prolongados e, portanto, torna-se essencial uma gestão sustentável da água, em particular, nos sistemas de rega. Este estudo tem como principal objetivo desenvolver modelos de previsão a curto prazo (com horizonte temporal até 6 dias), para as diferentes variáveis meteorológicas que têm impacto no processo de evapotranspiração. Assim, são propostos os modelos de espaço de estados que permitem lidar com séries temporais com um comportamento instável, que é uma característica predominante nos dados meteorológicos. Estes modelos são muito flexíveis permitindo incorporar componentes estocásticas bastante úteis na previsão a curto prazo, melhorando, assim, a qualidade preditiva das previsões obtidas. Os modelos propostos estão associados ao filtro de Kalman na obtenção das predições ótimas das variáveis não observáveis, uma vez que é um algoritmo recursivo que atualiza e melhora as previsões do vetor de estados em tempo real, sempre que novas observações ficam disponíveis.

Neste estudo são analisadas duas bases de dados distintas: a primeira corresponde a registos diários de variáveis meteorológicas, temperaturas máxima e mínima do ar, registadas numa estação meteorológica instalada numa quinta em Carrazeda de Ansiães, situada em Bragança, região Norte de Portugal, recolhidas no período de 20 de fevereiro a 11 de outubro de 2019; a segunda base de dados é referente às previsões provenientes do *website weatherstack.com* com horizonte temporal até 6 dias relativas às mesmas variáveis meteorológicas e no mesmo período temporal.

Apresenta-se uma comparação entre modelos de previsão, nomeadamente os modelos de espaço de estados associados ao filtro de Kalman, em particular o modelo de calibração, e os modelos de regressão linear simples, que constituem uma classe particular do modelo de calibração, considerando o estado determinístico.

Esta dissertação foi realizada no âmbito de uma Bolsa de Investigação (BI) do Projeto "TO CHAIR - Os Desafios Ótimos na Irrigação", cofinanciado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), através do SA&ICT do Programa Operacional de Competitividade e Internacionalização (POCI) - COMPETE 2020, do Portugal 2020, pela Fundação para a Ciência e a Tecnologia (FCT), através de Fundos Nacionais.

Palavras-chave: Irrigação; Séries temporais; Previsões a curto prazo; Variáveis meteorológicas; Calibração; Modelo de espaço de estados; Filtro de Kalman.

Abstract

Dry periods have been more frequent and prolonged in recent years due to climate change and, therefore, sustainable water management has become essential, particularly in irrigation systems. This study's main goal is to develop short-term forecasting models (with a time horizon of up to 6 days) for the different meteorological variables that have an impact on the evapotranspiration process. Thus, we propose state space models that allow dealing with time series with unstable behavior, which is a predominant feature in meteorological data. These models are very flexible and allow incorporating stochastic components, which are very useful in the short-term forecast, thus improving the predictive quality of the obtained forecasts. The proposed models are associated to the Kalman filter for obtaining the unobservable variables' optimal predictions, since it is a recursive algorithm that updates and improves the state vector forecasts in real time whenever new observations become available.

In this study, two distinct databases are analyzed: the first one corresponds to daily records of meteorological variables, maximum and minimum air temperatures recorded at a weather station installed in a farm in Carrazeda de Ansiães, in Bragança, in the northern region of Portugal, collected in the period between February 20, 2019 and October 11, 2019; the second database refers to the forecasts from the weatherstack.com website with a time horizon of up to 6 days for the same meteorological variables and same time period.

We present a comparison of the forecast models, namely the state space models associated with the Kalman filter, particularly the calibration model, and the simple linear regression models, which constitute a particular class of the calibration model when considering the deterministic state.

This dissertation was carried out as part of a Research Grant (RG) from Project "TO CHAIR – The Optimal Challenges in Irrigation", co-funded by the European Regional Development Fund (FEDER), through the SA&ICT of the Operational Program of Competitiveness and Internationalization (POCI) - COMPETE 2020, Portugal 2020, by the Foundation for Science and Technology (FCT), through National Funds.

Keywords: Irrigation; Time series; Short-term forecasts; Meteorological variables; Calibration; State space model; Kalman filter.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

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

AR - *Autoregressive model* (em português, modelo Autorregressivo)

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

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

BI - Bolsa de Investigação

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

CV - *Cross-Validation* (em português, Validação Cruzada)

EAM - Erro Absoluto Médio

EEAM - Erro Escalado Absoluto Médio

EM - Erro Médio

EPAM - Erro Percentual Absoluto Médio

EQM - Erro Quadrático Médio

FAC - Função de Autocorrelação

FACP - Função de Autocorrelação Parcial

FCT - Fundação para a Ciência e a Tecnologia

FEDER - Fundo Europeu de Desenvolvimento Regional

FK - Filtro de Kalman

IC - Intervalo de Confiança

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

MV - Método da Máxima Verosimilhança

QQ-plot - *Quantile-Quantile plot* (em português, gráfico Quantil-Quantil)

REQM - Raiz quadrada do Erro Quadrático Médio

RLS - Regressão Linear Simples

Capítulo 1

Introdução

A água é o principal recurso natural essencial à vida na Terra. Praticamente todas as atividades humanas, como a agricultura, produções industrial e energética, dependem deste recurso limitado. Além disso, a água está presente em todos os processos vitais dos seres vivos, tendo um papel importante na regulação dos ecossistemas e na manutenção através do ciclo da água.

No entanto, devido ao desenvolvimento económico, os recursos hídricos têm sido bastante explorados e a má gestão destes recursos tornou-se uma preocupação global. Embora quase 70% do planeta esteja coberto por água, estima-se que apenas 2,5% desta seja doce. E, mesmo assim, apenas 1% da água doce encontra-se em locais acessíveis ao ser humano, sendo que a restante encontra-se retida em glaciares e calotas polares (National Geographic, 2020).

O crescimento populacional, a poluição e as alterações climáticas também têm contribuído para a diminuição dos recursos hídricos do planeta. Segundo o Relatório Mundial das Nações Unidas sobre o Desenvolvimento dos Recursos Hídricos (2020) (WWAP, 2020), nos últimos cem anos o consumo mundial da água aumentou seis vezes e continua a crescer de forma constante a uma taxa de cerca de 1% todos os anos resultante do desenvolvimento económico e das mudanças nos padrões de consumo. Prevê-se que em 2050 a população mundial, atualmente de 7,6 mil milhões de indivíduos, chegará aos 9,8 mil milhões. Neste relatório também são discutidas algumas soluções para aumentar a disponibilidade dos recursos hídricos, como por exemplo a dessalinização da água e a recolha da humidade da atmosfera. O grande inconveniente da dessalinização é que, geralmente, esta técnica consome muita energia, contribuindo para o aumento das emissões de gases com efeito de estufa, caso a fonte de energia não seja renovável. As energias renováveis como a eólica, a solar fotovoltaica e certos tipos de energia geotérmica são, de longe, as melhores alter-

nativas energéticas a partir de uma perspetiva do consumo hídrico. A agricultura (incluindo irrigação, pecuária e aquicultura) é sem dúvida a atividade económica que mais água consome, utilizando cerca de 69% dos recursos hídricos disponíveis anualmente. A indústria consome 19%, e as residências particulares 12% (WWAP, 2019).

A seca é um fenómeno de origem climática provocada pela precipitação inferior à esperada sobre um determinado território, ao longo de um determinado período de tempo, levando à falta de água para satisfazer as necessidades essenciais existentes. Segundo o relatório da Associação Natureza Portugal em Associação com a World Wide Fund for Nature (2019) (ANP/WWF, 2019), grande parte dos cenários apontam para uma redução da disponibilidade da água em Portugal, principalmente na região Sul, causadas pelo aumento da temperatura que se tem verificado nos últimos anos. Esta situação provoca não só desiquilíbrios no processo de evaporação, como também leva à redução dos escoamentos superficiais e subterrâneos, alargamento da estação seca e concentração da precipitação em menos dias, tendência a longo prazo de redução da precipitação de inverno (já estatisticamente significativa para o mês de março), entre outras (ANP/WWF, 2019). É evidente que a seca causa impactos negativos que podem ser tanto a nível económico, como, por exemplo, a perda de produção e redução das áreas cultivadas, como a nível social e ambiental, onde a redução da quantidade e da qualidade da água disponível provoca o aumento da mortalidade de muitas espécies (perda de fauna e flora), o desequilíbrio nos povoamentos, sobretudo piscícolas, e a expansão de espécies exóticas.

As alterações climáticas constituem, num contexto global, um agravamento das ameaças associadas aos fenómenos meteorológicos naturais que afetam o ciclo da água e a disponibilidade hídrica (ANP/WWF, 2019). Desta forma, tornou-se fundamental encontrar as melhores soluções técnicas para melhorar a eficiência do uso da água, em particular, nos sistemas de rega. O uso mais eficiente da água na agricultura carece de um conhecimento primordial das caraterísticas do solo e das necessidades hídricas das plantas aí instaladas. Um dos principais fatores a ter em conta na estimativa dessas necessidades é a evapotranspiração, que é um processo pela qual a água é transferida do solo para a atmosfera por evaporação e pela transpiração das plantas. A intensidade da evapotranspiração depende de vários fatores que estão relacionados ao clima, como por exemplo a radiação solar, a humidade relativa do ar, a velocidade do vento, a precipitação e a temperatura, assim como da própria cultura, como por exemplo a dimensão das plantas e a fase do desenvolvimento em que se encontram. Os sistemas de irrigação mais comuns são a irrigação de

superfície, irrigação por aspersão e irrigação localizada (Moreira, 2019). Na irrigação de superfície, a água desloca-se ao longo do terreno através do fluxo da gravidade devido à orografia do solo e infiltra-se. Não recorre à bombagem, exceto para colocar a água à superfície do solo. Na irrigação por aspersão, a água é distribuída por toda a superfície do solo sob a forma de aspersão. A água é conduzida sob pressão por meio de bombeamento (implicando consumo de energia), ao longo de uma rede de tubagens até aos aspersores. Existem dois tipos de sistemas de irrigação por aspersão: os sistemas estacionários e os sistemas móveis. Nos sistemas estacionários, os aspersores permanecem numa posição fixa, enquanto que nos sistemas móveis, a água é aplicada enquanto os aspersores se movimentam (não são fixos). Na irrigação localizada, a água é aplicada sob pressão apenas nas zonas do solo em que se desenvolvem as raízes das plantas.

Este estudo foi realizado no âmbito de uma Bolsa de Investigação (BI) do Projeto "TO CHAIR - Os Desafios Ótimos na Irrigação, PTDC/MAT- APL/28247/2017, ref.ª: POCI-01-0145-FEDER-028247", cofinanciado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), através do SA&ICT do Programa Operacional de Competitividade e Internacionalização (POCI) - COMPETE 2020, do Portugal 2020, pela Fundação para a Ciência e a Tecnologia (FCT), através de Fundos Nacionais. O projeto envolve 4 Universidades, nomeadamente a Universidade do Minho, a Universidade de Aveiro, a Universidade do Porto e a Universidade de Trás-os-Montes e Alto Douro, incluindo 19 investigadores e vários bolseiros. Este projeto pretende analisar o comportamento da humidade no solo através da modelação, cujo principal objetivo baseia-se essencialmente na gestão eficiente dos recursos hídricos nos sistemas de rega. Neste projeto são, assim, analisados dados facultados pela Universidade de Trás-os-Montes e Alto Douro e pelo *website weatherstack*, com o objetivo último de estimar e prever adequadamente as perdas de água por evapotranspiração.

Nesta dissertação pretende-se desenvolver modelos de previsão a curto prazo (com horizonte temporal até 6 dias) para diferentes variáveis meteorológicas, nomeadamente temperatura máxima e a temperatura mínima, cujo principal objetivo foca-se essencialmente na obtenção de previsões mais precisas para as diferentes variáveis meteorológicas. Assim, são propostos os modelos com a formulação de espaço de estados que constituem uma ferramenta muito flexível para estudar fenómenos dinâmicos e sistemas em evolução que estabelecem uma relação entre a variável observada, a variável não observada (estado) e a componente aleatória, designados por modelos de calibração. A análise e a modelação de séries temporais na sua representação de espaço de estados são bastante eficazes no ponto de vista estocástico, cujos

principais objetivos baseiam-se na previsão a curto prazo e obtenção de estimativas filtradas ou alisadas. Estes modelos incorporam uma componente não observável, o estado, que precisa de ser estimada. Para isso, recorre-se ao filtro de Kalman (FK), que é um algoritmo de estimação recursivo que permite obter o estimador ótimo do vetor de estados, baseado na informação disponível até ao instante t e obter previsões a 1-passo, atualizando e melhorando as previsões do vetor de estados, em tempo real, quando novas observações ficam disponíveis.

Neste trabalho apresenta-se uma comparação entre dois modelos de previsão, nomeadamente os modelos de espaço de estados (MEE) associados ao FK, em particular o modelo de calibração, e os modelos de regressão linear simples(RLS), que constituem uma classe particular do modelo de calibração, considerando o estado determinístico. A formulação de um problema na representação de espaço de estados pretende evidenciar uma dependência funcional, dinâmica e estocástica entre componentes de um sistema, que pode ser representada a partir de duas equações: a primeira é chamada de equação de estado que traduz o modelo estocástico subjacente ao vetor de estados; a segunda é chamada de equação de observação que relaciona a variável observada com uma transformação linear do vetor de estados adicionada de um ruído branco.

Para a realização da análise estatística, recorreu-se ao *software* estatístico R (versão 3.5.1). Para a implementação dos MEE associados ao FK, foram criados novos códigos, na qual foram utilizados algumas funções implementadas no *package* "astsa", *Applied Statistical Time Series Analysis*, (versão 1.9), da autoria de David Stoffer (08/05/2019). Mais informações sobre este *package* podem ser encontradas em Shumway e Stoffer (2017, 2019).

1.1 Estado da Arte

Este trabalho surge no seguimento da linha de investigação de Costa (2019), cujas metodologias propostas para estudar as séries temporais de dados meteorológicos foram os modelos TBATS, que incorporam componentes de transformação Box-Cox, erros ARMA, tendência e componentes sazonais trigonométricas, e os modelos de regressão linear com erros correlacionados. A justificação para a aplicação destas metodologias deve-se ao facto da necessidade de lidar com a presença de correlação temporal forte e sazonalidade complexa. Ambos os modelos possuem uma estrutura flexível que permitem integrar as várias características dos dados, para além de poderem ser aplicadas em séries temporais não estacionárias.

Um dos principais objetivos do projeto "TO CHAIR" consiste em estimar e prever adequadamente as perdas de água por evapotranspiração através da modelação de séries meteorológicas que têm impacto no processo de evapotranspiração. Com a mesma finalidade, Lopes *et al.* (2016) propõem um modelo matemático com o objetivo de estudar e planear o uso eficiente da água na irrigação de uma determinada área agrícola, de forma a garantir que a cultura do campo seja mantida em bom estado de preservação. O modelo é baseado numa equação dinâmica que traduz a equação do balanço hidrológico. Refere ainda que, devido à imprevisibilidade do estado meteorológico, a precipitação pode ser difícil de estimar com precisão. Para dar conta dessa eventualidade, propuseram o recálculo da estratégia ótima de cada vez que novos dados são obtidos, que consiste basicamente no replaneamento ou recuo do horizonte temporal.

De acordo com Petris *et al.* (2009), os MEE fornecem uma ferramenta bastante flexível, mas ao mesmo tempo muito simples para analisar fenómenos dinâmicos e sistemas em evolução, e têm contribuído significativamente para estender os domínios clássicos de aplicação da análise de séries temporais para processos irregulares e não estacionários, para sistemas que evoluem em tempo contínuo, para dados multivariados contínuos e discretos. Além disso, refere que a classe mais geral dos MEE estende a análise aos sistemas dinâmicos não-Gaussianos e não-lineares.

Segundo Shumway e Stoffer (2017), o modelo surgiu na configuração de rastreamento de espaço, onde a equação de estado definia as equações de movimento para a posição ou estado de uma nave espacial e os dados refletiam as informações que podiam ser observadas a partir de um dispositivo de rastreamento, como a velocidade e o azimute[1].

Nas últimas décadas, o FK, aliado ao desenvolvimento tecnológico, tem sido uma ferramenta bastante poderosa no tratamento estatístico de modelos com a representação de espaço de estados que permite obter estimativas e previsões de variáveis não observáveis (vetor de estados) através de um conjunto de equações recursivas, onde os dados vão sendo atualizados sequencialmente a 1-passo sempre que uma nova observação é introduzida. De uma forma geral, o FK é um algoritmo recursivo que permite estimar um estado não observado a partir de um processo observado. Esta abordagem foi desenvolvida nos anos 60, quando Kalman (1960) publicou um artigo que descrevia uma solução recursiva para o problema de filtragem linear para dados discretos.

[1]Entende-se por azimute de uma dada direção o ângulo que essa direção faz com uma direção de referência, que na Topografia é a direção do Norte Cartográfico. O valor deste ângulo pode ser determinado de várias maneiras (Vicente, 1997).

Os modelos de regressão linear têm sido a abordagem mais aplicada quando um modelo de previsão é necessário. No entanto, Gonçalves e Costa (2013) salientam que é pouco provável que os modelos estatísticos com efeitos fixos produzam uma boa precisão preditiva, particularmente em situações onde a relação entre o preditor e a covariável não seja constante ao longo do tempo. Neste artigo, também referem que os modelos de espaço de estado linear têm o potencial de superar o modelo de regressão linear usual em termos de sua capacidade de incorporar a dinâmica temporal inerente ao procedimento em estudo.

Nas séries temporais de dados meteorológicos, é muito comum faltar um número significativo de dados e, quando os valores ausentes causam erros, Hyndman e Athanasopoulos (2018) referem que há pelo menos duas alternativas para lidar com o problema. Uma seria apenas utilizar todos os dados após o último valor ausente, supondo que existam observações suficientes para produzir previsões significativas. Outra alternativa seria substituir os valores em falta por estimativas. O FK permite modelar séries temporais com valores em falta. Por exemplo, Linroth (2014) aplicou o FK a dados obtidos desde janeiro de 1983 até dezembro de 2003, relativos à altura das ondas do mar, onde faltavam cerca de 17% dos dados. Para além da presença de valores em falta neste tipo de dados, existem também eventualidades que os perturbam. Costa e Gonçalves (2011) referem que os dados ambientais são naturalmente afetados pelas diferentes estações do ano e pelas alterações climáticas que se têm verificado de forma mais agravada nos últimos anos. Portanto, qualquer modelo de previsão deve ter em consideração esses dois fatores.

Os MEE possuem uma estrutura muito versátil e graças a esta característica, têm sido aplicados em diversas áreas, sobretudo na área ambiental. Por exemplo, Costa e Alpuim (2010, 2011) consideraram modelos de espaço de estado associados ao FK na calibração de observações da precipitação do radar meteorológico, com o objetivo de melhorar as estimativas da precipitação numa determinada área durante um período de tempo, relacionando as observações obtidas de radar meteorológico e de um pluviómetro (ou udómetro), onde o vetor de estados funciona como um fator de correção que varia ao longo do tempo. Uma vez que os dados relativos à precipitação não seguiam uma distribuição Normal, a aplicação do método da máxima verosimilhança para estimar os parâmetros do MEE não demonstrou ter um bom desempenho. Neste sentido, em alternativa propuseram uma abordagem independente da distribuição de base na qual, ao contrário do método da máxima verossimilhança, os parâmetros do modelo podem ser calculados sem assumir qualquer distribuição específica para os erros. Para avaliar o desempenho dos modelos,

recorreram à Raiz quadrada do Erro Quadrático Médio (REQM). No geral, os MEE mostraram ser uma abordagem eficiente para melhorar a precisão da estimativa do radar meteorológico da precipitação. Outro exemplo, Gonçalves *et al.* (2018) propuseram uma abordagem baseada em modelos estruturais de séries temporais com a representação de espaço de estados associados ao FK com o objetivo de analisar e avaliar a evolução temporal de séries de variáveis ambientais, em particular no contexto de um problema de monitorização da qualidade da superfície da água numa bacia hidrográfica, identificando tendências e possíveis mudanças na qualidade da água num contexto dinâmico de controlo. Além disso, ainda referem que a abordagem proposta permite obter resultados pertinentes relativamente à avaliação da qualidade da superfície da água e de pontos de mudança.

Para avaliar a qualidade da água, Dabrowski *et al.* (2018) apresentam um estudo comparativo, cujo objetivo é fornecer assistência aos criadores de camarão de viveiro no monitorização da qualidade da água com dados limitados, pois uma má gestão em termos da qualidade da água pode provocar grandes perdas nas criações. Os indicadores de qualidade utilizados foram o oxigénio dissolvido e o pH. Compararam o desempenho dos MEE Gaussiano linear e não linear, associados ao FK, cujos modelos foram baseados na estrutura dos dados e não nos processos subjacentes que os geraram. A medida de avaliação utilizada foi o Erro Percentual Absoluto Médio (EPAM) e concluíram que, no geral, o modelo linear Gaussiano obteve um melhor desempenho.

Álvarez *et al.* (2019) propuseram um novo algoritmo do FK que fornece uma análise estatística formal aos dados espaço-temporais com uma estrutura autorregressiva que permite captar tanto a dependência temporal quanto a estrutura de correlação espacial através da formulação de espaço de estados. Com o principal objetivo de realizar inferência estatística em termos de estimativa de parâmetros e previsão em locais não observados, foram comparadas a abordagem do FK com a previsão clássica de krigagem por meio de um estudo de simulação, onde mostraram, de uma forma geral, que o desempenho do FK foi superior tanto na estimação como na previsão dos dados espaço-temporais. Ao contrário da metodologia de krigagem, que utiliza apenas informações associadas ao instante atual t, o FK incorpora todas as informações disponíveis até ao instante t, não limitando as informações temporais gerais.

Achar *et al.* (2020) propõem uma nova metodologia para a previsão do tempo de chegada dos autocarros, em tempo real, cujos dados são de natureza espaço-temporal. O modelo proposto deteta a ordem desconhecida da dependência espacial

dos dados e, de seguida, captura as correlações temporais entre as viagens sucessivas em função da diferença do tempo. O modelo de previsão proposto for reescrito na formulação de espaço de estados linear, onde foi aplicado posteriormente o FK que permitiu obter previsões estatisticamente ótimas.

O método de otimização adotado neste trabalho é o método de Broyden-Fletcher-Goldfarb-Shanno (1970), usualmente conhecido como BFGS, que é um método numérico derivado dos métodos de otimização de Newton onde é assumido que a derivada parcial da função de log-verosimilhança em relação ao vetor dos parâmetros pode ser aproximada localmente através da expansão da série de Taylor em torno do ponto ótimo, onde são utilizadas as primeira e segunda derivadas, que correspondem ao gradiente e à matriz Hessiana, respetivamente. Este método iterativo é interrompido através de um critério de convergência definido de acordo com um determinado nível de tolerância. Shumway e Stoffer (2017) aplicaram o método BFGS a dados relativos à taxa de inflação trimestral para obter as estimativas de máxima verosimilhança dos parâmetros do modelo com a representação de espaço de estados que relaciona a taxa de inflação trimestral com a covariável relativa à taxa de juros trimestral.

1.2 Estrutura da Dissertação

Atendendo aos objetivos a serem alcançados, no Capítulo 2 faz-se uma introdução às séries temporais, expondo conceitos importantes, assim como são apresentados os principais modelos em séries temporais.

Nos Capítulos 3 e 4 é introduzida e desenvolvida a teoria relacionada aos MEE, nomeadamente a sua formulação, as propriedades inerentes importantes, onde são dados alguns exemplos de modelos estruturais. De seguida, apresenta-se o algoritmo do FK, assim como o alisamento e previsão. O Capítulo 4 termina com uma breve comparação entre os MEE e os modelos ARMA.

O Capítulo 5 é dedicado à estimação dos parâmetros do MEE, que é feita através do método de máxima verosimilhança.

No Capítulo 6 são descritas algumas medidas de avaliação e critérios de seleção utilizados para comparar o desempenho dos modelos. Também apresenta-se o método de validação cruzada para séries temporais e alguns testes estatísticos necessários para verificar se os pressupostos dos modelos são verificados.

No Capítulo 7 prossegue-se com a aplicação dos MEE às séries da temperatura máxima e mínima do ar. Inicialmente é feita uma breve apresentação da base de dados, seguida da análise exploratória de das variáveis meteorológicas em questão

e, por fim são exibidos os resultados resultantes do ajustado dos modelos, onde é feita uma análise comparativa em termos da capacidade preditiva, qualidade de ajustamento e análise dos resíduos.

Por fim, no Capítulo 8 são expostas as principais conclusões e propostas para o desenvolvimento do trabalho futuro.

Capítulo 2

Séries Temporais

Uma série temporal, $\{y_t,\ t = 1, 2, \ldots, n\}$, é definida como sendo um conjunto de observações indexadas no tempo de um dado fenómeno. A ordem e o momento aos quais cada uma das observações está associada são utilizados posteriormente para analisar e modelar as séries em estudo.

A análise de séries temporais tem uma vasta aplicabilidade em diferentes áreas, tais como na economia, medicina, meteorologia, engenharia, entre outros.

As séries temporais podem ser classificadas como discretas ou contínuas, dependendo do suporte do instante em que foram observadas. Quando as observações são medidas em pontos de tempo específicos, dizem-se que são discretas, quando são medidas praticamente em todos os instantes de tempo, dizem-se que são contínuas.

Normalmente, nas séries temporais discretas, as observações são registadas em intervalos de tempo igualmente espaçados, com separações temporais diárias, semanais, mensais, anuais, etc. No entanto, uma série temporal contínua pode ser facilmente transformada numa série discreta, discretizando os dados em instantes de tempo específicos. Ao longo desta dissertação, o estudo será limitado às séries temporais discretas.

Dada a natureza deste tipo de observações, os dados de séries temporais têm, em regra, como principal característica a presença de uma dependência temporal entre elas. Ou seja, numa série temporal, não se pode assumir que as observações são independentes entre si. A dependência das observações é um conceito primordial e é através dessa dependência que é feita a análise e a modelação das séries.

A análise de séries temporais pode ser feita de várias formas, dependendo dos objetivos as serem alcançados. De um modo geral, Murteira *et al.* (1993) destacam as quatro principais motivações:

a) **Descrição**. A descrição de uma série é uma tarefa primária cuja finalidade passa por compreender como a variável se altera e/ou se relaciona com uma ou mais variáveis ao longo de um determinado período de tempo. Pode ser feita através da construção de um cronograma, cálculo das medidas descritivas, como a média aritmética, variância, a taxa média de variação, etc;

b) **Explicação**. Encontrar o melhor modelo que permita explicar a evolução de uma série temporal;

c) **Previsão**. Antecipar a evolução no futuro das séries temporais em estudo, a partir do modelo que melhor se ajustou aos dados, de modo a obter previsões fiáveis das observações futuras;

d) **Controlo**. Uma série temporal pode traduzir uma característica quantitativa de uma produção em série. Admite-se que um processo está sob controlo se uma determinada característica se mantém dentro dos limites previamente estabelecidos; caso contrário, a produção será interrompida e terá de se procurar corrigir os fatores responsáveis pelo comportamento atípico.

Decomposição de uma série temporal

Geralmente, uma série temporal pode incorporar quatro componentes que podem ser separadas a partir dos dados observados. Esses componentes são: a tendência, a sazonalidade, a componente cíclica e a componente aleatória (Murteira *et al.*, 1993).

A **componente de tendência** (T_t) representa a variação, em média, ao longo do tempo, descrevendo um movimento regular e consistente durante períodos longos. Permite definir o comportamento padrão (crescente, decrescente, linear, não linear,...) presente na série. Esta componente pode ser consequência de valores observados que dependam de uma componente determinística que é função monótona do tempo.

A **componente de sazonalidade** (S_t) corresponde às variações que se repetem periodicamente, isto é, está relacionada com as oscilações que ocorrem na série durante um determinado período de tempo (semanal, mensal, anual, etc). As variações sazonais podem ser explicadas por diversos fatores, como por exemplo o clima e as

estações do ano, cujo efeito é intenso nas atividades tais como a agricultura, turismo, consumo de energia, costumes tradicionais, etc.

A **componente cíclica** (C_t) associa-se a fases alternadas de expansão e depressão que não apresentam qualquer tipo de periodicidade definida, podendo esconder uma evolução no tempo característica. Estes ciclos são dificilmente separáveis da tendência e, além disso, são difíceis de se prever. A maior parte das séries temporais económicas e financeiras, como por exemplo a série de produções e preços, mostram algum tipo de variação cíclica.

A **componente aleatória** (E_t) é a componente estocástica engloba tudo o que não se consegue definir ou modelar. Esta componente tem um papel muito importante nos modelos probabilísticos.

A análise de uma série temporal, de uma forma geral, considera uma das seguintes estruturas ou modelos

1. modelo aditivo
$$Y_t = T_t + S_t + C_t + E_t;$$

2. modelo multiplicativo
$$Y_t = T_t \cdot S_t \cdot C_t \cdot E_t;$$

3. modelos mistos
$$Y_t = (T_t + C_t)S_t + E_t \quad \text{ou} \quad Y_t = T_t \cdot S_t \cdot C_t + E_t.$$

O modelo que irá obter o melhor ajustamento varia de série para série e, quando a ideia de decomposição é aceite, a melhor opção será fazer várias tentativas até chegar ao modelo que produza a menor componente residual sem prejudicar a aleatoriedade (Murteira *et al.*, 1993). Um modelo aditivo é facilmente obtido a partir de um multiplicativo através da função logarítmica

$$\log(Y_t) = \log(T_t) + \log(S_t) + \log(C_t) + \log(E_t).$$

2.1 Processos Estocásticos Estacionários

Os modelos determinísticos não são os modelos mais adequados para estudar fenómenos dinâmicos observados no mundo real. O objetivo da teoria dos processos

estocásticos consiste em encontrar o modelo probabilístico que melhor descreva o comportamento de um determinado fenómeno, a fim de se obter boas previsões.

Assim, uma série temporal pode ser vista como uma realização de um processo estocástico cujo espaço de parâmetros é um conjunto discreto de índices que representa o tempo.

Definição 2.1.1. *Um processo estocástico é qualquer família ou coleção de variáveis aleatórias* $Y_t,\ t \in T$, *em que* T *é um conjunto de índices que representa o tempo.*

O conjunto de índices T chama-se **espaço de parâmetros** e o contradomínio das variáveis aleatórias Y_t designa-se por **espaço de estados**, que se representa por S. Para $T = \mathbb{Z}$ ou $T = \mathbb{N}$ diz-se que o processo é de tempo discreto. Caso $T = \mathbb{R}$ ou $T = \mathbb{R}^+$, diz-se que o processo é de tempo contínuo.

Um processo estocástico pode ser estacionário se o sistema se encontrar num estado de equilíbrio estatístico em torno de um nível médio fixo, isto é, o comportamento do processo é governado pela mesma lei de probabilidade ao longo do tempo (Murteira *et al.*, 1993). Se as características do processo forem alteradas ao longo do tempo, então o processo é não estacionário. Além disso, um processo estacionário pode ser classificado como estritamente estacionário ou estacionário de 2.ª ordem.

Definição 2.1.2. *Um processo estocástico* $\{Y_t,\ t \in T\}$ *diz-se estritamente estacionário ou fortemente estacionário se a distribuição conjunta de* $(Y_{t_1}, \ldots, Y_{t_n})$ *for igual à distribuição conjunta de* $(Y_{t_1+\delta}, \ldots, Y_{t_n+\delta})$ *qualquer que seja o n-úplo* $(t_1, \ldots, t_n)$ *e para qualquer* δ, *ou seja,*

$$F_{(Y_{t_1},\ldots,Y_{t_n})}(y_1, \ldots, y_n) = F_{(Y_{t_1+\delta},\ldots,Y_{t_n+\delta})}(y_1, \ldots, y_n),$$

para todos os pontos $(y_1, \ldots, y_n)$.

A definição de estritamente estacionário é demasiado exigente e difícil de se verificar na prática, pois exige o conhecimento de todas as distribuições marginais (Cordeiro, 2003). Para contornar esta dificuldade, é definida a estacionaridade com base na igualdade dos momentos e não na igualdade das distribuições, descrevendo o mesmo tipo de comportamento físico.

Definição 2.1.3. *Um processo* $\{Y_t,\ t \in T\}$ *diz-se estacionário de 2.ª ordem ou fracamente estacionário ou ainda estacionário para a covariância se todos os momentos até à 2.ª ordem de* $(Y_{t_1}, \ldots, Y_{t_n})$ *existem e são iguais aos momentos correspondentes*

até à 2.ª ordem de $(Y_{t_1+\delta}, \ldots, Y_{t_n+\delta})$. *Logo, num processo fracamente estacionário, tem-se*

1. *o valor médio não depende de t, isto é,* $E(Y_t) = \mu_t = \mu, \forall t$;
2. *a variância não depende de t, isto é,* $var(Y_t) = \sigma_t^2 = \sigma^2, \forall t$;
3. *a covariância entre* Y_{t_1} *e* Y_{t_2} *depende apenas do desfasamento* $t_2 - t_1$, *isto é,* $cov(Y_{t_1}, Y_{t_2}) = \gamma(|t_2 - t_1|)$.

Dado que a estacionaridade no sentido estrito é muito forte, apenas se exige a estacionaridade de 2.ª ordem, pelo que, a partir de agora, consideram-se os processos estacionários de 2.ª ordem designados simplesmente por processos estacionários.

Exemplo 2.1.1. *Um exemplo clássico do processo estacionário de 2.ª ordem é o processo de ruído branco. A maior importância deste processo não advém de si, mas da importância que representa na construção de outros processos estacionários relevantes na modelação de séries temporais. Um processo de ruído branco é constituído por uma sucessão de variáveis aleatórias com a mesma distribuição, média constante* $E(Y_t) = \mu$ *(usualmente* $\mu = 0$*), variância constante* $var(Y_t) = \sigma^2$ *e covariância* $cov(Y_{t_1}, Y_{t_2}) = 0$, *para todo* $t_1 \neq t_2$. *Na Figura 2.1 está representado uma trajetória simulada de um ruído branco Gaussiano com média nula e variância unitária de 200 observações.*

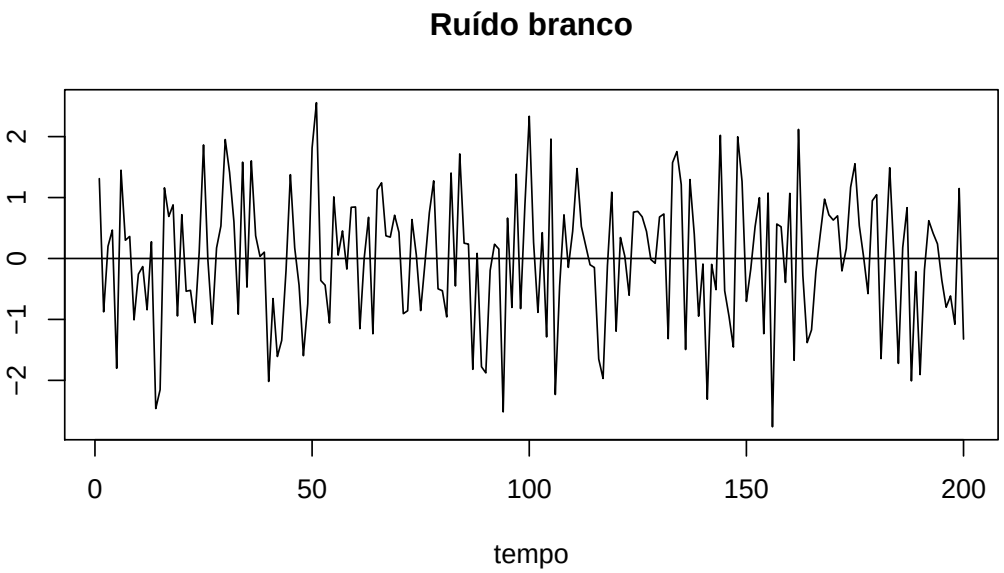


Figura 2.1: Simulação de um ruído branco Gaussiano.

Funções de autocovariância, autocorrelação e autocorrelação parcial

Definição 2.1.4. *Para um processo estacionário* $\{Y_t,\ t \in T\}$ *com valor médio* $E(Y_t) = \mu$ *e* $var(Y_t) = \sigma^2$*, define-se a função de autocovariância, com* $k \in \mathbb{Z}$

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

Para cada valor de k, a função de autocovariância mede a intensidade com que os pares de valores do processo, separados por um intervalo de tempo (*lag*) de amplitude k, se relacionam.

A função de autocovariância satisfaz as propriedades, com $k \in \mathbb{Z}$

i) $\gamma_0 = cov(Y_t, Y_t) = var(Y_t) = \sigma^2$;

ii) $|\gamma_k| \leq \gamma_0$;

iii) $\gamma_k = \gamma_{-k}$, ou seja, a função é par.

Definição 2.1.5. *Para um processo estacionário* $\{Y_t,\ t \in T\}$ *com valor médio* $E(Y_t) = \mu$ *e* $var(Y_t) = \sigma^2$*, define-se a função de autocorrelação (FAC), com* $k = 1, 2, \ldots$

$$\rho_k = corr(Y_t, Y_{t+k}) = \frac{\gamma_k}{\gamma_0} = \frac{cov(Y_t, Y_{t+k})}{var(Y_t)}.$$

Para cada valor de k, a função de autocorrelação quantifica a correlação entre pares de valores do processo separados por um intervalo de tempo de amplitude k. Além disso, dispõe das mesmas propriedades que a função de autocovariância, exceto para a propriedade i), onde $\rho_0 = 1$.

Na generalidade dos casos, o aumento de k leva ao decréscimo de ρ_k e de γ_k. É de esperar que a capacidade de memória do processo seja limitada e, portanto, a capacidade de retenção no momento $t+k$ seja reduzida do que se passou no momento t (Murteira *et al.*, 1993), ou seja, é de se esperar que a capacidade de memória do processo se vá desvanecendo ao longo do tempo. Assim, quando $k \to +\infty$, espera-se que $\rho_k \to 0$.

Quando se tem o interesse em estudar a correlação existente entre duas observações com desfazamento k, Y_t e Y_{t+k}, eliminando a dependência das variáveis intermédias, obtém-se a chamada **função de autocorrelação parcial** (FACP).

Definição 2.1.6. *O conjunto de autocorrelações parciais de desfasamento* k *é dado por* $\{\phi_{kk} : k = 1, 2, \ldots\}$*, onde*

$$\phi_{kk} = corr(X_t, X_{t+k} | X_{t+1}, X_{t+2}, \ldots, X_{t+k-1}) = \frac{|P_k^*|}{|P_k|},$$

e P_k^* *é a matriz quadrada (ordem* k*) de autocorrelações, onde a última coluna é substituída por* $[\rho_1 \ \rho_2 \ \dots \ \rho_k]^T$*. A matriz* P_k *é dada por*

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

A FACP satisfaz as propriedades:

i) $\phi_{11} = \rho_1$;

ii) $\phi_{22} = \dfrac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$;

iii) $\phi_{33} = \dfrac{\rho_3(1-\rho_1^2) + \rho_1(\rho_1^2 + \rho_2^2 - 2\rho_2)}{(1-\rho_2)(1+\rho_2-2\rho_1^2)}$.

Exemplo 2.1.2. *Um exemplo importante de um processo estocástico com tempo discreto é o passeio aleatório (random walk). Uma partícula em movimento com posição inicial* Y_0 *é observada numa sucessão discreta de pontos* $t = 0, 1, 2, \dots$*. No instante* $t = 1$*, a partícula apresenta um salto de amplitude* ε_1*, onde* ε_1 *é uma variável aleatória com uma dada função de distribuição. No instante* $t = 2$*, verifica-se novo salto de amplitude* ε_2*, onde* ε_2 *é uma variável aleatória não correlacionada com* ε_1*, mas com a mesma função de distribuição. E assim por diante. No instante* t*, a posição da partícula é dada por,*

$$Y_t = Y_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t \quad ou \quad Y_t = Y_{t-1} + \varepsilon_t,$$

onde ε_t*,* $t = 1, 2, 3, \dots$ *é uma sequência de variáveis aleatórias não correlacionadas (Murteira et al., 1993).*
Na Figura 2.2 estão representadas três simulações de um passeio aleatório com 200 observações, onde $Y_0 = 0$ *e* ε_t*,* $t = 1, 2, 3, \dots$ *variáveis aleatórias não correlacionadas onde* $\varepsilon_t \sim N(0, 1)$*.*

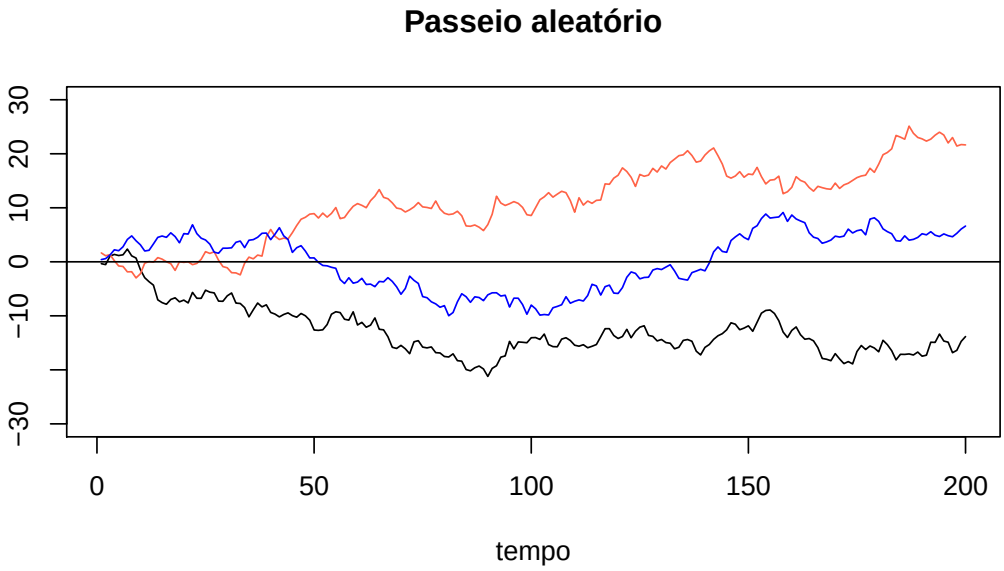


Figura 2.2: Simulação de três trajetórias do passeio aleatório.

2.2 Modelos Lineares para Séries Temporais Estacionárias

Na análise de séries temporais, a ideia de seleção de um modelo é de extrema importância pois é com base na estrutura probabilística, que reflete a estrutura subjacente à série temporal (processo estocástico), que se fazem previsões.

Um modelo de série temporal pode ser designado por linear ou não linear, dependendo se o valor atual da série é função linear das observações passadas ou não.

2.2.1 Processo Autorregressivo de Ordem p ($AR(p)$)

Definição 2.2.1. *Um processo $\{Y_t, t \in T\}$ diz-se ser um processo autorregressivo de ordem p, $AR(p)$, se satisfaz a equação*

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \tag{2.1}$$

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

Usando o operador atraso, B, a equação (2.1) pode ser reescrita

$$\Phi_p(B) Y_t = \varepsilon_t,$$

onde ε_t é um ruído branco e $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ é designado por **polinómio autorregressivo** de ordem p. O número de termos p define a ordem do processo autorregressivo (Cordeiro, 2003). Tendo em consideração as p raízes reais ou complexas, $G_1^{-1}, G_2^{-1}, \ldots, G_p^{-1}$, da equação característica $\Phi_p(B) = 0$, torna-se possível fatorizar o polinómio autorregressivo, que é dado por

$$\Phi_p(B) = \prod_{i=1}^{p}(1 - G_i B).$$

Condições de invertibilidade e estacionaridade

No geral, diz-se que um processo é invertível se admitir a representação autorregressiva, ou seja

$$\varepsilon_t = Y_t + \sum_{i=1}^{\infty} \phi_i Y_{t-i}.$$

Portanto, os processos autorregressivos são sempre invertíveis. Para que sejam estacionários, as raízes do polinómio autorregressivo têm que estar fora do círculo unitário, ou seja, que $|G_i| < 1$, para $i = 1, 2, \ldots, p$.

Função de autocovariância

Atendendo à definição de função de autovariância, tem-se

$$\gamma_k = E(Y_t Y_{t+k}) = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p}, \ k \geq 1.$$

Função de autocorrelação (FAC)

Dividindo a função de autocovariância por $\gamma_0 = \sigma^2$, obtém-se a expressão para a função de autocorrelação

$$\rho_k = \gamma_k/\gamma_0 = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}, \ k \geq 1.$$

Função de autocorrelação parcial (FACP)

Tento em conta a definição de função de autocorrelação parcial, a FACP do processo $AR(p)$ é dada por

$$\phi_{kk} = \begin{cases} \dfrac{|P_k^*|}{|P_k|}, & k = 1, 2, \ldots, p \\ 0, & k = p+1, p+2, \ldots \end{cases}.$$

Particularmente, no caso do processo autorregressivo de ordem 1, $AR(1)$, tem-se

$$Y_t = \phi Y_{t-1} + \varepsilon_t \text{ ou } (1 - \phi B)Y_t = \varepsilon_t,$$

onde ϕ é um número real e ε_t é um ruído branco. Além disso,

- Os processos $AR(1)$ são sempre invertíveis;
- Para que o processo seja estacionário, a variância de Y_t tem de se manter constante para qualquer t e, portanto, $|\phi| < 1$ e, consequentemente $\mu = E(Y_t) = 0$;
- Sendo o processo estacionário, a variância de Y_t é dada por $var(Y_t) = \dfrac{\sigma_\varepsilon^2}{1 - \phi^2}$;
- Sendo o processo estacionário, a função de autocovariância é dada por

$$\gamma_k = \frac{\phi^k \sigma_\varepsilon^2}{1 - \phi^2}, \quad k \geq 1;$$

- A função de autocorrelação (FAC) é dada por

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^k;$$

- Ambas as funções de autocovariância e FAC decrescem exponencialmente para zero com o aumento de k. Quando $\phi > 0$, as autocorrelações são sempre positivas e quando $\phi < 0$, estas alternam de sinal, começando com uma autocorrelação de primeira ordem negativa.

2.2.2 Processo de Médias Móveis de Ordem q $(MA(q))$

Definição 2.2.2. *O processo $\{Y_t,\ t \in T\}$ diz-se ser um processo de médias móveis de ordem q, MA(q) se satisfaz a equação*

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \tag{2.2}$$

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

Usando o operador atraso, B, a equação (2.2) pode ser reescrita

$$Y_t = \Theta_q(B)\varepsilon_t,$$

onde ε_t é um ruído branco e $\Theta_q(B) = 1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q$ é o **polinómio de médias móveis** de ordem q. Um processo de médias móveis de ordem q define-se em cada instante t, como a média ponderada das $q+1$ observações de um processo de ruído branco (Cordeiro, 2003).

Condições de invertibilidade e estacionaridade

Os processos de médias móveis são sempre estacionários. Além disso, o processo é invertível se a equação $\Theta_q(B) = 0$ tiver todas as suas raízes fora do círculo unitário.

Função de autocovariância

Tento em conta que $E(Y_t) = 0$ e variância $\gamma_0 = \left(1+\theta_1^2+\cdots+\theta_q^2\right)\sigma_\varepsilon^2$, a função de autocovariância é dada por

$$\gamma_k = \begin{cases} \sigma_\varepsilon^2\left(-\theta_k+\theta_{k+1}\theta_1+\cdots+\theta_q\theta_{q-k}\right), & 0<k\leq q \\ 0, & k>q. \end{cases}$$

Função de autocorrelação (FAC)

Para a FAC tem-se

$$\rho_k = \begin{cases} \dfrac{-\theta_k+\theta_{k+1}\theta_1+\cdots+\theta_q\theta_{q-k}}{1+\theta_1^2+\cdots+\theta_q^2}, & 0<k\leq q \\ 0, & k>q. \end{cases}$$

É de notar a queda brusca da FAC para $k \geq q+1$.

Função de autocorrelação parcial (FACP)

A FACP tem uma expressão complicada. Contudo, pode verificar-se que é majorada pela soma de duas exponenciais amortecidas quando as raízes do polinómio de médias móveis são reais e por uma sinusoide amortecida quando são complexas (Murteira *et al.*, 1993). No entanto, a FACP de um processo $MA(q)$ tem a mesma estrutura que a FAC de um processo $AR(p)$, que decai gradualmente para zero. Além disso, a FACP de um processo $AR(p)$ comporta-se como a FAC dos processos $MA(q)$, que decai bruscamente para zero. Esta correspondência traduz a dualidade existente entre os dois tipos de processos.

2.2.3 Processos Autorregressivos e de Médias Móveis ($ARMA(p,q)$)

Os processos autorregressivos e de médias móveis são modelos que combinam as duas representações: o processo autorregressivo e o processo de médias móveis.

Definição 2.2.3. *Um processo estacionário $\{Y_t, t \in T\}$ diz-se ser um processo autorregressivo e de médias móveis de ordens p e q, ARMA(p,q), se satisfaz a equação*

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \tag{2.3}$$

onde ε_t é um processo de ruído branco e $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ são constantes reais.

Usando o operador atraso, B, a equação (2.3) pode ser reescrita

$$\Phi_p(B) Y_t = \Theta_q(B) \varepsilon_t,$$

onde ε_t é um ruído branco, independente de Y_{t-k} para todo o $k \geq 1$, $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ é o polinómio autorregressivo de ordem p e $\Theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ é polinómio de médias móveis de ordem q.

Condições de invertibilidade e estacionaridade

Para que o processo seja estacionário, as raízes de $\Phi_p(B) = 0$ devem estar fora do círculo unitário. Para que o processo seja invertível, as raízes de $\Theta_q(B) = 0$ devem estar fora do círculo unitário.

Função de autocovariância

multiplicando a equação (2.3) por Y_{t-k} e tomando os valores esperados, vem

$$\gamma_k = \phi_1 \gamma_{k-1} + \cdots + \phi_p \gamma_{k-p}, \quad k \geq q + 1.$$

Função de autocorrelação (FAC)

Dividindo a função de autocovariância por $\gamma_0 = \sigma^2$, obtém-se a expressão para a FAC

$$\rho_k = \phi_1 \rho_{k-1} + \cdots + \phi_p \rho_{k-p}, \quad k \geq q + 1.$$

Função de autocorrelação parcial (FACP)

A FACP apresenta um decaimento gradual para zero majorado por uma soma de exponenciais e/ou sinusoides amortecidas, tendo um comportamento muito semelhante ao da FACP dos processos de médias móveis (Murteira *et al.*, 1993).

Na Tabela 2.1 encontram-se esquematizadas as propriedades preliminares para a identificação dos vários tipos de processos $AR(p)$, $MA(q)$ e $ARMA(p,q)$.

Tabela 2.1: Comparação da FAC e FACP dos vários processos $ARMA(p,q)$, adaptado de Murteira *et al.* (1993).

Processo	**FAC**	**FACP**
AR(p)	Decaimento exponencial e/ou sinusoidal para zero	Queda brusca para zero a partir de $k = p + 1$
MA(q)	Queda brusca para zero a partir de $k = p + 1$	Decaimento exponencial e/ou sinusoidal para zero
ARMA(p, q)	Decaimento exponencial e/ou sinusoidal para zero	Decaimento exponencial e/ou sinusoidal para zero

Capítulo 3

Modelos de Espaço de Estados

Os MEE são modelos com uma estrutura bastante flexível que permitem incorporar componentes não observadas de forma estocástica. Os modelos de regressão linear são um caso particular dos MEE, cujos parâmetros consideram-se determinísticos, ou seja, são invariantes no tempo.

Modelo de Regressão Linear Simples

Para que seja possível fazer previsões sobre uma variável a partir de uma ou mais covariáveis, é necessário que exista uma relação de causa-efeito, isto é, que a variação de uma possa ser explicada à custa da variação das outras (Reis, 2008).

Existem vários tipos de relação entre variáveis: linear, exponencial, logarítmica, etc. Para estudar o tipo de relação funcional existente, deve começar-se por fazer um diagrama de dispersão dos dados. Se essa relação for do tipo linear, os pontos devem encontram-se dispersos aleatoriamente em torno de uma reta.

No caso da RLS, o modelo traduz a relação afim entre a variável resposta, Y_t, e a variável explicativa, X_t, que pode ser descrita matematicamente através da seguinte equação

$$Y_t = \beta_0 + \beta_1 X_t + e_t, \quad t = 1, \ldots, n, \tag{3.1}$$

onde

- a variável aleatória Y_t representa a variável resposta ou dependente;
- a variável aleatória X_t representa a variável explicativa, preditor ou independente;
- β_0 e β_1 chamam-se coeficientes (parâmetros) de regressão: β_0 é a interseção da

reta com o eixo vertical e representa o valor esperado da variável Y_t quando a variável explicativa é nula e β_1 é o declive da reta e representa a variação do valor esperado de Y_t por cada incremento unitário na variável explicativa. Os coeficientes de regressão são constantes e, por regra, desconhecidos;

– e_t é a variável residual que inclui outros fatores explicativos de Y_t não incluídos em X_t e ainda erros de medição. É uma variável aleatória não observável porque depende dos coeficientes de regressão.

Modelo de regressão linear múltipla

Por vezes a variável resposta, Y_t, é frequentemente influenciada por mais do que uma variável explicativa. Por exemplo, o rendimento de uma colheita pode depender da quantidade de fertilizantes nitrogenados, fosfatados e potássicos utilizados. Estas variáveis são controladas pelo pesquisador, mas o rendimento também pode depender de outras variáveis que não são controláveis, como as associadas ao clima. Quando existem duas ou mais variáveis explicativas, o modelo é chamado de **modelo de regressão linear múltipla** e é dado por

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \cdots + \beta_p X_{t,p} + e_t, \quad t = 1, \ldots, n. \tag{3.2}$$

Como em (3.1), e_t representa a variação de Y que não é explicada pelas variáveis $X_{t,i}$ $(i = 1, \ldots, p)$. O coeficiente β_0 representa o valor esperado de Y_t quando as variáveis explicativas são simultaneamente nulas e β_i $(i = 1, \ldots, p)$ representa a variação do valor esperado de Y_t por cada incremento unitário na variável X_i, quando as restantes variáveis explicativas se mantêm constantes. É de realçar que o modelo em (3.2) é linear nos parâmetros β e não é necessariamente linear nas variáveis explicativas $X_{t,i}$ $(t = 1, 2, \ldots, n$ e $i = 1, \ldots, p)$.

As n igualdades deste modelo também podem apresentar-se sob a notação matricial. Com efeito, fazendo

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

obtém-se

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

onde $\boldsymbol{Y}$ é o vetor $n \times 1$ das variáveis resposta, $\boldsymbol{X}$ é a matriz $n \times (p+1)$ das observações das variáveis explicativas. Como o modelo tem um termo independente, β_0, a primeira coluna de $\boldsymbol{X}$ é constituída por uns. O vetor $\boldsymbol{\beta}$ tem dimensão $(p+1)\times 1$ e contém os coeficientes da regressão e $\boldsymbol{e}$ é o vetor de dimensão $n \times 1$ dos erros aleatórios.

Qualidade de ajustamento

Para avaliar o ajustamento do modelo aos dados, existem medidas do grau de associação ou correlação entre os Y_t e os $\hat{Y}_t$ $(t = 1, \ldots, n)$: o coeficiente de correlação amostral e o coeficiente de determinação.

O **coeficiente de correlação amostral** (também conhecido por coeficiente de correlação de Pearson) é dado por

$$r = \frac{\sum_{t=1}^{n} \left(Y_t - \overline{Y}\right)\left(\hat{Y}_t - \overline{\hat{Y}}\right)}{\sqrt{\sum_{t=1}^{n} \left(Y_t - \overline{Y}\right)^2 \sum_{t=1}^{n} \left(\hat{Y}_t - \overline{\hat{Y}}\right)^2}},$$

onde $\overline{Y}$ e $\overline{\hat{Y}}$ são as médias dos Y_t e dos $\hat{Y}$, respetivamente.

Este coeficiente mede o grau de associação linear entre duas variáveis e toma valores entre -1 e 1, sendo que 1 significa uma relação linear perfeita e positiva, enquanto que -1 é também uma relação linear perfeita mas negativa. Quando $r = 0$ significa apenas a ausência da relação linear entre as duas variáveis, podendo existir uma relação não linear entre as variáveis.

Outra medida do grau de relação linear entre duas variáveis, definida como sendo a média do produto das diferenças entre os valores de cada variável e a respetiva média, é a **covariância**, dada por

$$cov(Y, \hat{Y}) = \frac{1}{n-1} \sum_{t=1}^{n} \left(Y_t - \overline{Y}\right)\left(\hat{Y}_t - \overline{\hat{Y}}\right).$$

A grande desvantagem na utilização da covariância como medida de associação linear entre duas variáveis comparativamente ao coeficiente de correlação amostral é o facto de depender da unidade de medida das variáveis. Além disso, como o coeficiente de correlação amostral possui limites bem definidos, torna possível distinguir entre graus de associação elevados ou reduzidos.

O **coeficiente de determinação** é definido como sendo o quadrado do coefici-

ente de correlação amostral entre os Y_t e os $\hat{Y}_t$ $(t = 1, \ldots, n)$ dado por

$$r^2 = \frac{\left(\sum_{t=1}^{n} \left(Y_t - \overline{Y}\right)\left(\hat{Y}_t - \overline{\hat{Y}}\right)\right)^2}{\sum_{t=1}^{n} \left(Y_t - \overline{Y}\right)^2 \sum_{t=1}^{n} \left(\hat{Y}_t - \overline{\hat{Y}}\right)^2}.$$

Como $0 \leq r^2 \leq 1$, quando $r^2 \simeq 1$, melhor é o ajustamento.

No entanto, esta medida apresenta um inconveniente: quando se acrescenta ao modelo mais uma variável explicativa, qualquer que ela seja, o r^2 nunca decresce (para a mesma amostra) (Murteira *et al.*, 2015). O coeficiente de determinação pode ser utilizado para comparar modelos com o mesmo grau de complexidade, mas para comparar modelos com graus de complexidade distintos, deve utilizar-se o **coeficiente de determinação ajustado** dado por

$$r_a^2 = 1 - \frac{n-1}{n-p-1}(1 - r^2),$$

onde r^2 é o coeficiente de determinação, n é o número de observações e p é o número de variáveis explicativas.

Pressupostos do modelo de regressão linear

Nos modelos de regressão linear existem alguns pressupostos que se devem ter em conta

1. os erros aleatórios devem seguir uma distribuição aproximadamente Normal, ou seja, $e \sim N(0, \sigma^2)$;

2. o valor esperado dos erros aleatórios deve ser nulo, ou seja, $E(e_t) = 0,\ \forall t$;

3. a variância dos erros aleatórios deve ser constante ao longo do tempo (homocedasticidade), ou seja, $var(e_t) = \sigma^2, \forall t$ (variância constante e desconhecida);

4. Os erros aleatórios não devem ser autocorrelacionados, isto é, $cov(e_t, e_j) = 0$, para todo $t \neq j$.

3.1 Modelo de Espaço de Estados

Os modelos baseados na representação de espaço de estados apresentam uma vasta aplicabilidade nos problemas da atualidade e podem ser encontrados em diversas áreas tais como na Economia, Biologia, Engenharia de Controlo, *Machine Learning*, Processamento de Sinais, Análise de Séries Temporais, etc.

Historicamente, os MEE foram desenvolvidos na engenharia no início dos anos 60 com o objetivo de monitorizar e controlar os sistemas dinâmicos. As primeiras aplicações mais conhecidas foram nos programas aeroespaciais Apollo e Polaris, sendo amplamente desenvolvidos nos anos 70-80 (Petris *et al.*, 2009).

A formulação de espaço de estados assume que o desenvolvimento de um sistema ao longo do tempo é determinado por um processo de vetores $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_n}$ (estados), não observados, associados linearmente ou não a uma série de observações $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_n}$. Do ponto de vista da análise de séries temporais, a representação de espaço de estados formaliza a relação temporal entre as variáveis que são observadas, variáveis não observáveis (estados) e erros estocásticos (Costa, 2006).

Os MEE são modelos bastante flexíveis devido à sua capacidade de integrar várias caraterísticas dos dados que permitem analisar fenómenos dinâmicos que variam de forma significativa ao longo do tempo. Segundo Petris *et al.* (2009), são muito mais flexíveis do que os modelos ARMA no tratamento de séries temporais não estacionárias e na modelação de mudanças estruturais e, geralmente, são mais fáceis de interpretar.

Uma das principais mais-valias dos modelos com a representação de espaço de estados é a possibilidade de fazer-se inferências sobre os estados não observados e fazer previsões de futuras observações com base nos dados disponíveis. Uma das grandes vantagens destes modelos baseia-se na capacidade de atualizar de forma recursiva e em tempo real uma série temporal à medida que novas observações vão sendo disponibilizadas, melhorando as previsões fornecidas. Tendo em conta a natureza dinâmica dos dados meteorológicos, estes modelos conseguem captar de forma estocástica tal comportamento, devido à sua versatilidade e flexibilidade. Além disso, uma característica muito comum neste tipo de dados é a existência de valores em falta (NA), na qual os MEE lidam de forma particularmente simples com este tipo de lacunas.

No geral, o modelo linear Gaussiano de espaço de estados é caracterizado por duas equações: a **equação de observação** e a **equação de estado** (ou **equação de transição**). A equação de observação relaciona o vetor p-dimensional das variáveis observáveis $\boldsymbol{Y_t}$ com o vetor $\boldsymbol{\beta_t}$ não observável $m \times 1$, designado por **vetor de estados**.

A equação de observação é dada por

$$\boldsymbol{Y_t} = \boldsymbol{W_t}\boldsymbol{\beta_t} + \boldsymbol{e_t}, \quad t = 1, \ldots, n, \tag{3.3}$$

onde $\boldsymbol{W_t}$ é uma matriz $p \times m$ assumida como conhecida e $\boldsymbol{e_t}$ é um vetor $p \times 1$ dos erros independentes e identicamente distribuídos com distribuição Normal de média zero e matriz de covariância $\boldsymbol{H}$, ou seja

$$\boldsymbol{e_t} \sim N(\mathbf{0}, \boldsymbol{H}) \;\text{ e }\; E(\boldsymbol{e_t}\boldsymbol{e_s'}) = \mathbf{0}, \;\text{ para todo }\; t \neq s. \tag{3.4}$$

A equação de estado descreve a dinâmica do vetor de estados $\boldsymbol{\beta_t}$ que, embora não sejam observados, os elementos de $\boldsymbol{\beta_t}$ são atualizados no tempo segundo um processo autorregressivo de ordem um. A equação de estado é dada por

$$\boldsymbol{\beta_t} = \boldsymbol{\Phi}\boldsymbol{\beta_{t-1}} + \boldsymbol{\varepsilon_t}, \quad t = 1, \ldots, n, \tag{3.5}$$

onde $\boldsymbol{\Phi}$ é uma matriz $m \times m$ designada por **matriz autorregressiva** ou **matriz de transição** e $\boldsymbol{\varepsilon_t}$ é um vetor $m \times 1$ dos erros independentes e identicamente distribuídos com distribuição Normal de média zero e matriz de covariância $\boldsymbol{Q}$, ou seja

$$\boldsymbol{\varepsilon_t} \sim N(\mathbf{0}, \boldsymbol{Q}) \;\text{ e }\; E(\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_s'}) = \mathbf{0}, \;\text{ para todo }\; t \neq s. \tag{3.6}$$

As matrizes $\boldsymbol{W_t}$ e $\boldsymbol{H}$ da equação de observação (3.3) e as matrizes $\boldsymbol{\Phi}$ e $\boldsymbol{Q}$ da equação de estado (3.5) chamam-se **matrizes do sistema** e são inicialmente assumidas como conhecidas.

Nesta abordagem, existem duas assunções importantes a ter em conta

- o vetor de estados inicial, $\boldsymbol{\beta_1}$, tem distribuição Normal com média $\boldsymbol{a_1}$ e matriz de covariância $\boldsymbol{P_1}$, ou seja, $\boldsymbol{\beta_1} \sim N(\boldsymbol{a_1}, \boldsymbol{P_1})$;

- os erros $\boldsymbol{e_t}$ e $\boldsymbol{\varepsilon_t}$ são não correlacionados entre si e também são não correlacionados com o vetor de estados inicial, ou seja

$$E(\boldsymbol{e_t}\boldsymbol{\varepsilon_s'}) = \mathbf{0}, \;\; E(\boldsymbol{e_t}\boldsymbol{\beta_1'}) = \mathbf{0} \;\text{ e }\; E(\boldsymbol{\varepsilon_t}\boldsymbol{\beta_1'}) = \mathbf{0}, \;\text{ para todo }\; t, s = 1, \ldots, n. \tag{3.7}$$

Na prática, algumas ou até mesmo todas as matrizes $\boldsymbol{W_t}$, $\boldsymbol{\Phi}$, $\boldsymbol{H}$ e $\boldsymbol{Q}$ podem depender de um conjunto de parâmetros desconhecidos (Durbin e Koopman, 2001). Estes parâmetros são representados pelo vetor $\boldsymbol{\Theta}$. Nos MEE, a estrutura e dependência de (3.3) e (3.5) mantém-se, exceto as suposições de linearidade e normalidade,

que não são exigidas.

As vantagens da formulação de espaço de estados reside na facilidade com que é possível tratar várias configurações de dados em falta e na variedade de modelos que podem ser gerados a partir de (3.3) e (3.5). É possível gerar vários modelos com estruturas de efeitos fixos ou aleatórios, ou seja, que são constantes ou que variam ao longo do tempo fazendo escolhas apropriadas para a matriz $\boldsymbol{W_t}$ e para a estrutura de transição $\boldsymbol{\Phi}$ (Shumway e Stoffer, 2017).

3.2 Modelos Estruturais

Os modelos estruturais são uma classe geral de modelos que lidam com a tendência e a sazonalidade e podem ser representados na forma de espaço de estados. Nesta abordagem, são construídos submodelos para modelar cada componente não observável separadamente e depois são colocamos juntos para formar um só modelo. Os componentes não observáveis que ocorrem com maior frequência numa série temporal são:

- tendência (μ_t);
- componente sazonal (γ_t);
- componente cíclica (δ_t);
- componente aleatória ou erro (e_t).

Uma série temporal univariada $\{Y_t\}$, $t = 1, 2, \ldots, n$ pode ser decomposta de acordo com as suas componentes não observáveis da seguinte forma

$$Y_t = \mu_t + \gamma_t + \delta_t + e_t, \tag{3.8}$$

onde $e_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2)$.

A partir do modelo (3.8), derivam inúmeros modelos estruturais (Commandeur e Koopman, 2007; Baturin, 2016).

Exemplo 3.2.1. *O modelo de nível local (ou modelo linear dinâmico de primeira ordem ou passeio aleatório mais um erro) é um dos modelos estruturais mais simples, na qual considera o modelo (3.8), tomando $\mu_t = \alpha_t$, onde α_t é um passeio aleatório, não incorpora a componente sazonal e todas as variáveis aleatórias são normalmente distribuídas. Além disso, assume que o erro e_t tem variância constante σ_e^2. O modelo*

de nível local é dado pelo sistema de equações

$$\begin{aligned} Y_t &= \alpha_t + e_t, \quad e_t \sim N(0, \sigma_e^2) \\ \alpha_{t+1} &= \alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \end{aligned}$$

para $t = 1, \ldots, n$, onde α_t é o estado não observável, e_t é o erro da equação de observação e ε_t é o erro da equação de estado. Assume-se que e_t e ε_t são independentes.

Exemplo 3.2.2. *O modelo de tendência linear local (ou modelo linear dinâmico de segunda ordem) é obtido adicionando uma componente de tendência v_t à equação de estado, que é dado por*

$$\begin{aligned} Y_t &= \mu_t + e_t, \quad e_t \sim N(0, \sigma_e^2) \\ \mu_{t+1} &= \mu_t + v_t + \xi_t \quad \xi_t \sim N(0, \sigma_\xi^2) \\ v_{t+1} &= v_t + \zeta_t, \quad \zeta_t \sim N(0, \sigma_\zeta^2), \end{aligned}$$

para $t = 1, \ldots, n$, onde assume-se que os erros e_t, ξ_t e ζ_t são não correlacionados.

Exemplo 3.2.3. *O modelo de nível local com variável explicativa é obtido adicionando variáveis explicativas na equação de observação, caso seja do interesse estudar o efeito que outras variáveis possam ter sobre a série temporal em estudo. O modelo é dado por*

$$\begin{aligned} Y_t &= \mu_t + \beta_t X_t + e_t, \quad e_t \sim N(0, \sigma_e^2) \\ \mu_{t+1} &= \mu_t + \xi_t \quad \xi_t \sim N(0, \sigma_\xi^2) \\ \beta_{t+1} &= \beta_t + \tau_t, \quad \tau_t \sim N(0, \sigma_\tau^2), \end{aligned}$$

para $t = 1, \ldots, n$, onde assume-se que os erros e_t, ξ_t e τ_t são não correlacionados.

Capítulo 4

Filtragem, Alisamento e Previsão de Kalman

Nos MEE, o vetor de estados $\boldsymbol{\beta_t}$ não é observado e, portanto, deve ser estimado. Neste capítulo apresenta-se três problemas fundamentais associados ao MEE definido pelas equações

$$\boldsymbol{Y_t} = \boldsymbol{W_t}\boldsymbol{\beta_t} + \boldsymbol{e_t}, \quad \boldsymbol{e_t} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \boldsymbol{H}) \tag{4.1}$$

$$\boldsymbol{\beta_t} = \boldsymbol{\Phi}\boldsymbol{\beta_{t-1}} + \boldsymbol{\varepsilon_t}, \quad \boldsymbol{\varepsilon_t} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \boldsymbol{Q}) \tag{4.2}$$

para $t = 1, \ldots, n$. O objetivo é encontrar o melhor estimador linear de $\boldsymbol{\beta_t}$, no sentido em que possui o menor erro quadrático médio, com base das observações $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots$. A estimação de $\boldsymbol{\beta_t}$ condicional ao conhecimento de

1. $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_{t-1}}$ define o **problema de previsão**;

2. $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_t}$ define o **problema de filtragem**;

3. $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_n}$ $(n > t)$ define o **problema de alisamento**.

Cada um desses três problemas podem ser resolvidos recursivamente usando um conjunto apropriado de equações recursivas de Kalman (Brockwell e Davis, 1998).

4.1 Filtro de Kalman

O FK, desenvolvido por Kalman (1960) e Kalman e Bucy (1961), é um algoritmo recursivo que permite obter estimadores ótimos do vetor de estados $\boldsymbol{\beta_t}$, com base

na informação disponível até ao instante t. Pode ser aplicado a processos aleatórios estacionários e não estacionários.

Em certas aplicações na área da engenharia, o FK tem um papel extremamente relevante. Por exemplo, Harvey (1992) refere que, através deste algoritmo recursivo, é possível representar as coordenadas de um foguetão uma vez que permite obter as estimativas para o vetor de estados, sendo atualizado à medida que novas observações são disponibilizadas.

O FK aplica-se a modelos que admitem uma representação de espaço de estados e, além disso, permite obter estimadores ótimos quando o modelo se encontra totalmente especificado (isto é, quando todos os parâmetros são conhecidos), onde as variáveis se relacionam linearmente e os erros são ruídos brancos Gaussianos (Costa, 2006). Ainda assim, segundo Harvey (1992), mesmo quando o pressuposto da normalidade não é cumprido, o FK continua a fornecer estimadores ótimos dentro da classe de todos os estimadores lineares.

O objetivo do FK consiste em obter predições a 1-passo, tanto do vetor de estados como do vetor das variáveis observáveis, e atualizar o vetor de estados cada vez que é conhecida uma nova observação no instante $t-1$, ou seja, pretende-se obter a distribuição condicional do vetor de estados $\boldsymbol{\beta_t}$ $(t = 1, \ldots, n)$ com base nas observações disponíveis até ao instante $t-1$.

Partindo do pressuposto de que todas as distribuições consideradas no sistema são Normais e utilizando os resultados da distribuição Normal multivariada, segue que as distribuições condicionais e marginais também são Normais.

Considerando o modelo Gaussiano de espaço de estados definido em (4.1) e (4.2), seja $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_{t-1}})'$ o vetor das observações e $\boldsymbol{\beta_{t|t-1}}$ o estimador ótimo de $\boldsymbol{\beta_t}$ baseado na informação disponível até ao instante t, isto é,

$$\boldsymbol{\beta_{t|t-1}} = E[\boldsymbol{\beta_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}],$$

e $\boldsymbol{P_{t|t-1}}$ uma matriz $m \times m$ que representa a matriz de covariâncias do seu erro estimado, isto é

$$\boldsymbol{P_{t|t-1}} = E[(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t-1}})(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t-1}})'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}].$$

Considere ainda

$$\boldsymbol{\beta_{t|t}} = E[\boldsymbol{\beta_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \tag{4.3}$$

e

$$\boldsymbol{P_{t|t}} = E[(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}], \tag{4.4}$$

onde $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_t})'$.

Partindo das condições iniciais $\boldsymbol{\beta_{1|0}} = \boldsymbol{a_1}$ e $\boldsymbol{P_{1|0}} = \boldsymbol{P_1}$, das definições (4.3) e (4.4), a equação de previsão de $\boldsymbol{Y_t}$ é dada por

$$\boldsymbol{Y_{t|t-1}} = E[\boldsymbol{Y_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}] = \boldsymbol{W_t}\boldsymbol{\beta_{t|t-1}},$$

e, quando a observação $\boldsymbol{Y_t}$ fica disponível, a atualização da previsão de $\boldsymbol{\beta_t}$ e do respetivo EQM é dada por

$$\boldsymbol{\beta_{t|t}} = \boldsymbol{\beta_{t|t-1}} + \boldsymbol{K_t}(\boldsymbol{Y_t} - \boldsymbol{Y_{t|t-1}}), \tag{4.5}$$

$$\boldsymbol{P_{t|t}} = [\boldsymbol{I} - \boldsymbol{K_t}\boldsymbol{W_t}]\boldsymbol{P_{t|t-1}}, \tag{4.6}$$

onde

$$\boldsymbol{K_t} = \boldsymbol{P_{t|t-1}}\boldsymbol{W_t'}[\boldsymbol{W_t}\boldsymbol{P_{t|t-1}}\boldsymbol{W_t'} + \boldsymbol{H}]^{-1}, \quad t = 1, \ldots, n \tag{4.7}$$

é chamado de **ganho de Kalman** que é uma matriz $m \times p$ e $\boldsymbol{I}$ é a matriz identidade de ordem m. As equações (4.5) e (4.6) são conhecidas como **equações de atualização**.

A partir da equação de estado (4.2), é possível obter uma previsão para o vetor de estados no instante $t+1$, dada toda a informação disponível até ao instante t, da forma

$$\boldsymbol{\beta_{t+1|t}} = \boldsymbol{\Phi}\boldsymbol{\beta_{t|t}}, \tag{4.8}$$

$$\boldsymbol{P_{t+1|t}} = \boldsymbol{\Phi}\boldsymbol{P_{t|t}}\boldsymbol{\Phi'} + \boldsymbol{Q}. \tag{4.9}$$

As equações (4.8) e (4.9) são designadas por **equações preditivas** e, no geral, quando se pretende prever para $t > n$, estas equações são utilizadas com as condições iniciais $\boldsymbol{\beta_{n|n}}$ e $\boldsymbol{P_{n|n}}$.

As equações (4.8) e (4.9) são fáceis de demonstrar. Usando a definição de (4.3), a equação de estados (4.2), a condição (3.6) referida no Capítulo 3 e atendendo a que $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_t})'$, tem-se

$$\begin{aligned}
\boldsymbol{\beta_{t+1|t}} &= E[\boldsymbol{\beta_{t+1}}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= E[\boldsymbol{\Phi}\boldsymbol{\beta_t} + \boldsymbol{\varepsilon_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= \boldsymbol{\Phi}\boldsymbol{\beta_{t|t}} + E[\boldsymbol{\varepsilon_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= \boldsymbol{\Phi}\boldsymbol{\beta_{t|t}},
\end{aligned}$$

e utilizando a definição (4.4), tem-se

$$\begin{aligned}
\boldsymbol{P_{t+1|t}} &= E[(\boldsymbol{\beta_{t+1}} - \boldsymbol{\beta_{t+1|t}})(\boldsymbol{\beta_{t+1}} - \boldsymbol{\beta_{t+1|t}})'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= E[(\boldsymbol{\Phi\beta_t} + \boldsymbol{\varepsilon_t} - \boldsymbol{\Phi\beta_{t|t}})(\boldsymbol{\Phi\beta_t} + \boldsymbol{\varepsilon_t} - \boldsymbol{\Phi\beta_{t|t}})'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= E[(\boldsymbol{\Phi}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}}) + \boldsymbol{\varepsilon_t})((\boldsymbol{\Phi}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}}) + \boldsymbol{\varepsilon_t})'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= E[(\boldsymbol{\Phi}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}}) + \boldsymbol{\varepsilon_t})((\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})'\boldsymbol{\Phi}' + \boldsymbol{\varepsilon}_{\boldsymbol{t}}')|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= E[\boldsymbol{\Phi}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})'\boldsymbol{\Phi}' + \boldsymbol{\Phi}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})\boldsymbol{\varepsilon}_{\boldsymbol{t}}' + \\
&+ \boldsymbol{\varepsilon_t}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})'\boldsymbol{\Phi}' + \boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon}_{\boldsymbol{t}}'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}] \\
&= \boldsymbol{\Phi}E[(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}})'|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}]\boldsymbol{\Phi}' + \\
&+ \boldsymbol{\Phi}E[\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}]E[\boldsymbol{\varepsilon}_{\boldsymbol{t}}'] + \\
&+ E[\boldsymbol{\varepsilon_t}]E[\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t}}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}]'\boldsymbol{\Phi}' + E[\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon}_{\boldsymbol{t}}'] \\
&= \boldsymbol{\Phi P_{t|t}\Phi'} + \boldsymbol{Q},
\end{aligned}$$

para $t = 1, \ldots, n$. A verificação das equações (4.5), (4.6) e (4.7) requerem alguns resultados relativos à distribuição Normal multivariada, cujas demonstrações podem ser consultadas em Shumway e Stoffer (2017).

Outros conceitos muito importantes no FK são as **inovações** (erros de previsão)

$$\boldsymbol{\eta_{t|t-1}} = \boldsymbol{Y_t} - E(\boldsymbol{Y_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}) = \boldsymbol{Y_t} - \boldsymbol{W_t\beta_{t|t-1}} = \boldsymbol{Y_t} - \boldsymbol{Y_{t|t-1}},$$

e a respetiva matriz de covariâncias, dada por

$$\boldsymbol{\Sigma_{t|t-1}} = var(\boldsymbol{\eta_{t|t-1}}) = var[\boldsymbol{W_t}(\boldsymbol{\beta_t} - \boldsymbol{\beta_{t|t-1}}) + \boldsymbol{e_t}] = \boldsymbol{W_t P_{t|t-1} W_t'} + \boldsymbol{H}, \quad (4.10)$$

para $t = 1, \ldots, n$. Assume-se que $\boldsymbol{\Sigma_{t|t-1}} > 0$ (é definida positiva), o que é garantido, por exemplo, se $\boldsymbol{H} > 0$. No entanto, esta suposição não é estritamente necessária.

Após o FK ter processado todas as n observações, é devolvido o estimador ótimo do vetor de estados atual. Este estimador possui toda a informação necessária para se fazer predições ótimas dos valores futuros dos estados e das observações. Uma grande vantagem relativa a este filtro é que, sendo um algoritmo recursivo, devolve o valor de $\boldsymbol{\beta_{t|t}}$ através de $\boldsymbol{\beta_{t|t-1}}$ e recorrendo apenas à nova observação $\boldsymbol{Y_t}$ sem a necessidade de reprocessar todos os dados observados anteriores $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_{t-1}}$.

Resumidamente, tendo em conta o MEE definido pelas equações (4.1) e (4.2), o algoritmo do FK traduz-se no processo recursivo (Costa, 2006):

1. Iniciar com os valores

$$\boldsymbol{\beta_{1|0}} = \boldsymbol{a_1} \text{ e } \boldsymbol{P_{1|0}} = \boldsymbol{P_1};$$

2. Obter a previsão de $\boldsymbol{Y_t}$, dada toda a informação disponível até ao instante $t-1$

$$\boldsymbol{Y_{t|t-1}} = \boldsymbol{W_t}\boldsymbol{\beta_{t|t-1}};$$

3. Atualizar a previsão do vetor de estados $\boldsymbol{\beta_t}$ e do respetivo EQM, quando a observação $\boldsymbol{Y_t}$ fica disponível

$$\boldsymbol{\beta_{t|t}} = \boldsymbol{\beta_{t|t-1}} + \boldsymbol{K_t}(\boldsymbol{Y_t} - \boldsymbol{Y_{t|t-1}}) \text{ e } \boldsymbol{P_{t|t}} = [\boldsymbol{I} - \boldsymbol{K_t}\boldsymbol{W_t}]\boldsymbol{P_{t|t-1}};$$

   onde $\boldsymbol{\eta_{t|t-1}} = \boldsymbol{Y_t} - \boldsymbol{Y_{t|t-1}}$ é o erro de previsão no instante t e $\boldsymbol{K_t}$ é o ganho de Kalman dado por

$$\boldsymbol{K_t} = \boldsymbol{P_{t|t-1}}\boldsymbol{W_t'}[\boldsymbol{W_t}\boldsymbol{P_{t|t-1}}\boldsymbol{W_t'} + \boldsymbol{H}]^{-1};$$

4. Predizer o vetor de estados $\boldsymbol{\beta_{t+1}}$ e do respetivo EQM, dada toda a informação disponível até ao instante t

$$\boldsymbol{\beta_{t+1|t}} = \boldsymbol{\Phi}\boldsymbol{\beta_{t|t}} \text{ e } \boldsymbol{P_{t+1|t}} = \boldsymbol{\Phi}\boldsymbol{P_{t|t}}\boldsymbol{\Phi'} + \boldsymbol{Q};$$

5. Recomeçar no ponto 2.

Caso a equação de observação contenha uma constante aditiva, por exemplo, $\boldsymbol{Y_t} = \boldsymbol{\alpha} + \boldsymbol{W_t}\boldsymbol{\beta_t} + \boldsymbol{e_t}$, o algoritmo do FK deve ser atualizado na equação de previsão de $\boldsymbol{Y_t}$, no ponto 2, da forma

$$\boldsymbol{Y_{t|t-1}} = \boldsymbol{\alpha} + \boldsymbol{W_t}\boldsymbol{\beta_{t|t-1}}.$$

4.2 Alisamento de Kalman

Esta técnica de alisamento (*smoothing*) consiste em obter $\boldsymbol{\beta_t}$ com base em todas as n observações $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_n}$, ou seja, permite obter $\boldsymbol{\beta_{t|n}}$, onde $t \leqslant n$. Como é espectável, o uso total dos dados observados leva a um aumento da precisão dos

parâmetros estimados. Esta metodologia designa-se desta forma porque o gráfico de $\boldsymbol{\beta_{t|n}}$ é mais "liso" do que as previsões $\boldsymbol{\beta_{t|t-1}}$ e os filtros $\boldsymbol{\beta_{t|t}}$ (Shumway e Stoffer, 2019).

Para o MEE definido pelas equações (4.1) e (4.2), com as condições iniciais $\boldsymbol{\beta_{n|n}}$ e $\boldsymbol{P_{n|n}}$ dadas em (4.5) e (4.6), respetivamente, tem-se

$$\boldsymbol{\beta_{t-1|n}} = \boldsymbol{\beta_{t-1|t-1}} + \boldsymbol{J_{t-1}}(\boldsymbol{\beta_{t|n}} - \boldsymbol{\beta_{t|t-1}}), \tag{4.11}$$

$$\boldsymbol{P_{t-1|n}} = \boldsymbol{P_{t-1|t-1}} + \boldsymbol{J_{t-1}}(\boldsymbol{P_{t|n}} - \boldsymbol{P_{t|t-1}})\boldsymbol{J'_{t-1}}, \tag{4.12}$$

onde

$$\boldsymbol{J_{t-1}} = \boldsymbol{P_{t-1|t-1}}\boldsymbol{\Phi'}[\boldsymbol{P_{t|t-1}}]^{-1}, \tag{4.13}$$

para $t = n, n-1, \ldots, 1$. As demonstrações de (4.11), (4.12) e (4.13) podem ser consultada em Shumway e Stoffer (2017).

Em suma, as estimativas obtidas a partir do alisamento são mais precisas, comparadas às obtidas a partir das previsões ou à filtragem. Isto deve-se à quantidade de dados utilizados em cada recursão. No alisamento são utilizados todos os dados disponíveis, a par que nas previsões são utilizadas todas a informações disponíveis para t dado $t-1$ e no FK são utilizados os dados disponíveis até ao momento t. Consequentemente, os intervalos de confiança para $\boldsymbol{\beta_{t|n}}$ apresentarão menor variabilidade.

4.3 Previsão

Para calcular as previsões de uma série temporal na formulação de espaço de estados recorre-se ao FK, onde os dados vão sendo atualizados sequencialmente a 1-passo sempre que uma nova observação é introduzida. Considerando o modelo Gaussiano de espaço de estados definido pelas equações (4.1) e (4.2), o FK estima o vetor de estados no instante n, com base em todas as n observações, ou seja, estima $\boldsymbol{\beta_{n|n}}$. Assim, a última observação, $\boldsymbol{Y_n}$, pode ser utilizada para atualizar o vetor de estados no instante $n+1$ da seguinte forma:

$$\boldsymbol{\beta_{n+1|n}} = \boldsymbol{\Phi}\boldsymbol{\beta_{n|n}},$$

onde a previsão a 1-passo é dada pela expressão

$$\boldsymbol{Y_{n+1|n}} = \boldsymbol{W_{n+1}}\boldsymbol{\beta_{n+1|n}}.$$

Considerando agora o problema da previsão a h-passos, onde $h = 2, 3, \ldots$, substituindo repetidamente na equação de estado (4.2) no instante $n + h$, resulta

$$\boldsymbol{\beta_{n+h}} = \boldsymbol{\Phi}^h \boldsymbol{\beta_n} + \sum_{j=1}^{h-1} \boldsymbol{\Phi}^{h-j} \boldsymbol{\varepsilon_{n+j}} + \boldsymbol{\varepsilon_{n+h}}, \quad h = 2, 3, \ldots$$

donde

$$\boldsymbol{\beta_{n+h|n}} = E(\boldsymbol{\beta_{n+h}} | \widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}) = \boldsymbol{\Phi}^h \boldsymbol{\beta_{n|n}},$$

e a matriz de covariâncias é dada pela expressão

$$\boldsymbol{P_{n+h|n}} = \boldsymbol{\Phi}^h \boldsymbol{P_{n|n}} \boldsymbol{\Phi}^{h\prime} + \sum_{j=0}^{h-1} \boldsymbol{\Phi}^j \boldsymbol{Q} \boldsymbol{\Phi}^{j\prime}, \quad h = 2, 3, \ldots$$

A estimativa para $\boldsymbol{Y_{n+h}}$ pode ser obtida diretamente de $\boldsymbol{\beta_{n+h|n}}$. Tomando o valor esperado na equação de observação no instante $n + h$, obtém-se

$$\boldsymbol{Y_{n+h|n}} = E(\boldsymbol{Y_{n+h}} | \widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}) = \boldsymbol{W_{n+h}} \boldsymbol{\beta_{n+h|n}}, \quad h = 2, 3, \ldots$$

A matriz de covariâncias é, então, dada por

$$\boldsymbol{\Sigma_{n+h|n}} = \boldsymbol{W_{n+h}} \boldsymbol{P_{n+h|n}} \boldsymbol{W'_{n+h}} + \boldsymbol{H}, \quad h = 2, 3, \ldots$$

4.3.1 Intervalos de Previsão

Por vezes é desejável determinar não só o valor da previsão como também é desejável determinar os intervalos de previsão pois permitem quantificar a incerteza associada às previsões pontuais, dando uma ideia do quão confiável é a previsão. Geralmente, os intervalos de previsão baseiam-se no $\boldsymbol{\Sigma_{t|t-1}}$ pois fornecem uma estimativa da variância do erro de previsão a 1-passo. Também é importante salientar que os intervalos de previsão só são válidos se o modelo ajustado descrever de forma satisfatória a série temporal e os respetivos pressupostos forem válidos. Além disso, os erros de previsão devem seguir uma distribuição Normal com média nula (Makridakis *et al.*, 1998). Sob estas suposições e para o caso univariado, por simplificação dos cálculos, tem-se

$$\frac{Y_t - Y_{t|t-1}}{\sqrt{\Sigma_{t|t-1}}} \sim N(0, 1),$$

onde $\Sigma_{t|t-1}$ corresponde à variância das inovações dada por (4.10) para o caso univariado.

Considerando que o grau de confiança é de $1-\alpha$, com $0<\alpha<1$, obtém-se

$$P\left(-z_{1-\alpha/2} < \frac{Y_t - Y_{t|t-1}}{\sqrt{\Sigma_{t|t-1}}} < z_{1-\alpha/2}\right) = 1-\alpha,$$

donde

$$P\left(Y_{t|t-1} - z_{1-\alpha/2}\sqrt{\Sigma_{t|t-1}} < Y_t < Y_{t|t-1} + z_{1-\alpha/2}\sqrt{\Sigma_{t|t-1}}\right).$$

Portanto, o intervalo de previsão a $(1-\alpha)\times 100\%$ para Y_t é dado por

$$\left(Y_{t|t-1} - z_{1-\alpha/2}\sqrt{\Sigma_{t|t-1}}, Y_{t|t-1} + z_{1-\alpha/2}\sqrt{\Sigma_{t|t-1}}\right), \tag{4.14}$$

onde $z_{1-\alpha/2}$ é o quantil da distribuição Normal padrão e $1-\alpha$ é o grau de confiança do intervalo.

Segundo Harvey (1992), à medida que o horizonte temporal da previsão aumenta, também aumenta a incerteza associada às previsões e o intervalo de previsão.

De acordo com Commandeur e Koopman (2007), as previsões são, por natureza, sujeitas a mais incerteza do que qualquer outro valor estimado dentro do intervalo temporal da série observada e, portanto, é habitual ser um pouco menos conservador nos limites dos intervalos de previsão. Em vez dos valores usuais de 95%, costumam ser usados limites de confiança de 90% ou 85%, ou até mais baixos.

Para previsões a h-passos, o intervalo de previsão a $(1-\alpha)\times 100\%$ para Y_t é dado por

$$\left(Y_{t+h|t} - z_{1-\alpha/2}\sqrt{\Sigma_{t+h|t}}, Y_{t+h|t} + z_{1-\alpha/2}\sqrt{\Sigma_{t+h|t}}\right), \tag{4.15}$$

onde $z_{1-\alpha/2}$ é o quantil da distribuição Normal padrão e $1-\alpha$ é o grau de confiança do intervalo.

Para o cálculo dos intervalos de previsão (4.14) e (4.15), é exigido que os erros tenham distribuição Normal.

4.4 Inicialização do Filtro de Kalman

Até agora foi assumido por conveniência que o estado inicial $\boldsymbol{\beta_1} \sim N(\boldsymbol{a_1}, \boldsymbol{P_1})$, onde $\boldsymbol{a_1}$ e $\boldsymbol{P_1}$ são conhecidos. No entanto, na maioria dos casos reais alguns ou todos os elementos de $\boldsymbol{a_1}$ e $\boldsymbol{P_1}$ não são conhecidos. Nesta situação, existem duas alternativas: assume-se que os elementos correspondentes de $\boldsymbol{a_1}$ são fixos e estima-se por máxima verosimilhança ou assume-se que têm média zero e matriz de variância $k\mathbf{I}$, onde $\mathbf{I}$ é a matriz identidade e $k \to \infty$. Neste último caso, chama-se a distri-

buição de **difusa**. Este procedimento é conhecido como **Inicialização Difusa do Filtro de Kalman**. No entanto, Harvey *et al.* (2004) afirmam que as duas abordagens fornecem resultados muito próximos, onde as equações recursivas da filtragem e do alisamento resultantes têm uma estrutura semelhante àquelas com $\boldsymbol{a_1}$ e $\boldsymbol{P_1}$ conhecidas.

4.5 Abordagens de Espaço de Estados *versus* ARMA

Uma das áreas mais importantes que envolve a modelação de séries temporais é a teoria de controlo que trata do problema de manter o valor de alguma variável o mais próximo possível de um valor ideal num sistema sujeito a perturbações que, dada a sua natureza, o problema de controlo necessita de métodos adaptativos ou recursivos através dos quais novas observações vão sendo incorporadas no modelo, atualizando-o e tornando-o capaz de conduzir as melhores estimativas e previsões (Murteira *et al.*, 1993). Os MEE são uma ótima ferramenta que lidam com os problemas de controlo. Além disso, esta metodologia permite lidar com séries cujas observações foram perdidas ou feitas em intervalos irregulares. A vantagem dos MEE face aos modelos ARMA é que os modelos ARMA são insuficientes para lidar com os problemas de controlo uma vez que trabalham com base em amostras de dimensão fixa, onde os parâmetros do modelo vão sendo estimados a partir de dados observados em vez das estimativas serem adaptadas recursivamente à medida que novas observações vão sendo disponibilizadas.

A formulação dos MEE procuram modelar explicitamente as séries não estacionárias, tendo em conta as diferentes componentes que a constituem, tais como a tendência, a sazonalidade e a componente cíclica, juntamente com o efeito das variáveis explicativas, enquanto que nos modelos ARMA estas componentes precisam de ser removidas antes de qualquer análise, dada a necessidade da estacionaridade das séries (Durbin e Koopman, 2001).

Capítulo 5

Estimação dos Parâmetros do Modelo de Espaço de Estados

Considere o MEE definido pelas equações (4.1) e (4.2). Seja $\boldsymbol{\Theta} = \{\boldsymbol{\Phi}, \boldsymbol{H}, \boldsymbol{Q}\}$ o vetor dos parâmetros desconhecidos do modelo. Pretende-se estimar esses parâmetros e existem várias maneiras de o fazer. O método mais usual é o **método da máxima verosimilhança** que consiste em encontrar a função de verosimilhança do modelo e, em seguida, estimar as componentes de $\boldsymbol{\Theta}$ que maximize a função, ou seja, pretende-se maximizar a probabilidade das observações $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_n}$ em relação aos componentes do vetor $\boldsymbol{\Theta}$. Assume-se que o estado inicial segue uma distribuição Normal, $\boldsymbol{\beta_1} \sim N(\boldsymbol{a_1}, \boldsymbol{P_1})$, assim como as componentes aleatórias, $\boldsymbol{\varepsilon_t} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \boldsymbol{Q})$ e $\boldsymbol{e_t} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \boldsymbol{H})$. Por simplicidade, assume-se que $\boldsymbol{\varepsilon_t}$ e $\boldsymbol{e_t}$ são não correlacionados. A probabilidade é calculada usando as inovações

$$\boldsymbol{\eta_{t|t-1}} = \boldsymbol{Y_t} - \boldsymbol{W_t}\boldsymbol{\beta_{t|t-1}}, \quad t = 1, \ldots, n.$$

A função de verosimilhança pode ser determinada a partir dos valores obtidos pelo FK e é dada por

$$L(\boldsymbol{\Theta}; \widetilde{\boldsymbol{Y}}_{\boldsymbol{n}}) = \prod_{t=1}^{n} p(\boldsymbol{Y_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}), \tag{5.1}$$

onde $\widetilde{\boldsymbol{Y}}_{\boldsymbol{n}} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_n})'$ e $p(\boldsymbol{Y_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}})$ é a densidade de $\boldsymbol{Y_t}$ dado $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}$, que pode ser escrita como

$$p(\boldsymbol{Y_t}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{t-1}}) = (2\pi)^{-1/2}|\boldsymbol{\Sigma_{t|t-1}}(\boldsymbol{\Theta})|^{-1/2} \exp\{\boldsymbol{\eta_{t|t-1}}(\boldsymbol{\Theta})'\boldsymbol{\Sigma_{t|t-1}}(\boldsymbol{\Theta})^{-1}\boldsymbol{\eta_{t|t-1}}(\boldsymbol{\Theta})\}, \tag{5.2}$$

onde $\boldsymbol{\Sigma_{t|t-1}} = \boldsymbol{W_t P_{t|t-1} W_t'} + \boldsymbol{H}$ representa a matriz de covariâncias das inovações.

Aplicando o logaritmo a (5.1), a função de log-verosimilhança pode ser escrita como

$$\begin{aligned} \log L(\boldsymbol{\Theta}; \widetilde{\boldsymbol{Y}}_{\boldsymbol{n}}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n} \log|\boldsymbol{\Sigma_{t|t-1}}(\boldsymbol{\Theta})| - \\ - \frac{1}{2}\sum_{t=1}^{n} \boldsymbol{\eta_{t|t-1}}(\boldsymbol{\Theta})' \boldsymbol{\Sigma_{t|t-1}}(\boldsymbol{\Theta})^{-1} \boldsymbol{\eta_{t|t-1}}(\boldsymbol{\Theta}). \end{aligned} \tag{5.3}$$

Na função (5.3), destaca-se a dependência das inovações, $\boldsymbol{\eta_{t|t-1}}(\boldsymbol{\Theta})$, e da respetiva matriz de covariâncias, $\boldsymbol{\Sigma_{t|t-1}}(\boldsymbol{\Theta})$, em relação ao vetor de parâmetros $\boldsymbol{\Theta}$ que se pretende estimar.

Em algumas aplicações, quando o processo $\{\boldsymbol{\beta_t}\}_{t=1,\ldots,n}$ não é estacionário, os valores iniciais do FK $\boldsymbol{\beta_{1|0}}$ e $\boldsymbol{P_{1|0}}$ podem ser incluídos em $\boldsymbol{\Theta}$ de modo a que sejam estimados com base na amostra. No entanto, esta inclusão implica um aumento da complexidade do processo de otimização da log-verosimilhança. Como alternativa pode optar-se por valores iniciais não informativos ou por algoritmos mais específicos (Costa, 2006).

Uma vez que a função (5.3) é altamente não linear e também não é fácil de ser minimizada, a estimação dos parâmetros é feita através da maximização da função de log-verosimilhança recorrendo a algoritmos de otimização. Geralmente estes algoritmos partem de um conjunto inicial de valores $\boldsymbol{\Theta}$, realizam uma série de iterações e, para cada iteração é calculada um novo valor para a função de log-verosimilhança. Quando a função não sofrer grandes alterações, o algoritmo termina. Existem vários métodos numéricos que podem ser aplicados, mas talvez o mais conhecido seja o **algoritmo *Expectation-Maximization***. Este algoritmo foi desenvolvido por Dempster *et al.* (1977) e é utilizado para calcular estimativas de máxima verosimilhança dos parâmetros do modelo em dois passos: E-*step* (esperança), que consiste no cálculo do valor esperado da função de log-verosimilhança (5.3) e o M-*step*, que consiste em maximizar o valor esperado da função de log-verosimilhança (5.3). No entanto, este algoritmo apresenta alguns inconvenientes, nomeadamente a necessidade de um grande número de iterações até se verificar a convergência (convergência lenta) e é sensível à escolha inicial dos parâmetros.

Por outro lado, existem vários métodos numéricos que se baseiam no **método**

de Newton-Raphson na qual se resolve a equação

$$\partial_1(\boldsymbol{\Theta}) = \frac{\partial \log L(\boldsymbol{\Theta}; \widetilde{\boldsymbol{Y}}_{\boldsymbol{n}})}{\partial \boldsymbol{\Theta}} = 0, \tag{5.4}$$

aplicando a série de Taylor de primeira ordem tem-se

$$\partial_1(\boldsymbol{\Theta}) \simeq \tilde{\partial}_1(\boldsymbol{\Theta}) + \tilde{\partial}_2(\boldsymbol{\Theta})(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}), \tag{5.5}$$

para algum valor ótimo $\tilde{\boldsymbol{\Theta}}$, onde

$$\tilde{\partial}_1(\boldsymbol{\Theta}) = \partial_1(\boldsymbol{\Theta})|_{\boldsymbol{\Theta}=\tilde{\boldsymbol{\Theta}}}, \qquad \tilde{\partial}_2(\boldsymbol{\Theta}) = \partial_2(\boldsymbol{\Theta})|_{\boldsymbol{\Theta}=\tilde{\boldsymbol{\Theta}}},$$

com

$$\partial_2(\boldsymbol{\Theta}) = \frac{\partial^2 \log L(\boldsymbol{\Theta}; \widetilde{\boldsymbol{Y}}_{\boldsymbol{n}})}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'}. \tag{5.6}$$

Igualando (5.5) a zero obtém-se o valor corrigido $\bar{\boldsymbol{\Theta}}$, através da expressão

$$\bar{\boldsymbol{\Theta}} = \tilde{\boldsymbol{\Theta}} - \tilde{\partial}_2(\boldsymbol{\Theta})^{-1} \tilde{\partial}_1(\boldsymbol{\Theta}).$$

Este processo é repetido até convergir. $\partial_1(\boldsymbol{\Theta})$ é chamado de **gradiente** ou **vetor *score*** e determina a direção na qual se espera que a função de log-verosimilhança (5.3) cresça até encontrar um ponto onde assuma o valor máximo. $\partial_2(\psi)$ representa a **matriz Hessiana** que contém a informação sobre a curvatura da função (Durbin e Koopman, 2001). Na prática, o cálculo numérico do gradiente geralmente é possível. No entanto, o cálculo da hessiana requer, na maioria dos casos, um grande esforço computacional e geralmente é aproximada por diferentes métodos. Um exemplo é o **método BFGS** (Broyden-Fletcher-Goldfarb-Shannon), que faz a aproximação à inversa da matriz hessiana, diminuindo, assim, o custo computacional quando comparado com o cálculo exato da mesma.

Neste trabalho será utilizado o método BFGS, através da função *optim()*, implementada no *package stats* do *software* estatístico R.

Capítulo 6

Avaliação dos Modelos

"All models are wrong, but some models are more useful than others." (em português, *Todos os modelos estão errados, mas alguns são mais úteis do que outros*) é uma citação muito conhecida frequentemente atribuída a George Box. A ideia de que "todos os modelos estão errados" é de que um modelo, sendo uma representação simplificada da realidade, nunca representará o comportamento real exato, "mas alguns são mais úteis do que outros" pois, apesar de serem uma aproximação da realidade, alguns modelos conseguem captar melhor a estrutura inerente aos dados do que outros e, além disso, podem ser bastante úteis para explicar, prever e entender as várias componentes presentes nos dados. No entanto, é muito importante ter em atenção aos pressupostos na construção de um modelo porque os modelos só são realmente úteis quando esses pressupostos são verificados.

Para comparar diferentes MEE, é muito frequente recorrer a critérios de informação, como por exemplo o AIC ou o BIC, que fornecem uma medida da qualidade de ajustamento dos modelos. Complementarmente, é usual, também, analisar-se um conjunto de indicadores de desempenho baseados na precisão das previsões. Contudo, tornou-se prática comum o uso de técnicas de validação cruzada na regressão. Esses métodos também são utilizados na literatura para avaliar, por exemplo, regressões automáticas em séries temporais (Bergmeir e Benítez, 2012).

6.1 Medidas de Avaliação

Para avaliar o desempenho de um determinado modelo ou método de previsão, deve-se estudar a precisão das previsões obtidas. As medidas de avaliação são utilizadas para verificar o nível de precisão que o modelo de previsão possui. Assim, quanto mais próximas as previsões forem aos dados observados da série, menor será

o erro de previsão. Caso contrário, se os erros de previsão forem elevados, é sinal de que o modelo de previsão não é adequado. Segundo Makridakis *et al.* (1998), as medidas de avaliação são, na maioria dos casos de análise de previsões, o critério crucial da seleção de um método de previsão.

As medidas de avaliação podem ser classificadas em três tipos distintos: medidas dependentes da escala, medidas baseadas em erros percentuais e medidas de erros escalados.

Medidas dependentes da escala

Estas medidas de previsão são muito úteis para comparar diferentes métodos aplicados a conjunto de dados com a mesma escala e, portanto, não devem ser utilizadas para comparar dados com escalas diferentes, como por exemplo a temperatura (°C) e o comprimento (m). As medidas dependentes da escala mais utilizadas baseiam-se nos erros absolutos e nos erros quadráticos.

- **Erro Médio (EM)**
  O EM representa o valor médio dos desvios entre os valores observados Y_t e as previsões $\hat{Y}_t$ para os instantes $t = 1, 2, \ldots, n$, dada por

  $$\text{EM} = \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - \hat{Y}_t \right).$$

  No entanto, o EM não dá indicação relativamente ao tamanho dos erros típicos uma vez que os erros positivos e negativos tendem a compensar-se e, por isso, é provável que o EM produza um valor pequeno (Makridakis *et al.*, 1998). Na verdade, esta medida apenas indica se há sub ou sobreprevisão sistemática (viés de previsão).

- **Erro Quadrático Médio (EQM)**
  O EQM define-se como sendo o valor médio dos desvios ao quadrado entre os valores observados Y_t e as previsões $\hat{Y}_t$ para os instantes $1, 2, \ldots, n$, dado por

  $$\text{EQM} = \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - \hat{Y}_t \right)^2.$$

  A raiz do erro quadrático médio, $\text{REQM} = \sqrt{\text{EQM}}$, é uma das medidas mais utilizadas na literatura para reduzir a grandeza dos valores. O modelo com menor EQM (e, consequentemente, menor REQM) é considerado o mais adequado.

Segundo Chai e Draxler (2014), o REQM é uma medida de avaliação frequentemente utilizada para medir o desempenho dos modelos em estudos meteorológicos, qualidade do ar e pesquisas climáticas.

- **Erro Absoluto Médio (EAM)**
  O EAM representa o valor absoluto médio dos desvios entre os valores observados Y_t e as previsões $\hat{Y}_t$ para os instantes $t = 1, 2, \ldots, n$, dada por

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

Segundo Makridakis *et al.* (1998), o EAM tem a vantagem de ser facilmente interpretável. Por outro lado, o EQM tem a vantagem de ser mais fácil de manipular matematicamente e, portanto, é frequentemente utilizado na otimização estatística. Analogamente, o modelo com menor EAM é considerado o mais adequado.

Medidas baseadas em erros percentuais

Para superar o problema da dependência de escala, define-se o erro percentual (ou relativo) que é dado por $p_t = 100 \times (Y_t - \hat{Y}_t)/Y_t$. Os erros percentuais são frequentemente utilizados para comparar o desempenho das previsões entre diferentes séries.

- **Erro Percentual Médio (EPM)**
  O EPM representa o valor percentual médio dos desvios entre os valores observados Y_t e as previsões $\hat{Y}_t$ para os instantes $t = 1, 2, \ldots, n$, dada por

$$\text{EPM} = \frac{1}{n} \sum_{t=1}^{n} p_t = \frac{1}{n} \sum_{t=1}^{n} \frac{Y_t - \hat{Y}_t}{Y_t} \times 100 \quad (\%) .$$

  Analogamente, como acontece com o EM, é presumível que o EPM seja pequeno visto que os erros percentuais positivos e negativos tendem a compensar-se mutuamente.

- **Erro Percentual Absoluto Médio (EPAM)**
  O EPAM representa o valor percentual absoluto médio dos desvios entre os valores observados Y_t e as previsões $\hat{Y}_t$ para os instantes $t = 1, 2, \ldots, n$, dada por

$$\text{EPAM} = \frac{1}{n} \sum_{t=1}^{n} |p_t| = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right| \times 100 \quad (\%) .$$

No entanto, apesar destas medidas possibilitarem a comparação entre dados com diferentes escalas, têm a desvantagem de serem infinitas ou indefinidas quando $Y_t = 0$ e de tenderem para valores extremos quando Y_t está próximo de zero. Além disso, estas medidas assumem um zero não arbitrário (por exemplo, não se deve utilizar estas medidas para avaliar a precisão das previsões da temperatura nas escalas Celsius ou Fahrenheit, uma vez que os pontos zero de ambas escalas não refletem a ausência de temperatura).

Medidas de Erros Escalados

Estas medidas foram propostas por Hyndman e Koehler (2006), como uma alternativa às medidas baseadas em erros percentuais. Podem ser utilizadas para comparar a precisão das previsões entre séries expressas em diferentes escalas. A ideia destas medidas consiste em "escalar" os erros de previsão dividindo-os pelo EAM *in-sample*, isto é, calculado na amostra de treino através do método *naïve*[1]. O erro escalado é dado por $q_t = (Y_t - \hat{Y}_t)/Q_t$, onde Q_t representa o erro absoluto médio (EAM) da previsão *naïve*.

Para séries temporais sem sazonalidade, o erro escalado pode ser definido da seguinte forma

$$q_t = \frac{Y_t - \hat{Y}_t}{\frac{1}{n-1}\sum_{t=2}^{n} |Y_t - Y_{t-1}|}.$$

Para séries temporais com sazonalidade, o erro escalado utiliza previsões *naïve* sazonais e é dado por

$$q_t = \frac{Y_t - \hat{Y}_t}{\frac{1}{n-s}\sum_{t=s+1}^{n} |Y_t - Y_{t-s}|},$$

onde s corresponde ao período da sazonalidade.

- **Erro Escalado Absoluto Médio (EEAM)**
  O EEAM representa a média dos erros escalados absolutos e é definido por

$$\text{EEAM} = \frac{1}{n}\sum_{t=1}^{n} |q_t|.$$

[1] O método *naïve* é um método de referência simples e bastante comum que consiste em utilizar a última observação como previsão, ou seja, $Y_{t|t-1} = Y_{t-1}$, $t = 1, 2, \ldots, n$, onde n é a dimensão da amostra. Outro método frequentemente utilizado baseia-se em utilizar a média de todas as observações disponíveis como previsão.

Quando EEAM é inferior a 1, as previsões obtidas pelo método adotado são mais precisas, em média, do que as previsões do método *naïve*. Para comparar diferentes métodos de previsão, considera-se o mais preciso o que tiver menor EEAM.

O EEAM pode ser utilizado para comparar o desempenho das previsões entre séries com diferentes escalas. Além disso, possui uma escala significativa e são menos sensíveis a valores discrepantes. A única circunstância sob a qual essas medidas são infinitas ou indefinidas é quando todas as observações são iguais.

Atendendo às restrições destas medidas de avaliação, deve-se ter especial atenção na escolha delas, a fim de evitar tirar conclusões erradas.

Estatística U de Theil

A estatística U de Theil é uma medida de precisão que compara os métodos formais de previsão com o método *naïve* e, além disso, compara os erros de previsão envolvidos, penalizando mais os erros elevados em relação aos mais pequenos. Esta estatística é dada por (Makridakis *et al.*, 1998)

$$\mathrm{U} = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{\hat{Y}_{t+1}-Y_{t+1}}{Y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{Y_{t+1}-Y_t}{Y_t}\right)^2}}.$$

Assim, tem-se que se U = 1, o método *naïve* é tão eficiente quanto o método em avaliação; se U < 1, o método *naïve* é menos eficiente que o método em avaliação; se U > 1, o método *naïve* é mais eficiente que o método em avaliação e se U = 0, o método em avaliação é perfeito.

6.2 Critérios de Seleção

Muitas vezes, na análise de séries temporais podem surgir mais do que um modelo estatístico que descreva de forma satisfatória o comportamento da série em estudo e, portanto, é necessário escolher um modelo, de entre todos os possíveis modelos, que melhor se ajusta aos dados. Para medir a qualidade de ajustamento dos modelos, foram introduzidos os critérios de seleção, que têm por base a maximização da função de verosimilhança, penalizando os modelos com maior número de parâmetros.

- **Critério de informação de Akaike (AIC)**
  O critério de informação de Akaike (AIC), proposto por Akaike (1974), é um dos critérios de informação mais utilizados e é definido por

$$\mathrm{AIC} = -2\log L(\hat{\Theta}) + 2p, \tag{6.1}$$

  onde $L(\hat{\Theta})$ é a função de verosimilhança e p é o número de parâmetros do modelo.

- **Critério de informação Bayesiano (BIC)**
  Este critério, também conhecido por critério de informação de Schwarz (BIC), foi proposto por Schwarz (1978) e é semelhante ao AIC, mas a penalização é maior neste critério pela inclusão de parâmetros adicionais a serem estimados. É dado por

$$\mathrm{BIC} = -2\log L(\hat{\Theta}) + p\log(n), \tag{6.2}$$

  onde $L(\hat{\Theta})$ é a função de verosimilhança, p é o número de parâmetros do modelo e n é a dimensão da amostra.

Segundo estes critérios, considera-se que o melhor modelo de previsão é aquele que apresentar os menores valores de AIC ou BIC. Comparando estes dois critérios, a diferença entre o AIC e o BIC está no termo de penalização, sendo $2p$ no caso do AIC e $p\log(n)$ no caso de BIC. Repare que o BIC depende da dimensão da amostra, contrariamente ao AIC. Para $n > 8$ (ou seja, $\log(n) > 2$), a penalização do BIC é superior à penalização do AIC por cada inclusão de novos parâmetros no modelo.

Uma vez que os critérios AIC e BIC penalizam os modelos com muitos parâmetros, eles podem ser utilizados para comparar modelos com diferentes número de parâmetros (Durbin e Koopman, 2001).

Note-se que a complexidade de um modelo é geralmente definida pela quantidade de parâmetros que o constitui. Modelos mais complexos tendem a ter verosimilhança mais elevada. No entanto, o que se pretende é encontrar um modelo que seja o mais simples possível e, ao mesmo tempo, obter um bom ajustamento, ou seja, pretende-se encontrar um modelo parcimonioso (Box e Jenkins, 1976), que não contenha mais do que a quantidade necessária de complexidade.

Assim, através destes critérios obtém-se um compromisso entre a verosimilhança elevada (resíduos pequenos) e a complexidade do modelo (número reduzido de parâmetros).

6.3 Validação Cruzada

Um dos métodos mais utilizados para medir o desempenho de um determinado modelo preditivo é a **validação cruzada** (*Cross-Validation*). A ideia base deste método consiste em dividir os dados em dois ou mais subconjuntos. Parte dos dados, nomeadamente a **amostra de treino** é utilizada para treinar, ou seja, para identificar e estimar o modelo, enquanto a **amostra de teste** é reservada para avaliar a qualidade das previsões do modelo (Arlot e Celisse, 2010).

Existem vários métodos de validação cruzada (CV), mas certamente o procedimento mais conhecido é o *k-fold cross-validation* (Burman, 1989) que consiste em dividir aleatoriamente o conjunto de dados em k subconjuntos com a mesma dimensão, onde um subconjunto é escolhido para teste e os restantes $k-1$ são utilizados para treinar o modelo. Este processo é realizado k vezes de forma a que todos os subconjuntos tenham sido utilizados como amostra de teste. No final é calculada a precisão das previsões através dos erros obtidos. A questão da escolha de k permanece amplamente aberta, mesmo que possam ser dadas indicações para uma escolha apropriada (Arlot e Celisse, 2010).

No entanto, quando os dados são dependentes, como acontece nos dados de séries temporais, grande parte dos métodos de CV clássicos não podem ser aplicados devido aos pressupostos fundamentais dos dados serem independentes e identicamente distribuídos (i.i.d.). As séries temporais são geradas por processos que evoluem ao longo do tempo e a aleatorização nos algoritmos da CV clássicos, como por exemplo o k-fold CV, não preservam a ordem das observações. Além disso, a correlação temporal inerente e a potencial não estacionariedade dos dados, faz com que o uso da CV seja problemática, uma vez que não leva em conta esses problemas (Bergmeir *et al.*, 2018).

No entanto, Bergmeir *et al.* (2018) referem que, para modelos puramente autorregressivos, é possível aplicar a k-fold CV, desde que os modelos considerados apresentem erros não correlacionados.

Portanto, devido às características presentes nos dados de séries temporais, os métodos de CV devem ser ajustados. Para avaliar a qualidade das previsões obtidas, o procedimento mais utilizado consiste resumidamente em reservar uma parte final da série para avaliar a precisão das previsões (amostra de teste) e utilizar a restante da série para estimar quaisquer parâmetros do modelo de previsão (amostra de treino). Na Figura 6.1 está representada a abordagem descrita.

Figura 6.1: Série temporal dividida em amostra de treino (pontos verdes) e amostra de teste (pontos vermelhos). Fonte: Hyndman e Athanasopoulos (2018).

Geralmente, a dimensão da amostra de teste é cerca de 20% da dimensão da amostra total, embora este valor dependa da dimensão da amostra e do horizonte temporal da previsão (Hyndman e Athanasopoulos, 2018). A dimensão da amostra de teste deve ser, pelo menos, igual ao horizonte temporal máximo da previsão necessária. Note que os dados pertencentes à amostra de teste não são utilizados para a construção do modelo.

Esta abordagem pode ser aplicada a dados de séries temporais uma vez que, como assumimos que o futuro depende do passado, a dependência natural dos dados é respeitada (Bergmeir e Benítez, 2012).

No entanto, esta abordagem apresenta algumas desvantagens. Um modelo que se ajuste bem aos dados da amostra de treino não significa que produzirá necessariamente previsões adequadas, uma vez que um ajuste perfeito pode ser obtido através de um modelo não parcimonioso. Além disso, para séries não estacionárias, esta abordagem pode ser enganosa, pois os dados da amostra de teste podem ser muito diferentes dos dados da amostra de treino e, portanto, o futuro desconhecido também pode ser diferente da amostra de treino, da amostra de teste ou de ambas. Outro inconveniente deste método, principalmente para séries com poucas observações, é o facto de não utilizar a totalidade dos dados disponíveis para ajustar o modelo. Devido a estas dificuldades, é prática comum na previsão de séries temporais assumir séries temporais estacionárias (Bergmeir e Benítez, 2012).

Segundo Bergmeir *et al.* (2018), o facto desta abordagem de dividir uma série na amostra de treino e de teste ser a mais aplicada no contexto das séries temporais, está relacionada com o facto do ajustamento dos modelos padrão, como por exemplo o alisamento exponencial ou os modelos ARIMA, serem totalmente iterativos, no sentido em que iniciam a estimação no início da série.

Inoue e Kilian (2006) mostram que os critérios de seleção AIC e BIC apresentam um desempenho assintoticamente melhor do que a abordagem de dividir uma série temporal na amostra de treino e de teste e, além disso, mencionam o facto desta abordagem falhar em utilizar apenas uma parte dos dados disponíveis, perdendo informações potencialmente importantes.

Posto isto, Hyndman (2014) e Hyndman e Athanasopoulos (2018) propõem uma

versão mais sofisticada na qual intitulam de **validação cruzada de séries temporais**. Nesta abordagem, os autores sugerem utilizar várias amostras de treino diferentes onde, por cada iteração, cada uma contém mais uma observação do que a amostra anterior e cada amostra de teste correspondente é constituída por apenas uma observação. As observações da amostra de treino, como ocorrem antes das observações que constituem a amostra de teste, nenhuma observação futura é utilizada na construção da previsão.

Na Figura 6.2 está representado a série da amostra de treino (a verde) e amostras de teste (a vermelho). Para calcular a precisão das previsões, são calculadas as medidas de avaliação a 1-passo.

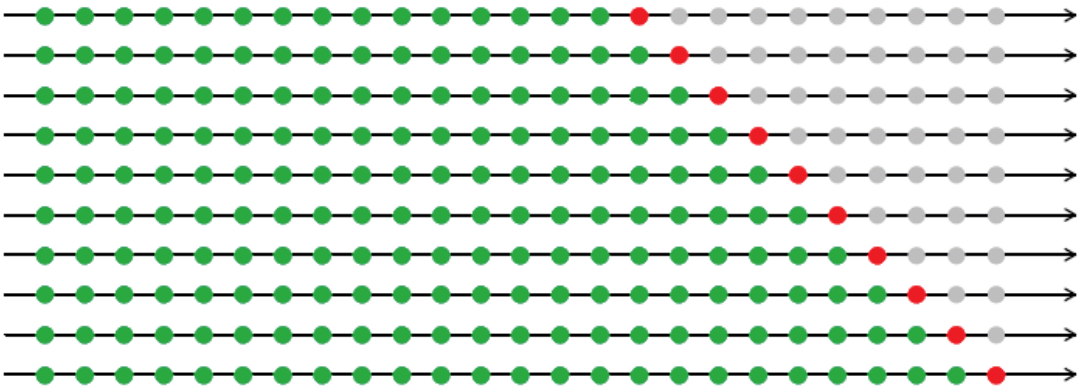


Figura 6.2: Validação cruzada de séries temporais baseadas nas previsões a 1-passo. Pontos verdes - amostras de treino; pontos vermelhos - amostra de teste; pontos cinzentos são ignorados. Fonte: Hyndman (2014).

Nesta abordagem de validação cruzada, é necessário definir um tamanho mínimo para a amostra de treino porque muitas vezes não é possível obter previsões confiáveis se não houver dados suficientes na amostra de treino para estimar o modelo escolhido. O tamanho mínimo da amostra de treino depende da complexidade do modelo que se deseja ajustar (Hyndman, 2014). Suponhamos que são necessárias k observações para obter previsões confiáveis. Então, o procedimento funciona da seguinte forma:

1. Selecionar a observação no tempo $k+i$ na amostra de teste e utilizar as observações nos tempos $1, 2, \ldots, k+i-1$ para estimar o modelo de previsão. Calcular o erro de previsão para o tempo $k+i$;

2. Repetir o passo 1 para $i = 1, 2, \ldots, n-k$, onde n representa o número total de observações;

3. Calcular as medidas de avaliação com base nos erros obtidos.

Através desta abordagem, uma boa maneira de encontrar o melhor modelo de previsão é escolher o que tiver o menor REQM (Hyndman e Athanasopoulos, 2018).

Uma desvantagem da validação cruzada de séries temporais comparativamente à abordagem de amostra de treino/teste, é que esta envolve apenas uma etapa de treino e de teste, enquanto a validação cruzada de séries temporais envolve essas etapas para todas as subamostras, sendo mais exigente computacionalmente. No entanto, Hyndman (2014) recomenda o uso de validação cruzada em séries temporais sempre que possível, em vez da simples divisão da amostra de treino/teste.

Existem outras abordagens de CV aplicadas à estimativa dos erros de previsão em ambiente dependente na qual têm sido estudadas extensivamente na literatura devido à sua capacidade de selecionar modelos no contexto não paramétrico. Várias técnicas de CV podem ser encontrados na literatura, destacando-se, por exemplo, a *Blocked* CV (Bergmeir *et al.*, 2014; Cerqueira *et al.*, 2019), h-Block CV (Burman *et al.*, 1994; Racine, 1997) e hv-Block CV (Racine, 2000).

6.4 Análise dos Resíduos

A análise dos resíduos permite avaliar a qualidade de ajustamento de um modelo e identificar observações que não são bem explicadas pelo modelo. Espera-se que um bom modelo, em termos da qualidade de ajustamento, gere resíduos com um comportamento próximo ao de um ruído branco, no caso dos modelos de regressão linear, e os resíduos padronizados, no caso dos MEE. Ou seja, espera-se que cumpram os pressupostos de média nula, variância constante e não correlação. Além disso, torna-se pertinente verificar se os resíduos apresentam uma distribuição aproximadamente Normal. Esta condição é necessária para a construção de intervalos de previsão e para a verificação das expressões anteriormente obtidas para estimadores de máxima verosimilhança. Daqui em diante, sempre que se referir aos resíduos, serão os resíduos no caso dos modelos de regressão linear e resíduos padronizados, no caso dos MEE, sendo, neste caso, a partir de (4.10), dado por

$$\eta^{\dagger}_{t|t-1} = \frac{\eta_{t|t-1}}{\sqrt{\Sigma_{t|t-1}}}.$$

- **Normalidade**

  A análise do pressuposto de normalidade dos resíduos pode ser feita através de representações gráficas, como o histograma, que deve aproximar-se do comportamento da função densidade de uma distribuição Normal e o *QQ-plot*, que re-

laciona os quantis de uma amostra com os quantis teóricos de uma distribuição teórica. Os pontos devem ficar próximos a uma reta, permitindo visualizar os desvios da normalidade. Além da análise gráfica, pode-se verificar a condição da normalidade dos resíduos recorrendo a testes estatísticos tais como o teste de Shapiro-Wilk, mais indicado para amostras de pequenas dimensões (menos de 30 observações), e o teste de Kolmogorov-Smirnov, em que ambos testam a mesma hipótese nula H_0 : *"Os erros seguem uma distribuição Normal"*.

- **Não correlação**

  Analogamente à condição da normalidade, a análise da não correlação dos resíduos pode ser feita por inspeção gráfica através da FAC e FACP amostrais, cujas estruturas de correlação devem aproximar-se ao de um ruído branco, tendo autocorrelações não significativamente diferentes de zero, devendo estar dentro dos limites do intervalo $\pm 1,96/\sqrt{n}$. Além disso, a avaliação do comportamento dos resíduos pode ser feita através de testes estatísticos, como os de *Portmanteau*, mais especificamente o teste de Ljung-Box, que testa se as primeiras k autocorrelações são conjuntamente nulas. A estatística de teste Q é dada por

  $$Q = n(n+2)\sum_{i=1}^{k}\frac{\hat{\rho}_i^2}{n-i},$$

  onde $Q \sim \chi^2_{k-p}$, sendo n o número de observações, p o número de parâmetros a serem estimados e k corresponde ao número de autocorrelações a serem testadas. Hyndman e Athanasopoulos (2018) sugerem fazer $k = 10$ para dados não sazonais e $k = 2m$ para dados sazonais, onde m corresponde ao período de sazonalidade. Caso $k > n/5$, sugerem utilizar $k = n/5$. Em caso de rejeição de H_0, conclui-se que o modelo não é apropriado.

- **Média nula e homocedasticidade**

  Para além das duas condições mencionadas na qual os resíduos devem cumprir, também devem ter média nula e variância constante ao longo do tempo (homocedasticidade). Para testar se a média dos resíduos é nula, recorre-se ao teste t para o valor médio. Note-se que este teste só pode ser aplicado caso os pressupostos de normalidade sejam verificados. No entanto, se a distribuição dos resíduos for aproximadamente simétrica e a dimensão da amostra for superior a 20 observações, o teste t pode ser aplicado. Para averiguar a condição de homocedasticidade, pode-se recorrer à análise gráfica através da representação dos resíduos ao longo do tempo.

Capítulo 7

Análise e Previsão das Séries de Dados Meteorológicos

Neste capítulo são analisadas séries temporais relativas às variáveis meteorológicas, ajustando vários modelos de calibração considerando modelos com o fator de calibração determinístico (em particular, os modelos de RLS) e estocástico (que correspondem aos MEE associados ao FK). Para isso, neste estudo é feita uma breve apresentação dos dados, seguida de uma análise exploratória. Esta etapa é muito importante pois permite identificar e compreender o comportamento dos dados que darão informações relevantes para a aplicação das metodologias já expostas. Este passo também permite identificar algumas relações existentes entre as variáveis, como por exemplo o horizonte temporal e a temperatura máxima e miníma prevista.

Para a realização da análise estatística, recorreu-se ao *software* estatístico R (versão 3.5.1). Para a implementação dos MEE, foram criados novos códigos, na qual foram utilizados algumas funções implementadas no *package* "astsa" (*Applied Statistical Time Series Analysis*), da autoria de David Stoffer (15/05/2020). Em toda a análise estatística, considera-se o nível de significância de 5%.

7.1 Caracterização da Base de Dados

Os dados utilizados neste estudo provêm de duas bases de dados distintas: a primeira corresponde a registos diários de duas variáveis meteorológicas observadas: temperatura máxima do ar (°C) e temperatura mínima do ar (°C), registadas numa estação meteorológica portátil instalada numa quinta chamada Senhora da Ribeira, em Carrazeda de Ansiães, situada no distrito de Bragança, na região Norte de Portugal. Estas observações foram recolhidas no período entre 20 de fevereiro de 2019 e 11 de outubro de 2019, perfazendo um total de 234 registos; a segunda base de dados é referente às previsões provenientes do *website weatherstack* com horizonte temporal até 6 dias relativas às mesmas variáveis meteorológicas: temperatura máxima do ar (°C) e temperatura mínima do ar (°C).

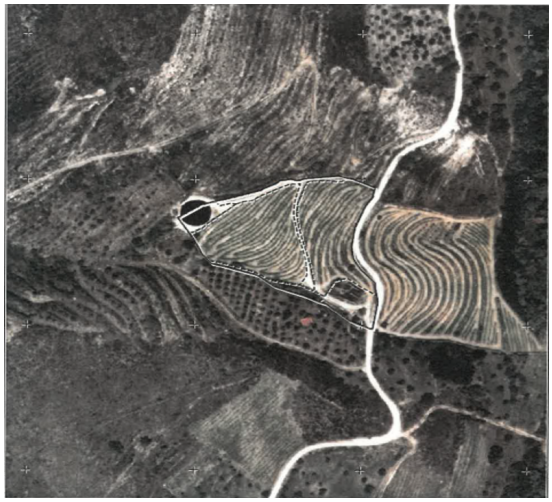


Figura 7.1: Quinta Senhora da Ribeira, Bragança.

A base de dados correspondente às observações registadas na quinta apresenta valores em falta nas duas variáveis meteorológicas desde 24/04/2019 até 01/05/2019 e de 29/09/2019 a 30/09/2019. Esta situação é bastante comum em dados reais devido a diversos fatores como, por exemplo, avaria dos dispositivos de observação como sensores na estação meteorológica, perda de dados, entre outros. Lidar com séries temporais com valores em falta pode ser bastante problemático e, por vezes, podem induzir a enviesamentos no modelo de previsão. Além disso, grande parte dos algoritmos são projetados para dados completos. Para lidar com este problema, uma solução passa por fazer a imputação dos dados, que consiste em preencher esses valores ausentes com base em algumas propriedades dos dados.

7.2 Análise Exploratória dos Dados

7.2.1 Temperatura Máxima

Na Figura 7.2 estão representadas as séries temporais da temperatura máxima observada juntamente com as respetivas previsões com horizonte temporal de 1 até 6 dias. É possível notar que a série da temperatura máxima observada (linha

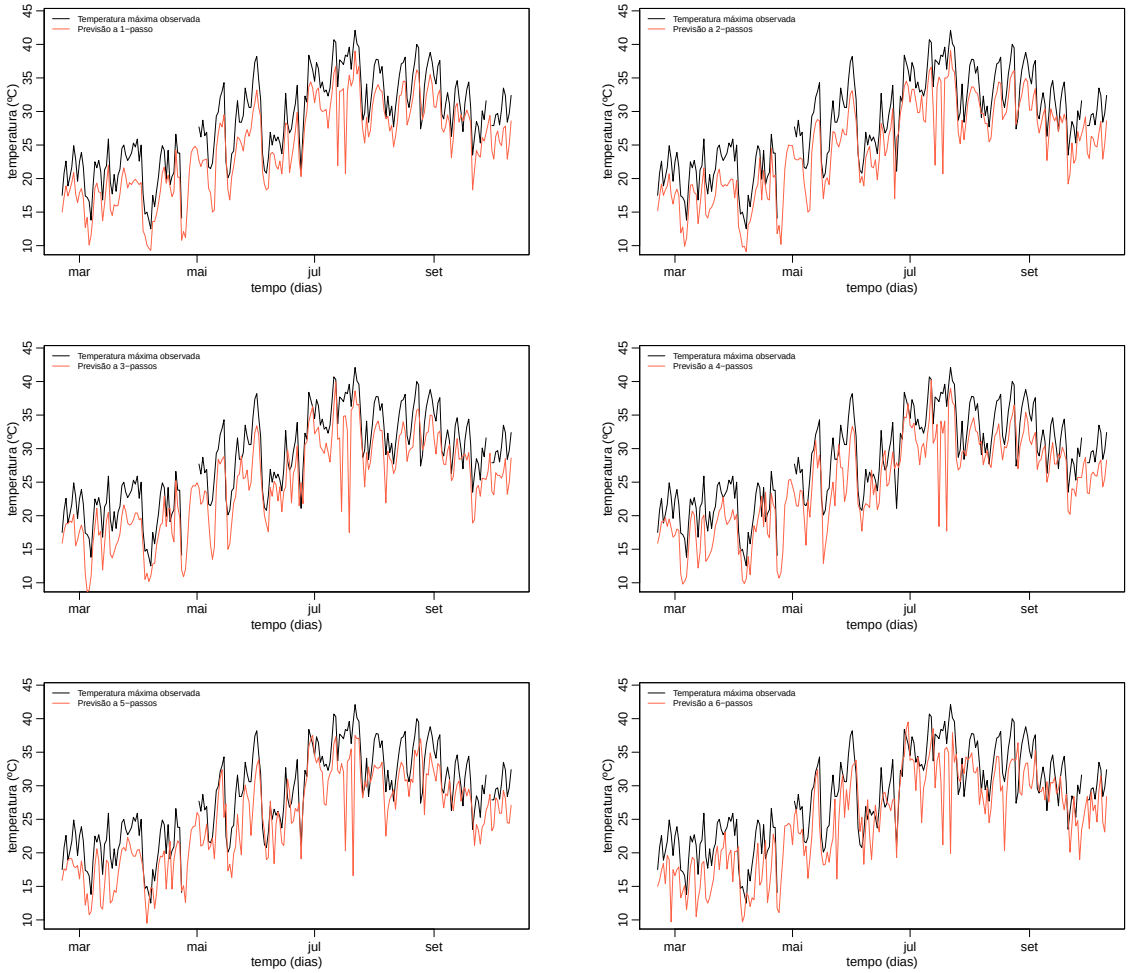


Figura 7.2: Séries da temperatura máxima observada e das respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias provenientes do *website*.

a preto) apresenta valores em falta. Como referido na secção 8.1, estão em falta as observações desde 24/04/2019 até 01/05/2019 e de 29/09/2019 a 30/09/2019. Para completar a série, isto é, imputar os valores ausentes, adotou-se um método simples de interpolação linear através do ajustamento de modelos de RLS aos dados: para completar os dados entre 24/04/2019 e 01/05/2019, inclusive, ajustou-se uma reta de regressão aos dados compreendidos entre 01/04/2019 e 31/05/2019; para completar os dados dos dias 29/09/2019 e 30/09/2019, o procedimento foi o mesmo, mas ajustando o modelo aos dados compreendidos entre 01/09/2019 e 11/10/2019. Nestes modelos, a variável resposta corresponde à temperatura máxima observada na quinta em Carrazeda de Ansiães e a variável independente corresponde à temperatura máxima observada em Vila Real, nomeadamente

$$T_{t,CA}^{M} = \alpha + \beta T_{t,VR}^{M} + a_t,$$

onde $T_{t,CA}^{M}$ e $T_{t,VR}^{M}$ representam as temperaturas máximas em Carrazeda de Ansiães e Vila Real, respetivamente. A escolha desta variável independente deve-se à proximidade geográfica entre Vila Real e Bragança.

Analisando a Figura 7.3, constata-se uma pequena diferença entre os *box plots* da temperatura máxima observada e das previsões, cuja mediana da temperatura máxima observada é ligeiramente superior às das previsões.

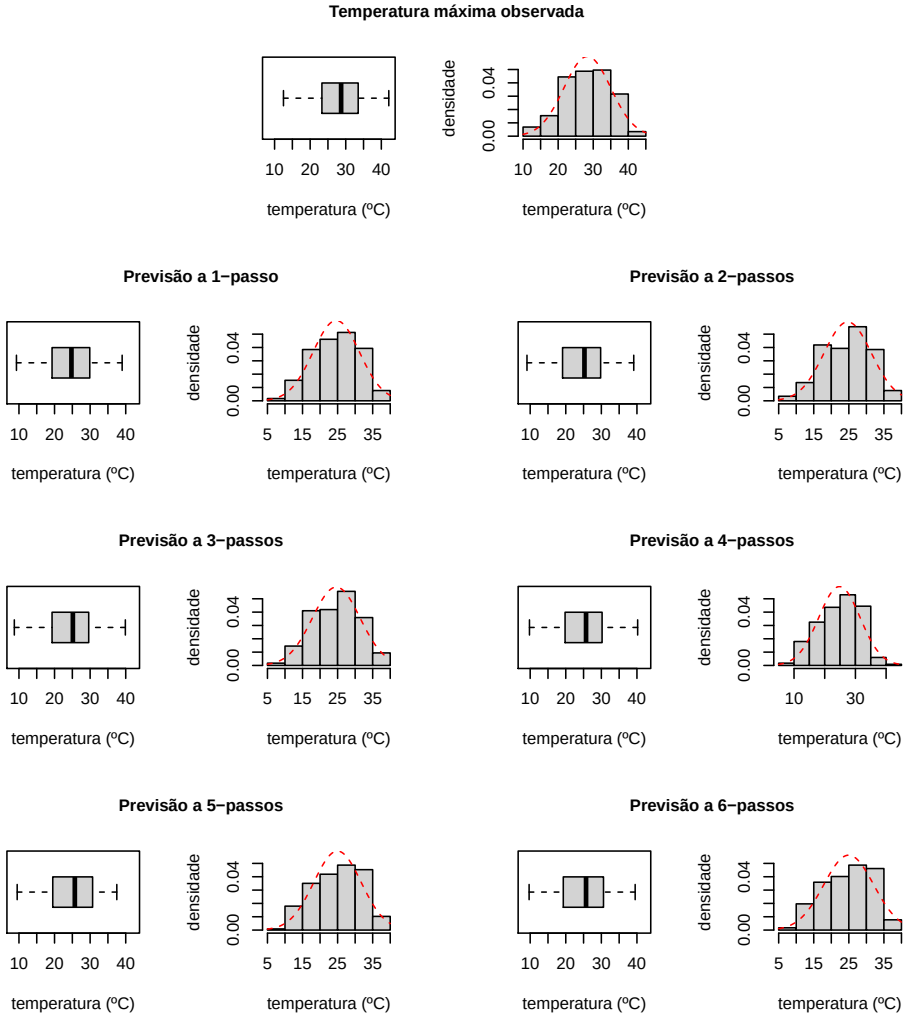


Figura 7.3: *Box plots* e histogramas da temperatura máxima observada e das respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias provenientes do *website*.

Também é possível observar que nenhuma variável em estudo possui *outliers*. Relativamente aos histogramas, sugerem que as variáveis são aproximadamente si-

métricas, cujas distribuições assemelham-se à da distribuição Normal, podendo ser comprovado através do teste estatístico Kolmogorov-Smirnov (ver Tabela A.1, Apêndice A).

Comparando as estatísticas descritivas presentes na Tabela 7.1, é notória a discrepância entre a temperatura máxima observada, denotada por Y_t^M, e as respetivas previsões, denotadas por $W_{t,(h)}^M$, $h = 1, \ldots, 6$ dias. Por exemplo, o mínimo de Y_t^M é de 12,50°C, enquanto que das previsões são de aproximadamente 9°C (uma diferença de 3°C, aproximadamente). O mesmo acontece com o máximo. A média de Y_t^M é de 28,45°C, enquanto que das previsões, $W_{t,(h)}^M$, a média anda em torno de 24,8°C.

Tabela 7.1: Estatísticas descritivas da temperatura máxima observada, Y_t^M, e das respetivas previsões a h-passos, $W_{t,(h)}^M$, $h = 1, \ldots, 6$ dias do *website*.

	Y_t^M	$W_{t,(1)}^M$	$W_{t,(2)}^M$	$W_{t,(3)}^M$	$W_{t,(4)}^M$	$W_{t,(5)}^M$	$W_{t,(6)}^M$
mínimo	12,50	9,30	9,10	8,70	9,80	9,50	9,70
1.º quartil	23,45	19,32	19,12	19,32	19,80	19,50	19,30
mediana	28,70	24,75	25,25	25,15	25,70	25,70	25,70
média	28,45	24,61	24,67	24,61	24,80	24,96	24,96
3.º quartil	33,50	29,90	29,77	29,60	30,18	30,68	30,60
máximo	42,10	39,00	39,10	39,90	40,20	37,50	39,50
desvio padrão	6,65	6,62	6,74	6,79	6,74	6,73	7,10
coeficiente de variação (%)	23,38	26,90	27,33	27,59	27,19	26,97	28,44
dados omissos	10	0	0	0	0	0	0

Os dados correspondentes à temperatura máxima observada parecem ser mais homogéneos do que as previsões, uma vez que possuem menor coeficiente de variação (23,38°C). O 1.º quartil de Y_t^M é de 23,45°C, o que significa que cerca de 25% dos dias, no período de observação, apresentaram temperaturas máximas inferiores a 23,45°C.

De acordo com a Figura 7.4, existem evidências de uma forte relação linear entre a temperatura máxima observada e as respetivas previsões para os diferentes horizontes temporais até 6 dias, uma vez que todas as correlações são superiores a 0,80.

De facto, os valores de prova do teste de correlação linear de Pearson são inferiores ao nível de significância (5%), levando à rejeição da hipótese nula da inexistência de correlação linear entre as variáveis. Por outro lado, à medida que o horizonte temporal aumenta, a correlação entre as previsões e a temperatura máxima observada diminui, o que é bastante intuitivo.

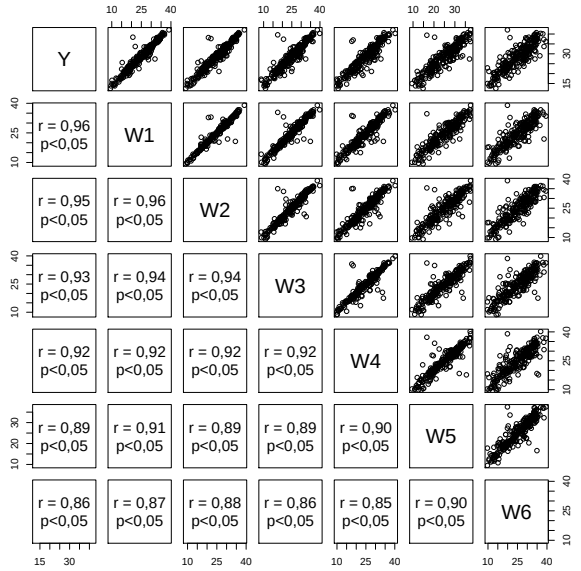


Figura 7.4: Coeficientes de correlação linear de Pearson (r) e valores de prova (p) entre a temperatura máxima observada, Y_t^M, e as respetivas previsões a h-passos, $W_{t,(h)}^M$, $h = 1, \ldots, 6$ dias.

7.2.2 Temperatura Mínima

Na Figura 7.5 estão representadas as séries da temperatura mínima observada com as respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias. Analogamente à série da temperatura máxima, a série da temperatura mínima observada (linha a preto) apresenta valores em falta desde 24/04/2019 até 01/05/2019 e de 29/09/2019 a 30/09/2019. Para completar a série, procedeu-se com a mesma abordagem que foi feita na temperatura máxima, onde foram ajustados modelos de RLS aos dados: para completar os dados entre 24/04/2019 e 01/05/2019, inclusive, ajustou-se uma reta de regressão aos dados compreendidos entre 01/04/2019 e 31/05/2019; para completar os dados dos dias 29/09/2019 e 30/09/2019, o procedimento foi o mesmo, mas ajustando o modelo aos dados compreendidos entre 01/09/2019 e 11/10/2019. Mais uma vez, nestes modelos, a variável resposta corresponde à temperatura mínima observada na quinta e a variável independente corresponde à temperatura mínima observada em Vila Real, nomeadamente

$$T_{t,CA}^m = \alpha + \beta T_{t,VR}^m + a_t,$$

onde $T_{t,CA}^m$ e $T_{t,VR}^m$ representam as temperaturas mínimas em Carrazeda de Ansiães e Vila Real, respetivamente.

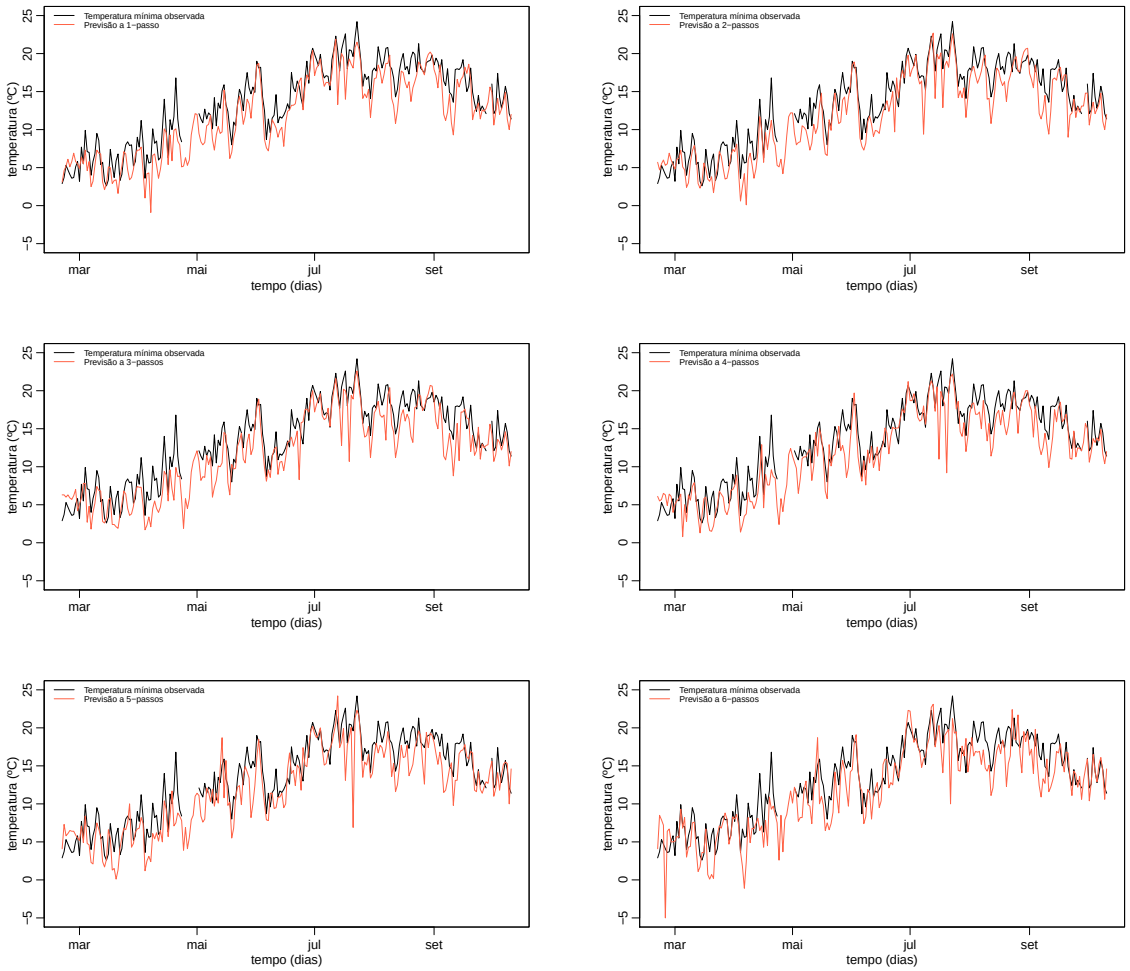


Figura 7.5: Séries da temperatura mínima observada e das respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias provenientes do *website*.

Na Figura 7.6 encontram-se representados os *box plots* e os respetivos histogramas da variável temperatura mínima observada e das previsões para os diferentes horizontes temporais até 6 dias.

Por observação gráfica, deteta-se uma pequena diferença entre os *box plots* da temperatura mínima observada e das previsões, onde a mediana da temperatura mínima observada é ligeiramente superior às das previsões. Também é possível observar que nenhuma variável em estudo possui *outliers*.

Relativamente aos histogramas, parece não existir uma relação discernível entre o aumento do horizonte temporal e a simetria dos histogramas.

Para testar a normalidade destas variáveis, recorreu-se ao teste estatístico Kolmogorov-Smirnov (ver Tabela A.2, Apêndice A).

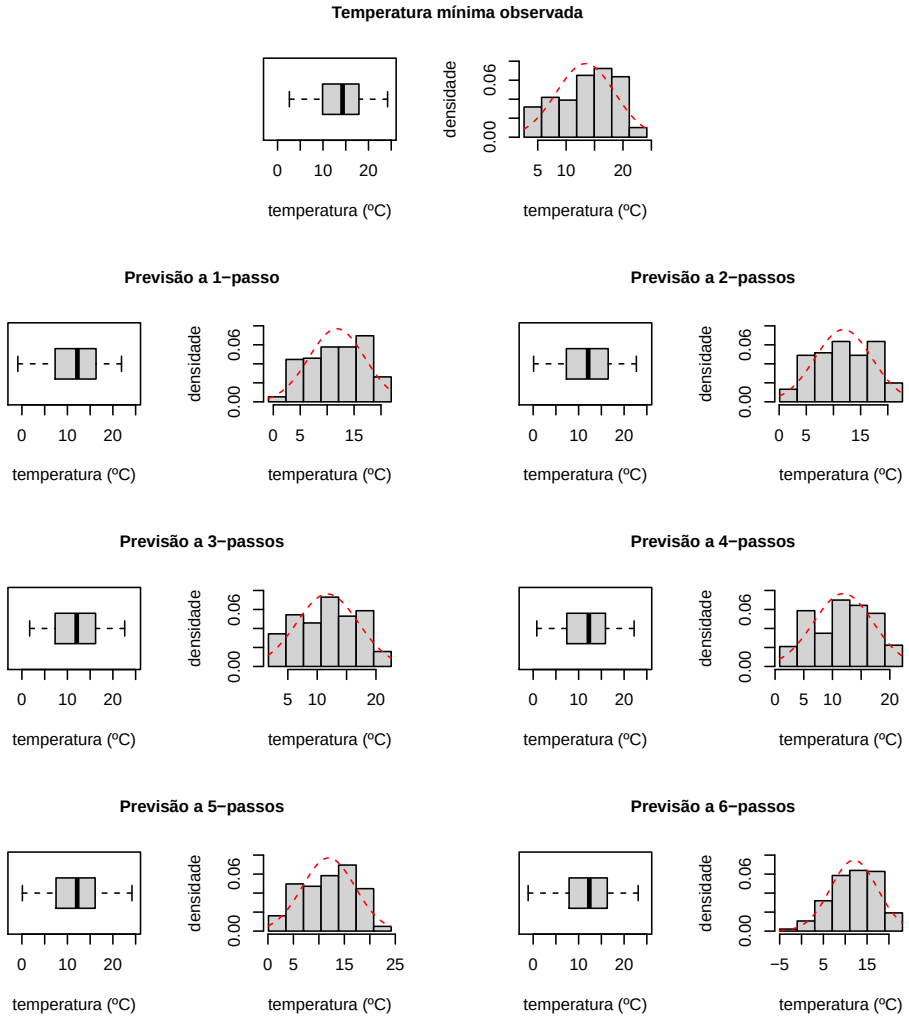


Figura 7.6: *Box plots* e histogramas da temperatura mínima observada e das respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias provenientes do *website*.

De acordo com as estatísticas descritivas presentes na Tabela 7.2 relativas à temperatura mínima observada, denotada por Y_t^m, e às respetivas previsões a h-passos, representadas por $W_{t,(h)}^m$, $h = 1, \ldots, 6$ dias, é possível observar que o valor mínimo de Y_t é bem diferente das respetivas previsões, sendo o valor mínimo observado corresponde à previsão a 3-passos (1,70°C) que se aproxima mais do valor mínimo observado. No geral, também é notória a discrepância entre a temperatura mínima observada e as respetivas previsões. A variável correspondente à previsão a 6-passos apresenta maior dispersão relativa em relação à média, uma vez que é a variável que possui maior coeficiente de variação (44,56%). O 1.º quartil de Y_t^m é de 9,90°C, o que significa que cerca de 25% dos dias, no período de observação, apresentam temperaturas mínimas inferiores a 9,90°C.

Analisando a Figura 7.7, existem evidências de uma forte relação linear entre a

Tabela 7.2: Estatísticas descritivas da temperatura mínima observada, Y_t^m, e das respetivas previsões a h-passos, $W_{t,(h)}^m$, $h = 1, \ldots, 6$ dias do *website*.

	Y_t^m	$W_{t,(1)}^m$	$W_{t,(2)}^m$	$W_{t,(3)}^m$	$W_{t,(4)}^m$	$W_{t,(5)}^m$	$W_{t,(6)}^m$
mínimo	2,60	-0,90	0,10	1,70	0,80	0,10	-5,00
1.º quartil	9,90	7,30	7,30	7,30	7,43	7,50	7,93
mediana	14,30	12,20	12,10	12,05	12,25	12,15	12,35
média	13,52	11,80	11,84	11,76	11,90	11,92	12,00
3.º quartil	17,90	16,27	16,50	16,15	15,90	16,10	16,28
máximo	24,20	21,90	22,70	22,60	22,20	24,20	23,10
desvio padrão	5,21	5,19	5,25	5,23	5,20	5,18	5,35
coeficiente de variação (%)	38,55	44,00	44,30	44,46	43,73	43,47	44,56
dados omissos	10	0	0	0	0	0	0

temperatura mínima observada e as respetivas previsões para os diferentes horizontes temporais. Os valores de prova do teste de correlação linear de Pearson são inferiores ao nível de significância (5%), o que leva à rejeição da hipótese nula da ausência de correlação linear entre as variáveis. Todas as correlações são superiores a 0,80. Observa-se também que as correlações entre as previsões e a temperatura mínima observada diminuem com o aumento do horizonte temporal.

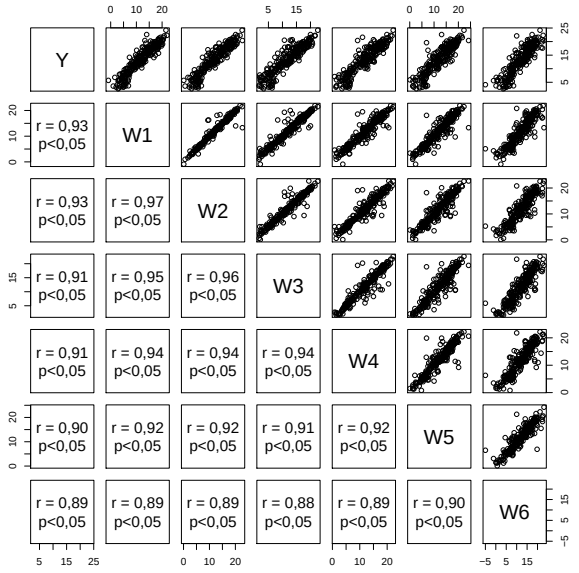


Figura 7.7: Coeficientes de correlação linear de Pearson (r) e valores de prova (p) entre a temperatura mínima observada, Y_t^m, e as respetivas previsões a h-passos, $W_{t,(h)}^m$, $h = 1, \ldots, 6$ dias.

7.3 Aplicação dos Modelos de Calibração às Séries das Variáveis Meteorológicas

Modelo de regressão linear simples

Como ponto de partida no processo de modelação, foram considerados modelos de Regressão Linear Simples Com constante aditiva (RLSC_h), dado por

$$Y_t = \alpha + \beta W_{t,(h)} + e_t,$$

e modelos de Regressão Linear Simples Sem constante aditiva (RLSS_h), dado por

$$Y_t = \beta W_{t,(h)} + e_t,$$

para $t = 1, \ldots, n$, onde Y_t é a variável da temperatura máxima/mínima observada no dia t, $W_{t,(h)}$ corresponde à previsão a h-passos proveniente do *website weatherstack*, onde h representa o horizonte temporal, com $h = 1, \ldots, 6$ dias, α (no caso de RLSC_h) e β são os parâmetros de regressão (constantes) e o erro e_t é uma variável aleatória que representa o erro aleatório.

Ambos os modelos de RLS consideram os parâmetros determinísticos ao longo do tempo e, por vezes não conseguem captar a dinâmica natural presente nos dados de séries temporais meteorológicas. Os MEE surgem nesse sentido, pois são modelos flexíveis devido à sua natureza estocástica dada a sua estrutura dinâmica.

Modelos de calibração baseados na representação de espaço de estados

Os modelos de calibração são baseados na representação de espaço de estados que permitem lidar com uma variedade de problemas na análise de séries temporais, pois são processos que possuem dependência temporal e, além disso, precisam de alguma dinâmica estocástica.

Nesta abordagem, foram considerados os Modelos de Espaço de Estados Sem constante aditiva (MEES_h) que permitem relacionar os valores observados da temperatura máxima/mínima, Y_t, com as respetivas previsões obtidas do *website weatherstack*, $W_{t,(h)}$, dadas pelas equações

$$Y_t = \beta_t W_{t,(h)} + e_t, \tag{7.1}$$

$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t, \tag{7.2}$$

para $t = 1, \ldots, n$, onde Y_t é a variável da temperatura máxima/mínima observada no dia t e $W_{t,(h)}$ é a previsão h-passos, registada no tempo $t-h$, com $h = 1, \ldots, 6$ dias. Assume-se que:

- o processo $\{\beta_t\}_{t=1,\ldots,n}$ segue um processo autorregressivo de ordem 1, ou seja, $\{\beta_t\}_{t=1,\ldots,n} \sim AR(1)$ e $|\phi| < 1$;
- o processo $\{\beta_t\}_{t=1,\ldots,n}$ tem média μ, $E(\beta_t) = \mu$, e variância $var(\beta_t) = \dfrac{\sigma_\varepsilon^2}{1-\phi^2}$;
- e_t é o erro aleatório que corresponde a um processo de ruído branco, ou seja, $E(e_t) = 0$ e $E(e_t e_s) = 0$, $\forall t, s \in \{1, \ldots, n\}, t \neq s$, com distribuição $e_t \sim N(0, \sigma_e^2)$;
- ε_t é o erro aleatório que corresponde a um processo de ruído branco, ou seja, $E(\varepsilon_t) = 0$ e $E(\varepsilon_t \varepsilon_s) = 0$, $\forall t, s \in \{1, \ldots, n\}, t \neq s$, com distribuição $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$;
- os erros e_t e ε_t são não correlacionados, ou seja, $E(e_t \varepsilon_s) = 0, \forall t, s \in \{1, \ldots, n\}$.

Neste modelo, quando $\mu > 1$, pode-se concluir que as previsões do *website* tendem a substimar, em média, a variável meteorológica observada correspondente. Caso contrário, quando $\mu < 1$, pode-se concluir que as previsões do *website* sobrestimam a variável meteorológica observada correspondente.

Nesta dissertação também considerou-se o Modelo de Espaço de Estados Com constante aditiva (MEEC$_h$), dado por

$$Y_t = \alpha + \beta_t W_{t,(h)} + e_t, \tag{7.3}$$

$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t, \tag{7.4}$$

para $t = 1, \ldots, n$, onde Y_t é a variável da temperatura máxima/mínima observada no dia t e $W_{t,(h)}$ é a respetiva previsão a h-passos, registada no tempo $t-h$, com $h = 1, \ldots, 6$ dias, α e β_t são os parâmetros da equação de observação, onde α é determinístico e β_t é o declive estocástico.

A diferença entre os modelos de calibração definidos pelas equações (7.1) - (7.2) e (7.3) - (7.4) está na componente aditiva α, que neste caso representa o erro sistemático, assumida como constante, cujo efeito do valor esperado é descrito por um efeito aditivo. Ou seja, nesta situação, o fator de calibração não coincide com o rácio entre os valores observados, Y_t, e os valores previstos, $W_{t,(h)}$, mas sim com o rácio entre a diferença dos valores observados e do valor α e os valores previstos.

Os pressupostos listados para o modelo de calibração dadas pelas equações (7.1) e (7.2) também são válidos para o modelo definido pelas equações (7.3) e (7.4). O objetivo da aplicação deste modelo aos dados é compará-lo com o modelo de calibração dadas pelas equações (7.1) e (7.2), a fim de verificar se a adição de uma componente estocástica na equação de observação melhoraria em termos de ajustamento.

Nas Figuras 7.8 e 7.9 estão representados os *box plots* dos rácios entre os valores observados e as respetivas previsões correspondentes à temperatura máxima e à temperatura mínima, respetivamente, para os diferentes horizontes temporais.

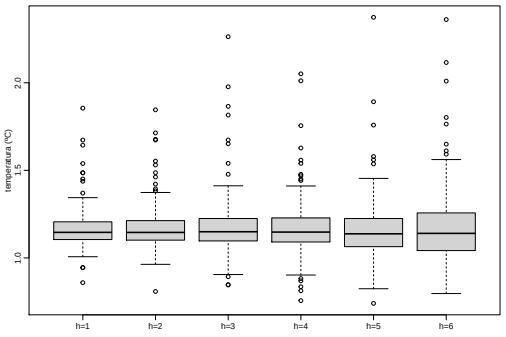


Figura 7.8: *Box plots* do rácio entre a temperatura máxima observada e as respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias.

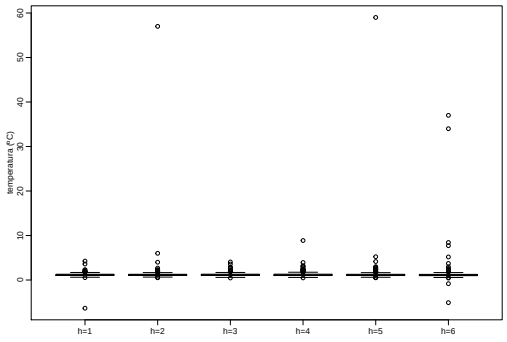


Figura 7.9: *Box plots* do rácio entre a temperatura mínima observada e as respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias.

Através da análise gráfica, percebe-se que existem previsões que diferem a larga amplitude dos valores observados, dada presença de *outliers* relativos ao rácio entre os valores observados e os valores previstos do *website weatherstack*.

Na Figura 7.8 é possível observar o efeito que o horizonte temporal tem sobre a precisão das previsões da temperatura máxima, dado que a dispersão aumenta com o aumento do horizonte temporal, o que é bastante intuitivo visto que a incerteza associada às previsões tende a aumentar com o aumento do horizonte temporal. Também é percetível que as medianas não se diferem significativamente, assumindo um valor de aproximadamente 1,2°C.

Por outro lado, na Figura 7.9, observam-se *outliers* bastante discrepantes. Isto acontece uma vez que, como correspondem a rácios entre temperaturas mínimas, estão envolvidos valores inferiores à unidade e portanto geram rácios de valores elevados.

Perante esta situação, procedeu-se, então, pela remoção dos *outliers* dos rácios e substituir os valores em falta por estimativas através da interpolação linear (ver Tabelas A.3 e A.4, Apêndice A).

Decidiu-se aplicar o modelo de calibração definido pelas equações (7.3) e (7.4),

apesar do rácio entre os valores observados e os valores previstos no *website* não coincidirem com o fator de calibração deste modelo. Assim, torna-se possível comparar vários modelos uma vez que são aplicados sobre a mesma base de dados.

A interpolação linear é um método que calcula o valor aproximado de $f(x)$ através de uma função linear dada por (Samarin, 2012)

$$L(x) = \alpha(x - x_1) + \beta,$$

onde os parâmetros α e β são escolhidos de forma a que os valores de $L(x)$ sejam iguais aos valores de $f(x)$, dados os pontos x_1 e x_2 que são conhecidos, ou seja,

$$L(x_1) = f(x_1), \qquad L(x_2) = f(x_2).$$

Estas condições são satisfeitas por

$$L(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) + f(x_1),$$

que aproxima a função $f(x)$ no intervalo $[x_1, x_2]$

$$f(x) - L(x) = \frac{f''(\xi)}{2}(x - x_1)(x - x_2), \quad \xi \in [x_1, x_2].$$

A substituição dos valores em falta por estimativas foi realizada recorrendo à função *na.interp()*, da autoria de R. Hyndman, implementada no package *forecast* do *software* R. Esta função faz interpolação linear para séries temporais não sazonais. Caso a série seja sazonal, é feita uma decomposição STL[1]. Na primeira etapa, a componente sazonal é removida da série temporal, de seguida é feita uma interpolação linear para imputar os valores e posteriormente a componente sazonal é adicionada novamente (Moritz *et al.*, 2015).

Moritz *et al.* (2015) compararam experimentalmente seis funções de imputação do *software* R, onde foram consideradas seis séries temporais. O primeiro passo consistiu em excluir valores do conjunto de dados usando quatro proporções diferentes de valores em falta. De seguida foram aplicadas as funções de imputação e avaliaram o seu desempenho através de duas medidas de avaliação (REQM e EPAM) para os valores imputados e concluíram que, das implementações testadas, a função

[1] A decomposição STL (*Seasonal-Trend decomposition procedure based on Loess*) é um método que decompõe uma série temporal em três componentes: tendência, a componente sazonal e o ruído.

na.interp() foi uma das que mostrou os melhores resultados.

7.3.1 Temperatura Máxima

Na Figura 7.10 estão representadas as séries temporais da temperatura máxima observada e as respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias, provenientes do *website weatherstack*. De um modo geral, é notável a subestimação das previsões em relação à temperatura máxima observada, sendo bastante evidente no mês de julho, onde as previsões chegam a discernir cerca de 15°C, dando indícios de possíveis *outliers*.

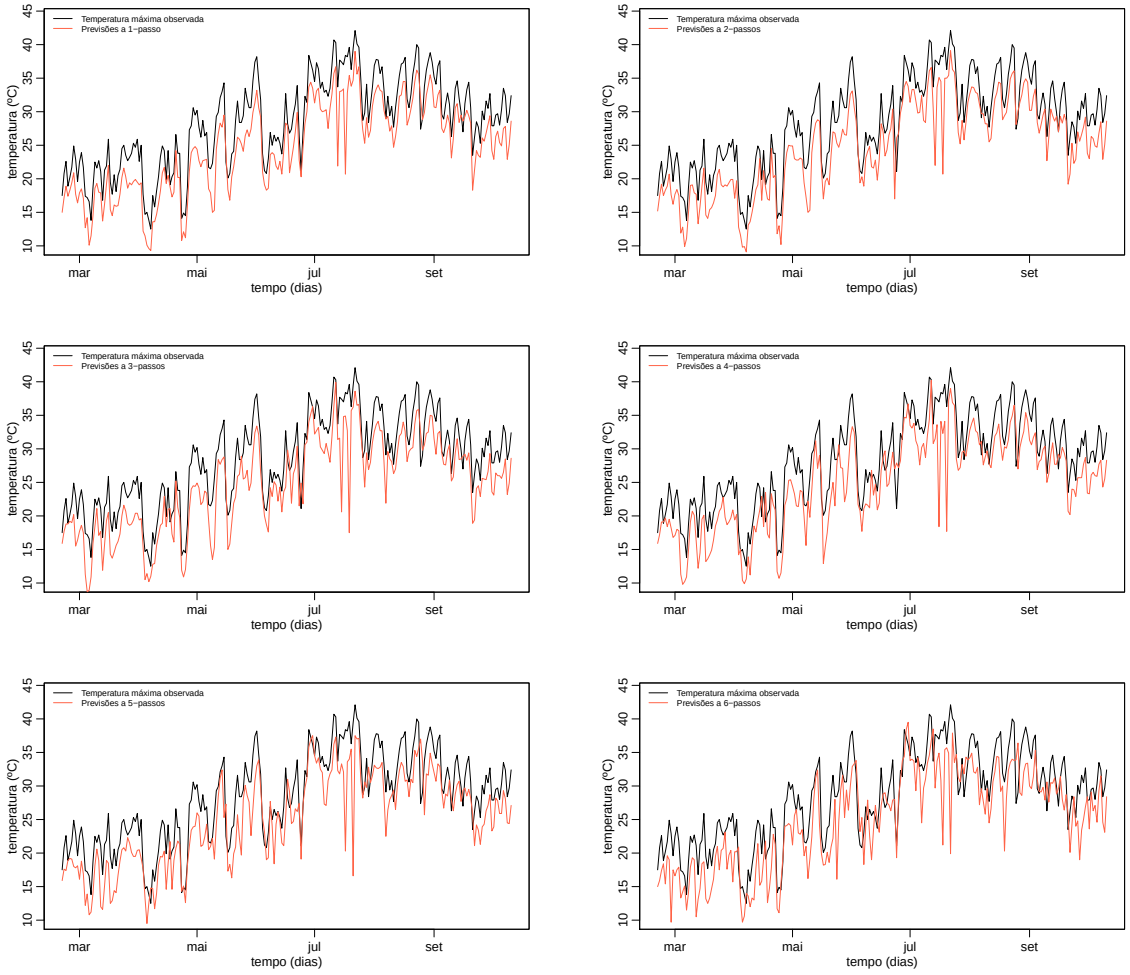


Figura 7.10: Séries da temperatura máxima observada e das respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias.

Analisando a Tabela 7.3, é possível tirar algumas conclusões

- Nos quatro modelos, as estimativas e os erros padrão de σ_e tendem a aumentar com o aumento do horizonte temporal. Por outro lado, a log-verosimilhança tende a diminuir;

- Para ambos os modelos de calibração com o fator de calibração estocástico, MEES_h e MEES_h^*, os valores estimados de ϕ estão muito próximos da unidade (um processo $AR(1)$ é estacionário se $|\phi| < 1$);

- As estimativas do valor médio, no modelo MEES_h, do processo $\{\beta_t\}_{t=1,\ldots,n}$, μ, são de aproximadamente 1,13°C, coincidindo com as estimativas de β no modelo de regressão linear RLSS_h. Isto indica que, em média, a temperatura máxima observada é cerca de 13% superior em relação às previsões. Por outro lado, as estimativas e os erros padrão de σ_e são mais elevados no modelo de regressão linear. As mesmas conclusões podem ser tiradas comparando os modelos RLSS_h^* e MEES_h^*.

Na Tabela 7.4 estão apresentadas algumas medidas de avaliação e critérios de seleção referentes aos quatro modelos, cujas estimativas dos parâmetros encontram-se na Tabela 7.3. No geral, percebe-se que

- Nos quatro modelos, os critérios AIC e BIC tendem a aumentar com o aumento do horizonte temporal, assim como as medidas REQM, EAM, EEAM e U-Theil. O coeficiente de determinação, r^2, e o coeficiente de determinação ajustado, r_a^2, tendem a diminuir;

- O modelo MEES_h produziu os menores valores de AIC e BIC do que o modelo RLSS_h, assim como obteve menores valores de REQM, EAM, EEAM e U-Theil. No entanto, o modelo mais simples (RLSS_h) obteve um melhor ajustamento, visto que tem menores valores de r_a^2;

- O modelo MEES_h^* obteve menores valores para a REQM, EAM, EEAM e U-Theil, assim como produziu menores valores para os critérios de seleção AIC e BIC. No entanto, o modelo RLSS_h^* obteve valores mais elevados de r_a^2 para $h = 1, 3, 4, 5$ dias.

Tabela 7.3: Estimativas dos parâmetros e respetivos erros padrão dos quatro modelos para a série da temperatura máxima, onde Y_t^M representa a temperatura máxima observada, $W_{t,(h)}^M$ corresponde às previsões a h-passos do *website* e $W_{t,(h)}^{M*}$ corresponde às previsões com a substituição dos *outliers* do rácio $Y_t^M/W_{t,(h)}^M$, $h = 1, \ldots, 6$ dias.

		h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \beta W_{t,(h)}^M + e_t$							
β	estimativa	1,1390	1,1342	1,1345	1,1267	1,1180	1,1106
	erro padrão	0,0055	0,0062	0,0075	0,0077	0,0088	0,0103
σ_e	estimativa	2,1591	2,4107	2,9351	3,0311	3,4642	4,0962
	erro padrão	0,1398	0,1559	0,1899	0,1964	0,2247	0,2648
	logL	-511,6332	-537,4309	-583,4917	-591,0225	-622,2736	-661,4873
$Y_t^M = \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,9859	0,9861	0,9849	0,9796	0,9080	0,9387
	erro padrão	0,0115	0,0097	0,0144	0,0196	0,0758	0,0545
μ	estimativa	1,1297	1,1247	1,1322	1,1187	1,1316	1,1299
	erro padrão	0,0387	0,0382	0,0392	0,0404	0,0183	0,0270
σ_ε	estimativa	0,0056	0,0060	0,0060	0,0058	0,0212	0,0225
	erro padrão	0,0032	0,0027	0,0043	0,0086	0,0102	0,0112
σ_e	estimativa	1,9285	2,1574	2,7785	2,8852	3,2421	3,8000
	erro padrão	0,0956	0,1045	0,1338	0,1569	0,1726	0,2017
	logL	-279,0373	-304,9940	-362,5232	-370,1581	-404,1322	-441,4972
$Y_t^M = \beta W_{t,(h)}^{M*} + e_t$							
β	estimativa	1,1320	1,1258	1,1295	1,1219	1,1138	1,1037
	erro padrão	0,0044	0,0048	0,0056	0,0061	0,0073	0,0087
σ_e	estimativa	1,7234	1,9078	2,1840	2,3983	2,8929	3,4805
	erro padrão	0,1116	0,1236	0,1415	0,1556	0,1879	0,2254
	logL	-458,8993	-482,6831	-514,3246	-536,2211	-580,0963	-623,3697
$Y_t^M = \beta_t W_{t,(h)}^{M*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,9795	0,9584	0,9829	0,9817	0,9821	0,6063
	erro padrão	0,0221	0,0361	0,0145	0,0150	0,0086	0,1922
μ	estimativa	1,1407	1,1441	1,1350	1,1253	1,0966	1,1253
	erro padrão	0,0309	0,0261	0,0328	0,0320	0,0274	0,0152
σ_ε	estimativa	0,0083	0,0161	0,0073	0,0075	0,0036	0,0745
	erro padrão	0,0051	0,0071	0,0032	0,0034	0,0050	0,0273
σ_e	estimativa	1,4004	1,4099	1,9392	2,1650	2,7347	2,5573
	erro padrão	0,0861	0,1065	0,0962	0,1065	0,1363	0,4305
	logL	-211,8309	-227,9199	-282,5388	-307,3497	-356,2259	-398,8004

Tabela 7.4: Comparação das medidas e de critérios de seleção dos quatro modelos para a série da temperatura máxima.

	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \beta W_{t,(h)}^M + e_t$						
REQM	2,1544	2,4055	2,9289	3,0247	3,4568	4,0875
EAM	1,4723	1,6694	2,0048	2,1331	2,4793	3,0681
EEAM	0,5815	0,6594	0,7919	0,8425	0,9793	1,2119
U-Theil	0,6066	0,6453	0,7380	0,7126	0,7446	0,8323
r^2	**0,9257**	**0,9094**	**0,8616**	**0,8452**	**0,7914**	**0,7363**
r_a^2	**0,9257**	**0,9094**	**0,8616**	**0,8452**	**0,7914**	**0,7363**
AIC	1027,2660	1078,8620	1170,9830	1186,0450	1248,5470	1326,9750
BIC	1034,1770	1085,7720	1177,8940	1192,9560	1255,4580	1333,8850
$Y_t^M = \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,9754**	**2,2060**	**2,8398**	**2,9316**	**3,3927**	**3,9831**
EAM	**1,2459**	**1,3745**	**1,8716**	**1,9971**	**2,4365**	**2,9679**
EEAM	**0,4921**	**0,5429**	**0,7392**	**0,7888**	**0,9624**	**1,1723**
U-Theil	**0,5003**	**0,5371**	**0,6771**	**0,6470**	**0,6746**	**0,7368**
r^2	0,9228	0,9038	0,8524	0,8353	0,7874	0,7250
r_a^2	0,9228	0,9038	0,8524	0,8353	0,7874	0,7250
AIC	**566,0746**	**617,9880**	**733,0464**	**748,3162**	**816,2644**	**890,9944**
BIC	**579,8959**	**631,8093**	**746,8677**	**762,1375**	**830,0857**	**904,8157**
$Y_t^M = \beta W_{t,(h)}^{M*} + e_t$						
REQM	1,7197	1,9037	2,1794	2,3931	2,8867	3,4731
EAM	1,3037	1,4536	1,7106	1,8685	2,2720	2,7737
EEAM	0,5149	0,5742	0,6757	0,7380	0,8974	1,0956
U-Theil	0,5897	0,6611	0,6595	0,7599	0,8503	0,9520
r^2	**0,9527**	0,9414	**0,9222**	**0,8994**	**0,8503**	0,7986
r_a^2	**0,9527**	0,9414	**0,9222**	**0,8994**	**0,8503**	0,7986
AIC	921,7986	969,3662	1032,6492	1076,4422	1164,1926	1250,7394
BIC	928,7092	976,2768	1039,5598	1083,3528	1171,1032	1257,6500
$Y_t^M = \beta_t W_{t,(h)}^{M*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,4970**	**1,5986**	**2,0233**	**2,2522**	**2,7741**	**3,3447**
EAM	**1,1009**	**1,1783**	**1,5499**	**1,7114**	**2,1969**	**2,6521**
EEAM	**0,4348**	**0,4654**	**0,6122**	**0,6760**	**0,8677**	**1,0475**
U-Theil	**0,4533**	**0,4772**	**0,5632**	**0,6611**	**0,7737**	**0,7765**
r^2	0,9526	**0,9452**	0,9171	0,8942	0,8424	**0,8028**
r_a^2	0,9526	**0,9452**	0,9171	0,8942	0,8424	**0,8028**
AIC	**431,6618**	**463,8398**	**573,0776**	**622,6994**	**720,4518**	**805,6008**
BIC	**445,4831**	**477,6611**	**586,8989**	**636,5207**	**734,2731**	**819,4221**

Na Figura 7.11 estão representadas as séries da temperatura máxima observada com as respetivas previsões a h-passos, obtidas no *website weatherstack*, e as previsões a h-passos calibrada. A parte superior corresponde ao modelo MEES_1 e a parte inferior ao modelo MEES_6. No geral, é notória uma melhoria nas previsões calibradas em relação às previsões dadas pelo *website*, onde é possível perceber que as previsões calibradas a 1-passo sobrepõem melhor a série da temperatura máxima observada em comparação com a previsão a 6-passos calibrada. Esta situação já era de esperar visto que as previsões com menor horizonte temporal tendem a ser mais

precisas.

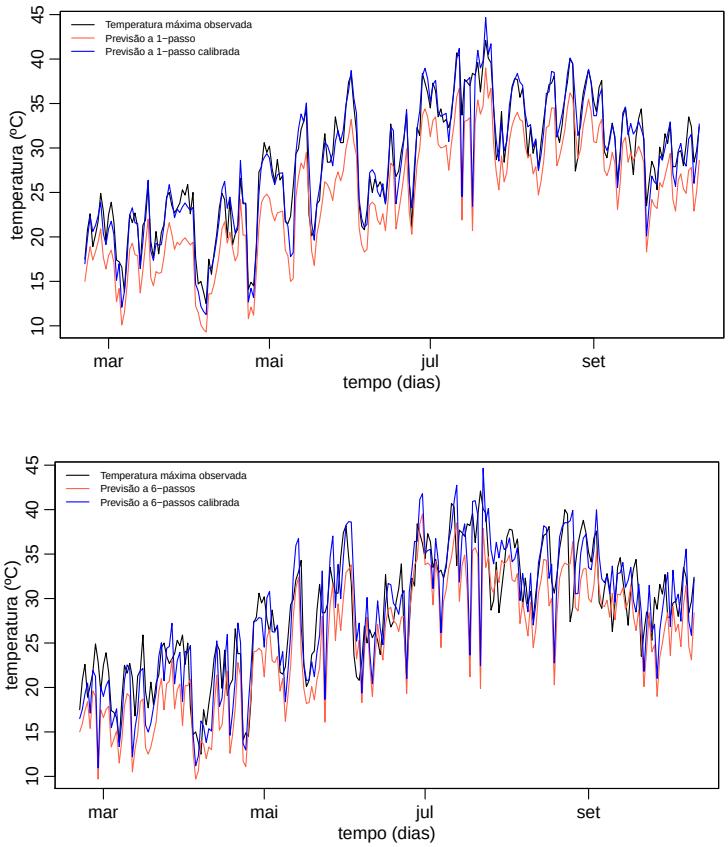


Figura 7.11: Séries da temperatura máxima observada (a preto), previsões do *website* (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEES_1, gráfico inferior - MEES_6.

Na parte superior da Figura 7.12 estão representadas as previsões do fator de calibração, $\beta_{t|t-1}$, para $t = 1, \ldots, 234$ dias, como uma linha contínua, obtidas através do FK e os limites do erro, $\beta_{t|t-1} \pm 1,96\sqrt{P_{t|t-1}}$, como linhas tracejadas. No gráfico do meio estão representados os valores filtrados, $\beta_{t|t}$, para $t = 1, \ldots, 234$ dias, como uma linha e os limites do erro, $\beta_{t|t} \pm 1,96\sqrt{P_{t|t}}$ como linhas tracejadas. No gráfico inferior estão representados, como uma linha, o fator de calibração alisado, $\beta_{t|n}$ para $t = 1, \ldots, 234$ dias e $n = 234$, obtidos através do alisamento de Kalman e os limites do erro, $\beta_{t|n} \pm 1,96\sqrt{P_{t|n}}$.

Por observação gráfica, verifica-se que o comportamento de $\beta_{t|n}$ é mais suave do que o comportamento de $\beta_{t|t-1}$ e $\beta_{t|t}$ para ambos os modelos. Isto acontece porque as estimativas obtidas a partir do alisamento são mais precisas em relação às obtidas a partir da previsão ou da filtragem, uma vez que utiliza a totalidade dos dados. Além disso, os intervalos (limites do erro) para o modelo de calibração com $h = 6$ dias (lado direito) apresentam maior variabilidade do que os intervalos para o modelo de

calibração com $h = 1$ dia (lado esquerdo). Isto deve-se ao facto que as previsões a 1-passo são mais precisas do que as previsões a 6-passos, tendo em conta de que a incerteza associada às previsões tende a aumentar com o aumento do horizonte temporal. Por outro lado, verifica-se que nos três gráficos correspondentes ao modelo a 1-passo (lado esquerdo), as estimativas para β_t oscilam entre 1,1 e 1,2, a passo que no modelo a 6-passos (lado direito), oscilam entre 1,0 e 1,2.

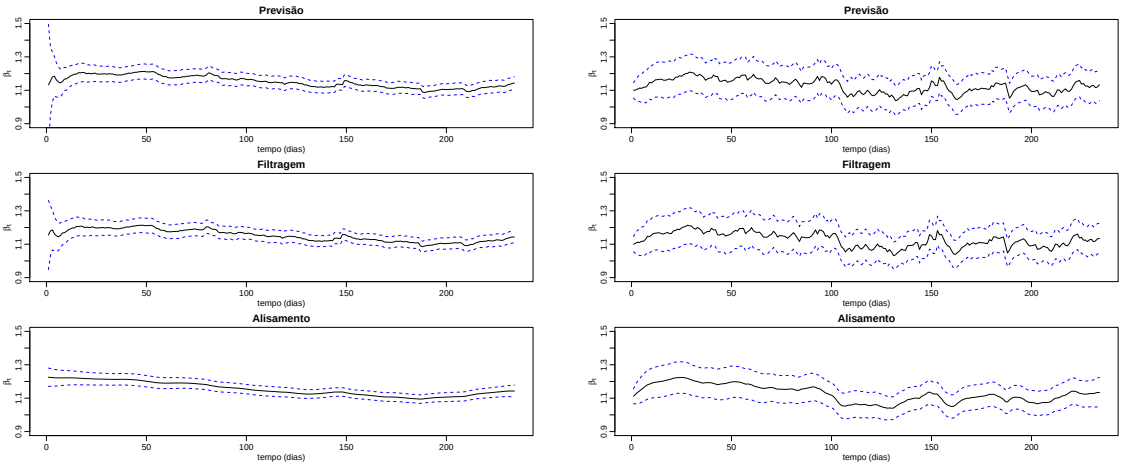


Figura 7.12: Parte superior - previsões $\beta_{t|t-1}$, centro - filtragem $\beta_{t|t}$, parte inferior- alisamento $\beta_{t|n}$ para a temperatura máxima; lado esquerdo - MEES_1, lado direito - MEES_6.

Para validar um modelo de previsão, é essencial fazer a análise das inovações (resíduos) a fim de verificar se têm um comportamento similar ao de um ruído branco, tendo que cumprir os pressupostos da normalidade, independência, média nula e variância constante (homocedasticidade) oo longo do tempo. A condição da normalidade pode ser verificada através da análise gráfica do histograma dos resíduos e através de um teste estatístico que, neste caso, o teste escolhido é o teste Kolmogorov-Smirnov, tendo em conta à dimensão da amostra ($n = 234$ observações).

Analisando o histograma dos resíduos (Figura 7.13) correspondente ao modelo MEES_1 (lado esquerdo), sugere que os resíduos apresentam uma cauda pesada à direita, sendo o suficiente para que o teste Kolmogorov-Smirnov rejeite a hipótese da normalidade dos erros (valor de prova de 0,0059). No caso no modelo de calibração MEES_6 (lado direito), sugere que os resíduos têm um comportamento próximo ao da função densidade da distribuição Normal, sendo comprovado pelo teste Kolmogorov-Smirnov (valor de prova de 0,3080). No entanto, para ambos os modelos, verifica-se que no gráfico *QQ-plot*, existem pontos nos extremos que não estão sobrepostos sobre a reta nos extremos.

O pressuposto da média nula aparenta ser válido para ambos os modelos. No

entanto, relativamente ao pressuposto da homocedasticidade dos erros, a análise gráfica sugere que, para ambos os modelos, a variância não é constante, apresentando valores mais elevados no início de julho.

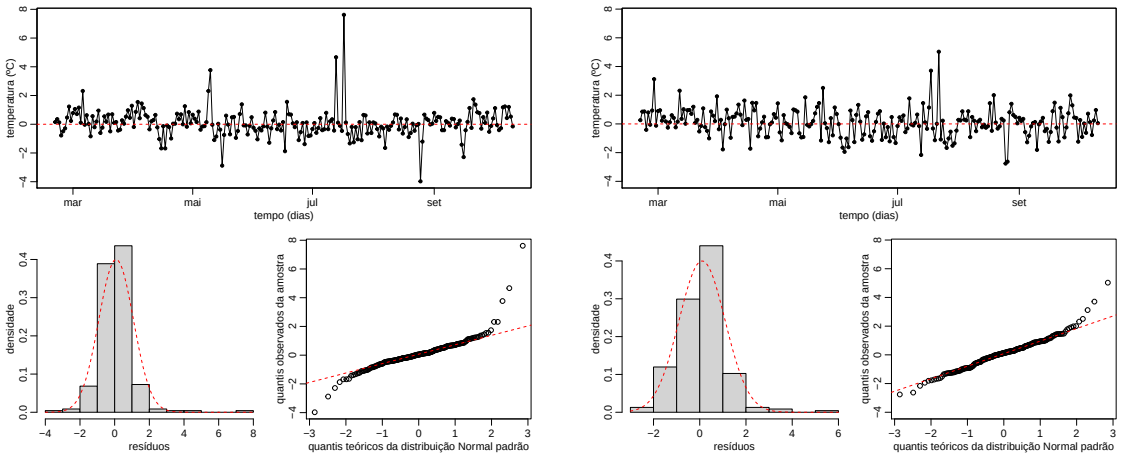


Figura 7.13: Série, histograma e *QQ-plot* dos resíduos da temperatura máxima; lado esquerdo - MEES_1, lado direito - MEES_6.

Relativamente ao pressuposto da independência dos erros, fez-se o teste Ljung-Box, aplicado à série dos resíduos (inovações), fazendo variar k entre 7 e 17, onde k representa o número de autocorrelações a serem testadas. Segundo os resultados do teste, a hipótese da independência é rejeitada para todos os valores de k, exceto para $k = 17$ (valor de prova de 0,0534) para o modelo a 1-passo e exceto para $k = 17$ (valor de prova de 0,0529) para o modelo a 6-passos. De facto, a FAC e a FACP dos resíduos (Figura 7.14) apresentam correlações significativas para ambos os modelos de calibração. Portanto, o pressuposto da independência não é verificado.

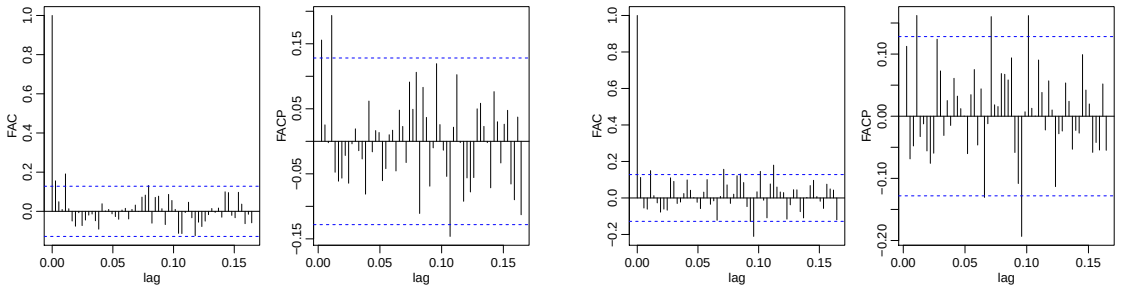


Figura 7.14: FAC e FACP dos resíduos da temperatura máxima; lado esquerdo - MEES_1, lado direito - MEES_6.

Segundo a Figura 7.15, onde estão representadas as séries da temperatura máxima observada (a preto), das previsões a 1-passo do *website* com a substituição

dos *outliers* do rácio $Y_t^M/W_{t,(h)}^M$ por estimativas através da interpolação linear (a vermelho) e das respetivas previsões a 1-passo calibrada (azul), correspondentes aos modelos MEES_1^* (gráfico superior) e MEES_6^* (gráfico inferior), é notória uma melhoria nas previsões calibradas em relação às previsões dadas pelo *website* com a substituição de observações, onde é possível perceber que as previsões calibradas a 1-passo acompanham melhor a série da temperatura máxima observada do que a previsões a 6-passos calibradas.

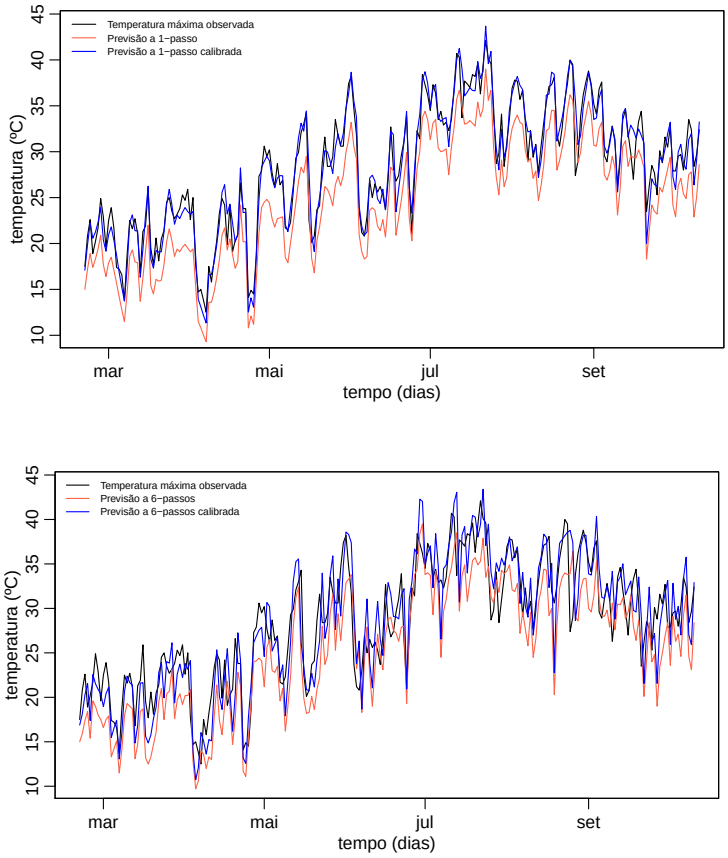


Figura 7.15: Séries da temperatura máxima observada (a preto), previsões do *website* com a substituição dos *outliers* do rácio $Y_t^M/W_{t,(h)}^M$ (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEES_1^*, gráfico inferior - MEES_6^*.

Analisando os gráficos da Figura 7.16, a principal característica que se destaca são as amplitudes dos intervalos entre o modelo de calibração a 1-passo (esquerda) e o modelo de calibração a 6-passos (direita), onde se verifica que as margens de erro para o modelo a 6-passos são bastante superiores comparativamente ao modelo a 1-passo. Além disso, os comportamentos de $\beta_{t|t-1}$ (previsão), $\beta_{t|t}$ (filtragem) e $\beta_{t|n}$ (alisamento) são mais suaves no modelo a 1-passo. Isto deve-se ao facto das previsões com maior horizonte temporal terem, geralmente, maior erro associado.

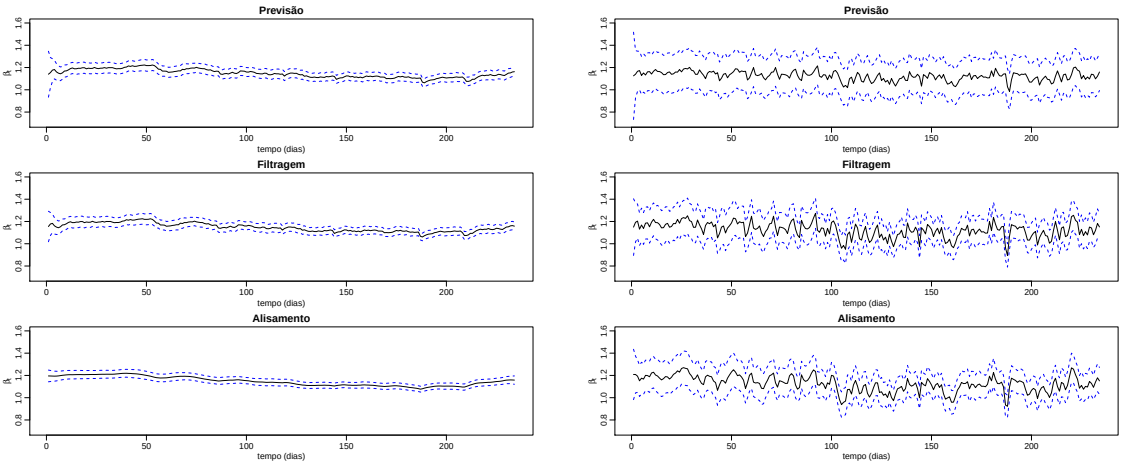


Figura 7.16: Parte superior - previsões $\beta_{t|t-1}$; centro - filtragem $\beta_{t|t}$; parte inferior- alisamento $\beta_{t|n}$ para a temperatura máxima; lado esquerdo - MEES_1^*, lado direito - MEES_6^*.

Procedendo com a análise das inovações (resíduos), de acordo com a Figura 7.17, verifica-se que o histograma associado aos resíduos do modelo a 1-passo (esquerda) apresenta uma cauda pesada à esquerda, não sendo o suficiente para rejeitar a hipótese da normalidade dos erros do teste Kolmogorov-Smirnov (valor de prova de 0,1105); o histograma dos resíduos associado ao modelo a 6-passos (direita) apresenta uma ligeira assimetria à direita, não sendo também o suficiente para rejeitar a hipótese da distribuição Normal dos erros (valor de prova de 0,8338). Analisando o *QQ-plot* para ambos os modelos, verifica-se que os pontos encontram-se alinhados, havendo uns desvios no extremo para o modelo a 1-passo.

De acordo com a representação gráfica da série dos resíduos, esta parece apresentar uma distribuição em torno de zero, sendo comprovada a não rejeição da hipótese de média nula pelo teste t para o valor esperado (valor de prova de 0,5824 para o modelo a 1-passo e 0,8438 para o modelo a 6-passos). Relativamente à variância dos erros, a análise gráfica sugere que a variabilidade dos resíduos para o modelo a 6-passos é superior comparativamente ao modelo a 1-passo.

Recorreu-se ao teste Ljung-Box para testar o pressuposto da independência dos erros, fazendo variar k, que corresponde ao número de autocorrelações a serem testadas, entre 7 e 17. Segundo os resultados do teste, a hipótese da independência para o modelo a 1-passo não é rejeitada para nenhum valor de k, apresentando valores de prova entre 0,0941 ($k = 8$) e 0,4579 ($k = 17$). No entanto, para o modelo a 6-passos, a hipótese da independência dos erros é rejeitada para todos os valores de k.

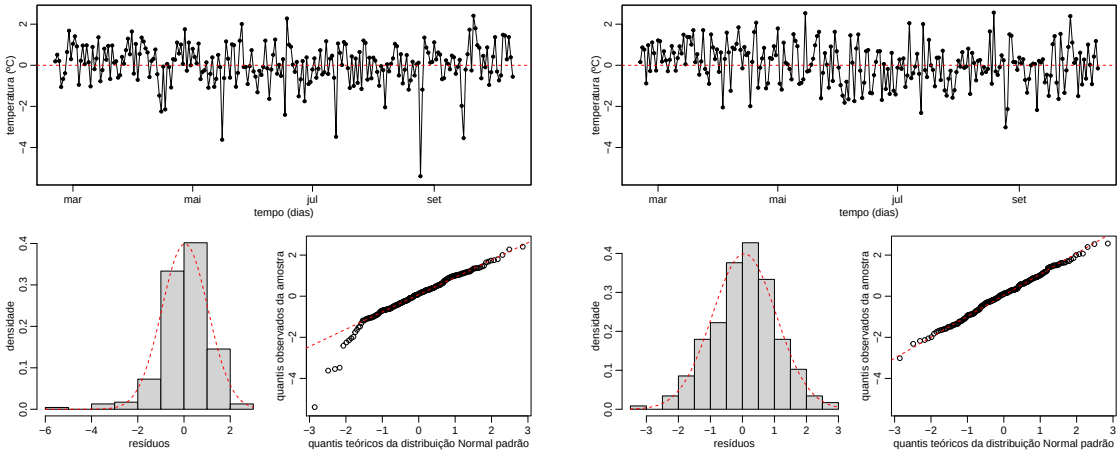


Figura 7.17: Série, histograma e *QQ-plot* dos resíduos; lado esquerdo - MEES_1^*, lado direito - MEES_6^*.

Os resultados do teste Ljung-Box estão de acordo com a análise da FAC e a FACP dos resíduos (Figura 7.14), uma vez que, para o modelo a 1-passo (esquerda), não apresentam correlações significativas, contrariamente ao que se sucede no modelo a 6-passos (direita), apresentando correlações significativas para vários lags.

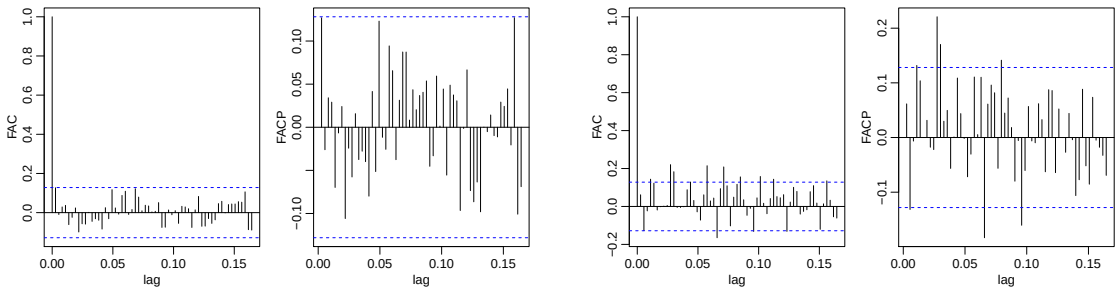


Figura 7.18: FAC e FACP dos resíduos da temperatura máxima; lado esquerdo - MEES_1^*, lado direito - MEES_6^*.

Tendo em conta a Tabela 7.5, verifica-se que

– Nos quatro modelos, as estimativas e os erros padrão de σ_e tendem a aumentar com o aumento do horizonte temporal, a passo que a log-verosimilhança tende a diminuir;

– O parâmetro α é determinístico e representa o erro sistemático, que tende a aumentar com o aumento do horizonte temporal. As estimativas de β, no caso dos modelos de regressão linear, e as estimativas de μ, no caso dos modelos de calibração, tendem a diminuir com o aumento do horizonte temporal;

– Comparando os modelos RLSC_h e MEEC_h, é possível ver que as estimativas do valor médio do processo $\{\beta_t\}_{t=1,\ldots,n}$, μ, variam entre os 0,64-0,96°C, coincidindo com as estimativas de β no modelo de regressão linear, que variam entre os 0,81-0,97°C. Neste caso, representa a variação média na temperatura máxima por cada incremento unitário nas previsões. Por outro lado, as estimativas e os erros padrão de σ_e são mais elevados no modelo de regressão linear. As mesmas conclusões podem ser tiradas comparando os modelos RLSC_h^* e MEEC_h^*. Para o modelo MEEC_h^*, verifica-se que as estimativas do valor médio do processo $\{\beta_t\}_{t=1,\ldots,n}$, μ, assumem valores muito próximos de 1°C, principalmente nos primeiros horizontes temporais, também coincidindo com as estimativas de β no modelo RLSC_h^*;

– Após a transformação de $W_{t,(h)}$, $W_{t,(h)}^*$, o modelo MEEC_h^* obteve estimativas menores para ϕ, do que o modelo MEEC_h.

Na Tabela 7.6 estão apresentadas algumas medidas de avaliação e critérios de seleção referentes aos quatro modelos cujas estimativas dos parâmetros encontram-se na Tabela 7.5. No geral, percebe-se que

– Nos quatro modelos, os critérios AIC e BIC tendem a aumentar com o aumento do horizonte temporal, assim como as medidas REQM, EAM, EEAM e U-Theil. Porém, r^2 e r_a^2 tendem a diminuir;

– O modelo MEEC_h obteve menores valores de AIC e BIC do que o modelo RLSC_h, assim como obteve maioritariamente os menores valores de REQM, EAM, EEAM e U-Theil. Também obteve melhor ajustamento, uma vez que apresenta maior r_a^2.

– O modelo MEEC_h^* produziu menores valores de AIC e BIC, assim como obteve os menores valores de REQM, EAM, EEAM e U-Theil do que o modelo RLSC_h^*. Além disso, obteve maior r_a^2.

Tabela 7.5: Estimativas dos parâmetros e respetivos erros padrão dos quatro modelos com a componente aditiva determinística α para a série da temperatura máxima, onde Y_t^M representa a temperatura máxima observada, $W_{t,(h)}^M$ corresponde às previsões a h-passos do *website* e $W_{t,(h)}^{M*}$ corresponde às previsões com a substituição dos *outliers* do rácio $Y_t^M/W_{t,(h)}^M$, $h = 1, \ldots, 6$ dias.

		h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \alpha + \beta W_{t,(h)}^M + e_t$							
α	estimativa	4,4775	5,0579	5,8903	5,7791	6,3297	8,2067
	erro padrão	0,4596	0,5000	0,6126	0,6566	0,7681	0,8222
β	estimativa	0,9693	0,9434	0,9120	0,9097	0,8816	0,8063
	erro padrão	0,0180	0,0196	0,0240	0,0256	0,0297	0,0317
σ_e	estimativa	1,8227	2,0125	2,4873	2,6301	3,0535	3,4335
	erro padrão	0,1189	0,1313	0,1623	0,1716	0,1992	0,2240
	logL	-471,5038	-494,6869	-544,2513	-557,3135	-592,2371	-619,6877
$Y_t^M = \alpha + \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,7943	0,7924	0,8522	0,8690	0,7408	0,8258
	erro padrão	0,1051	0,1315	0,1289	0,0949	0,1674	0,1167
α	estimativa	4,7240	5,2587	6,3762	7,2620	7,3375	12,1181
	erro padrão	0,5356	0,5531	0,8755	1,1049	1,1565	2,2116
μ	estimativa	0,9602	0,9363	0,8936	0,8495	0,8398	0,6392
	erro padrão	0,0223	0,0224	0,0355	0,0460	0,0480	0,0957
σ_ε	estimativa	0,0153	0,0122	0,0109	0,0211	0,0327	0,0547
	erro padrão	0,0064	0,0073	0,0078	0,0086	0,0150	0,0200
σ_e	estimativa	1,6912	1,9313	2,4071	2,3959	2,7800	2,7383
	erro padrão	0,0955	0,1041	0,1282	0,1370	0,1901	0,3057
	logL	-253,6134	-278,5940	-328,1716	-338,0121	-374,8660	-398,3939
$Y_t^M = \alpha + \beta W_{t,(h)}^{M*} + e_t$							
α	estimativa	3,6466	3,9587	4,4192	4,4142	5,1008	6,6844
	erro padrão	0,3735	0,4130	0,4721	0,5433	0,6621	0,7403
β	estimativa	0,9946	0,9775	0,9633	0,9569	0,9242	0,8576
	erro padrão	0,0146	0,0160	0,0184	0,0210	0,0255	0,0283
σ_e	estimativa	1,4540	1,6182	1,8647	2,1206	2,5870	3,0005
	erro padrão	0,0948	0,1056	0,1216	0,1383	0,1688	0,1957
	logL	-418,6214	-443,6505	-476,8316	-506,9204	-553,4452	-588,1389
$Y_t^M = \alpha + \beta_t W_{t,(h)}^{M*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,2778	0,4762	0,2988	0,2959	0.4197	0,4063
	erro padrão	0,1067	0,0986	0,1370	0,0967	0,1297	0,1058
α	estimativa	3,4330	3,7801	4,3864	4,4442	5,4486	7,3357
	erro padrão	0,3651	0,4586	0,4844	0,5442	0,7589	0,9085
μ	estimativa	1,0038	0,9856	0,9647	0,9555	0,9094	0,8289
	erro padrão	0,0161	0,0203	0,0206	0,0237	0,0320	0,0396
σ_ε	estimativa	0,0476	0,0474	0,0501	0,0669	0,0635	0,0914
	erro padrão	0,0072	0,0072	0,0108	0,0099	0,0140	0,0139
σ_e	estimativa	0,7329	0,8889	1,2941	1,1588	1,8671	1,6087
	erro padrão	0,2289	0,1831	0,2324	0,2918	0,2814	0,3955
	logL	-193,9268	-213,6194	-256,9069	-283,9959	-333,1306	-361,6854

Tabela 7.6: Comparação das medidas e dos critérios de seleção dos quatro modelos para a série da temperatura máxima.

	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \alpha + \beta W_{t,(h)}^M + e_t$						
REQM	1,8149	2,0039	2,4767	2,6189	3,0404	3,4188
EAM	1,1636	1,2816	1,6645	**1,8134**	2,1978	2,6056
EEAM	0,4596	0,5062	0,6575	**0,7163**	0,8681	1,0292
U-Theil	0,5437	0,5919	**0,6954**	**0,7083**	0,8360	**0,9580**
r^2	0,9257	0,9094	0,8616	0,8452	0,7914	0,7363
r_a^2	0,9254	0,9090	0,8610	0,8445	0,7905	0,7352
AIC	949,0076	995,3738	1094,5026	1120,6270	1190,4742	1245,3754
BIC	959,3736	1005,7398	1104,8686	1130,9930	1200,8402	1255,7414
$Y_t^M = \alpha + \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,7752**	**1,9828**	**2,4552**	**2,5553**	**2,9850**	**3,2757**
EAM	**1,1336**	**1,2534**	**1,6576**	1,8343	**2,1471**	**2,5769**
EEAM	**0,4478**	**0,4951**	**0,6547**	0,7245	**0,8481**	**1,0178**
U-Theil	**0,5262**	**0,5840**	0,6976	0,7210	**0,8192**	1,0001
r^2	**0,9289**	**0,9113**	**0,8640**	**0,8529**	**0,7990**	**0,7585**
r_a^2	**0,9286**	**0,9109**	**0,8634**	**0,8523**	**0,7981**	**0,7575**
AIC	**517,2268**	**567,1880**	**666,3432**	**686,0242**	**759,7320**	**806,7878**
BIC	**534,5034**	**584,4646**	**683,6198**	**703,3008**	**777,0086**	**824,0644**
$Y_t^M = \alpha + \beta W_{t,(h)}^{M*} + e_t$						
REQM	1,4478	1,6112	1,8567	2,1115	2,5759	2,9876
EAM	1,0456	1,1571	1,4279	1,6055	2,0420	2,3693
EEAM	0,4130	0,4571	0,5640	0,6341	0,8065	0,9358
U-Theil	0,5055	0,5789	0,6181	0,7455	0,8966	1,0023
r^2	0,9527	0,9414	0,9222	0,8994	0,8503	0,7986
r_a^2	0,9525	0,9411	0,9219	0,8990	0,8497	0,7977
AIC	843,2428	893,3010	959,6632	1019,8408	1112,8904	1182,2778
BIC	853,6088	903,6670	970,0292	1030,2068	1123,2564	1192,6438
$Y_t^M = \alpha + \beta_t W_{t,(h)}^{M*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,4171**	**1,5130**	**1,8334**	**2,0624**	**2,5186**	**2,8589**
EAM	**1,0157**	**1,1164**	**1,3954**	**1,5624**	**1,9544**	**2,2816**
EEAM	**0,4012**	**0,4410**	**0,5511**	**0,6171**	**0,7719**	**0,9012**
U-Theil	**0,4736**	**0,5056**	**0,5858**	**0,6875**	**0,8205**	**0,8583**
r^2	**0,9548**	**0,9484**	**0,9242**	**0,9041**	**0,8569**	**0,8156**
r_a^2	**0,9546**	**0,9482**	**0,9239**	**0,9037**	**0,8563**	**0,8148**
AIC	**397,8536**	**437,2388**	**523,8138**	**577,9918**	**676,2612**	**733,3708**
BIC	**415,1302**	**454,5154**	**541,0904**	**595,2684**	**693,5378**	**750,6474**

Através da análise gráfica da Figura 7.19, é percetível uma melhoria das previsões calibradas em relação às previsões dadas pelo *website*. Além disso, as previsões

calibradas a 1-passo sobrepõem melhor a série da temperatura máxima observada do que a série das previsões a 6-passos calibrada, o que é bastante intuitivo dado que as previsões mais próximas tendem a ser mais precisas.

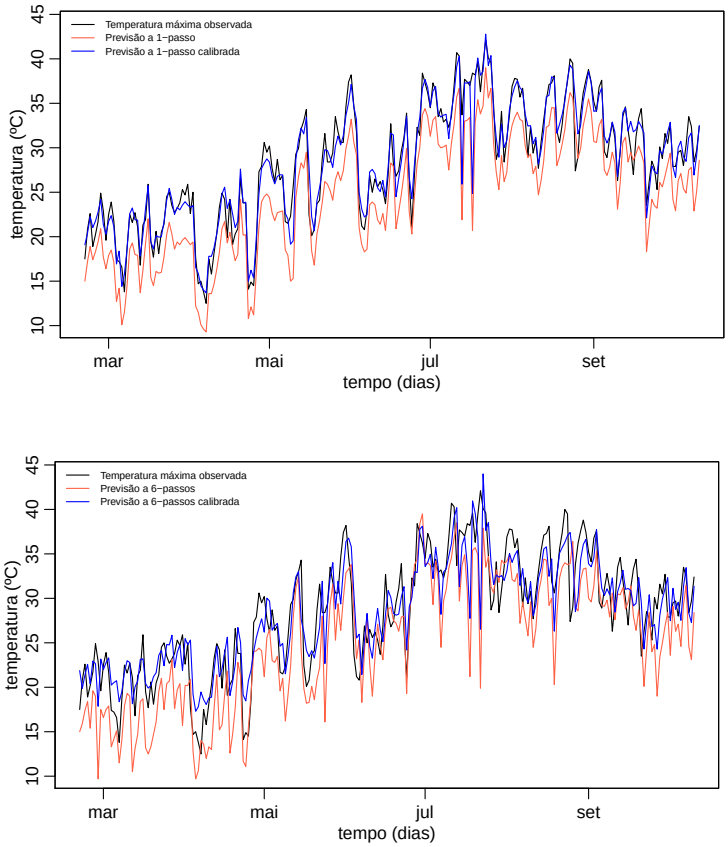


Figura 7.19: Séries da temperatura máxima observada (a preto), previsões do *website* (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEEC_1, gráfico inferior - MEEC_6.

Na parte superior da Figura 7.20 estão representadas as previsões do estado, $\beta_{t|t-1}$, no gráfico do meio estão os valores filtrados, $\beta_{t|t}$, e no gráfico inferior estão representados os valores alisados, $\beta_{t|n}$ para $t = 1, \ldots, 234$ dias.

Através da análise, verifica-se que os intervalos (limites do erro) para o modelo de calibração com $h = 6$ dias (lado direito) apresentam uma margem de erro bastante superior comparativamente aos intervalos para o modelo de calibração com $h = 1$ dia (lado esquerdo). Além disso, o comportamento dos três gráficos do lado esquerdo é mais irregular. É de notar que à medida que o horizonte temporal aumenta, também aumenta a incerteza associada às previsões e, consequentemente, os limites do erro associado também aumentam.

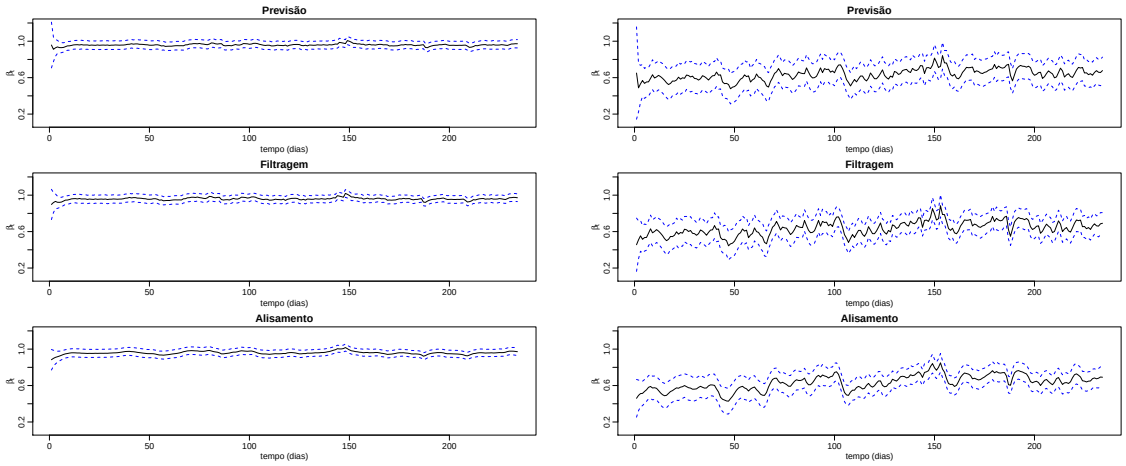


Figura 7.20: Parte superior - previsões $\beta_{t|t-1}$; centro - filtragem $\beta_{t|t}$; parte inferior- alisamento $\beta_{t|n}$ para a temperatura máxima; lado esquerdo - MEEC_1, lado direito - MEEC_6.

Para verificar se um modelo é apropriado, é necessário que os resíduos cumpram os pressupostos da normalidade, independência, média nula e variância constante ao longo do tempo.

Analisando o histograma dos resíduos (Figura 7.13) correspondente ao modelo MEEC_1 (lado esquerdo), sugere que os resíduos apresentam uma distribuição aproximadamente simétrica. No entanto, no *QQ-plot* existem pontos que estão bastante afastados de uma reta, sendo o suficiente para que o teste Kolmogorov-Smirnov rejeite a hipótese da normalidade dos erros (valor de prova de 0,0076). No caso do modelo MEEC_6 (lado direito), o histograma dos resíduos sugere que os resíduos têm um comportamento próximo ao da função densidade da distribuição Normal, além de que no *QQ-plot*, os pontos estão dispostos maioritariamente sobre uma linha reta, tendo apenas 2 pontos no extremo que se desviam. O teste Kolmogorov-Smirnov não rejeita a hipótese da normalidade dos erros (valor de prova de 0,8206).

Analisando ambos os gráficos da série dos resíduos, parecem apresentar uma distribuição em torno de zero, não apontando para a rejeição da média nula dos erros. Relativamente ao pressuposto da homocedasticidade dos erros, a análise gráfica das séries dos resíduos sugerem que, para ambos os modelos, a variância não é constante.

Em relação ao pressuposto da independência dos erros, aplicou-se o teste Ljung-Box à série dos resíduos (inovações), onde foram testadas de 7 a 17 autocorrelações. Segundo o teste, a hipótese da independência dos erros para o modelo a 1-passo é rejeitada para os valores de $k = 7, 8, 9, 10, 14$, apresentando valores de prova entre 0,0035 ($k = 7$) e 0,1287 ($k = 17$). Para o modelo a 6-passos, a hipótese da independência dos erros é rejeitada para todos os valores de k.

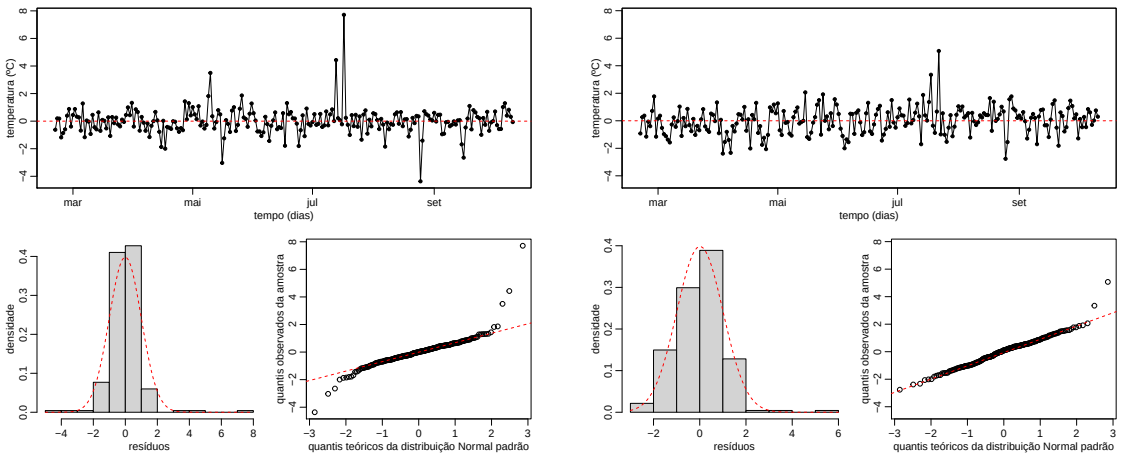


Figura 7.21: Série, histograma e *QQ-plot* dos resíduos da temperatura máxima; lado esquerdo - MEEC_1, lado direito - MEEC_6.

Relativamente à FAC e a FACP dos resíduos, representadas na Figura 7.22, apresentam correlações significativas para alguns lags em ambos os modelos de calibração. Portanto, o pressuposto da independência dos erros não é verificado.

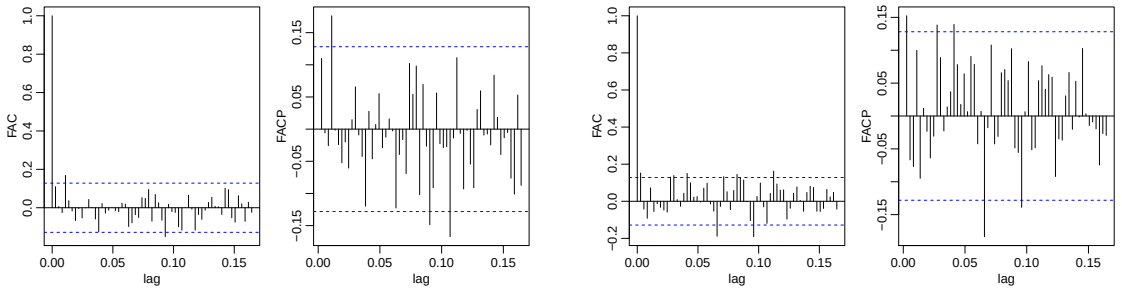


Figura 7.22: FAC e FACP dos resíduos da temperatura máxima; lado esquerdo - MEEC_1, lado direito- MEEC_6.

Analisando os gráficos da Figura 7.23, verifica-se que, no geral, as previsões calibradas parecem aproximar-se mais da temperatura máxima observada do que as previsões dadas pelo *website* com a substituição de observações dos *outliers* do rácio $Y_t^M / W_{t,(h)}^M$, onde é possível perceber que as previsões a 1-passo calibradas têm uma melhor *performance*, visto que sobrepõem melhor a série da temperatura máxima observada (gráfico superior).

Através da análise gráfica da Figura 7.24, percebe-se que o comportamento de $\beta_{t|n}$ (alisamento) para ambos os modelos é mais irregular comparativamente ao comportamento de $\beta_{t|t-1}$ (previsão). Além disso, os comportamentos dos três gráficos do modelo a 6-passos (direita) aparentam ser mais irregulares do que no modelo a

1-passo (esquerda), além de possuírem maior variabilidade.

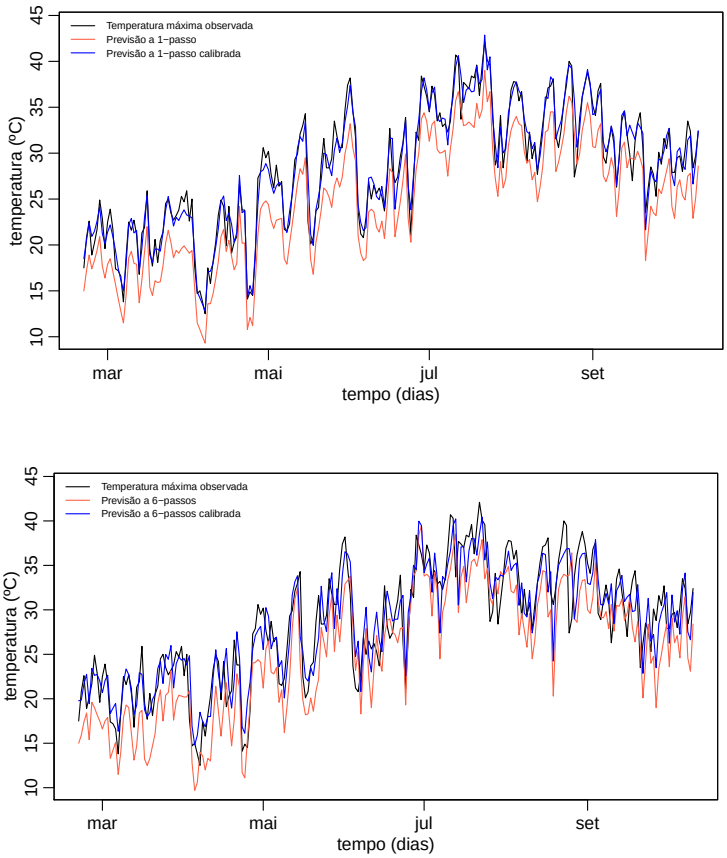


Figura 7.23: Séries da temperatura máxima observada (a preto), previsões do *website* com a substituição dos *outliers* do rácio $Y_t^M/WM_{t,(h)}$ (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEEC_1^*, gráfico inferior - MEEC_6^*.

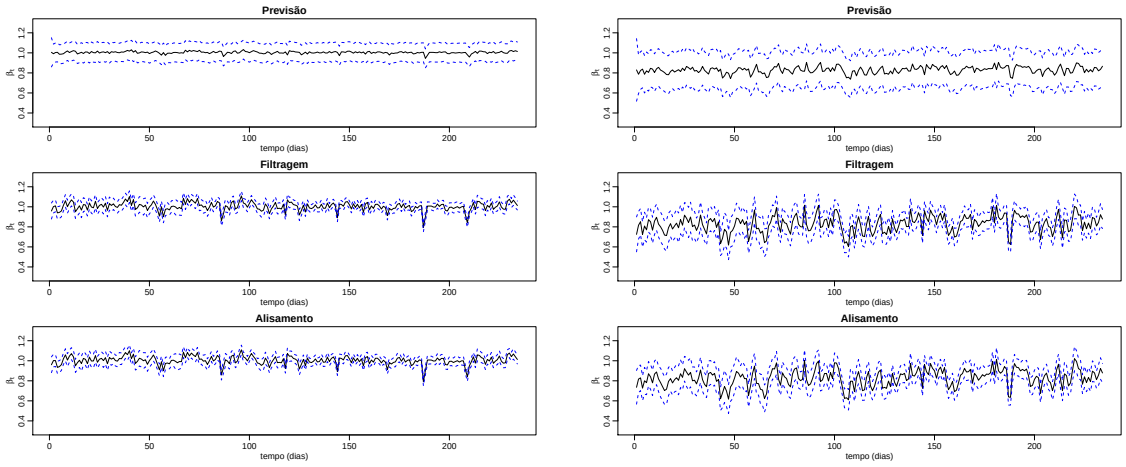


Figura 7.24: Parte superior - previsões $\beta_{t|t-1}$; centro - filtragem $\beta_{t|t}$; parte inferior- alisamento $\beta_{t|n}$ para a temperatura máxima; lado esquerdo - MEEC_1^*, lado direito - MEEC_6^*.

Relativamente à análise dos resíduos, na Figura 7.25, verifica-se que o histograma associado aos resíduos do modelo a 1-passo (esquerda) apresenta uma cauda bastante evidente à esquerda, que, no entanto, não foi suficiente para rejeitar a hipótese da normalidade dos erros do teste Kolmogorov-Smirnov (valor de prova de 0,1699); o histograma dos resíduos associado ao modelo a 6-passos (direita) apresenta uma assimetria à direita, não sendo também o suficiente para rejeitar a hipótese da distribuição Normal dos erros (valor de prova de 0,4369). Analisando o *QQ-plot* para ambos os modelos, verifica-se que os pontos encontram-se alinhados, havendo alguns pontos no extremo no modelo a 1-passo que se distanciam da reta.

De acordo com a representação gráfica da série dos resíduos (Figura 7.25), esta parece apresentar uma distribuição em torno de zero, sendo comprovada a não rejeição da hipótese de média nula pelo teste t para o valor esperado (valor de prova de 0,8764 para o modelo a 1-passo e 0,9981 para o modelo a 6-passos). Relativamente à variância dos erros, a análise gráfica sugere que a variabilidade dos resíduos para o modelo a 6-passos é superior comparativamente ao modelo a 1-passo. Além disso, a variabilidade resíduos de ambos os modelos não aparentam apresentar um comportamento constante ao longo do tempo.

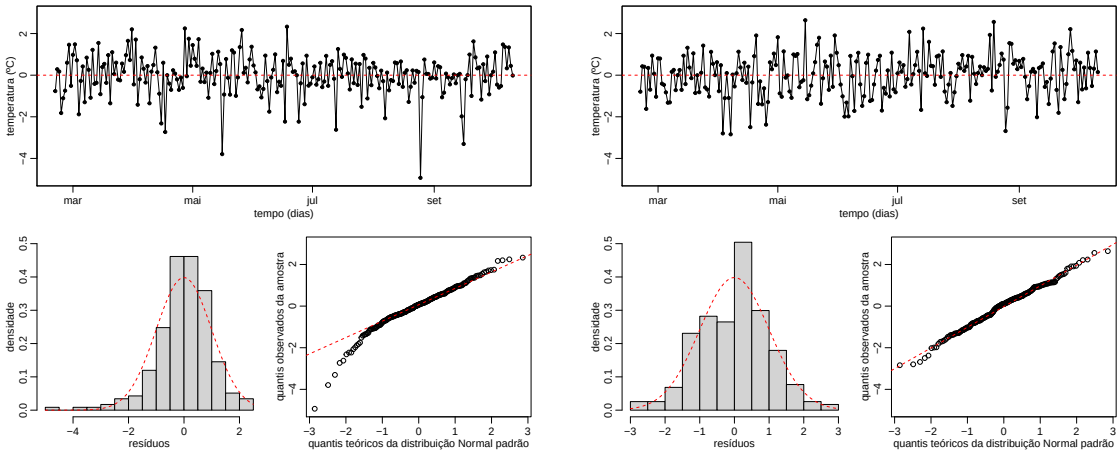


Figura 7.25: Série, histograma e *QQ-plot* dos resíduos da temperatura máxima; lado esquerdo - MEEC_1^*, lado direito - MEEC_6^*.

Para verificar o pressuposto da independência dos erros, o teste Ljung-Box é aplicado à série dos resíduos, fazendo variar k (número de autocorrelações) entre 7 a 17, não rejeitando a hipótese da independência dos erros para nenhum valor de k no modelo a 1-passo, apresentando valores de prova entre 0,2632 ($k = 7$) e 0,8219 ($k = 14$). No modelo a 6-passos, a hipótese da independência dos erros é rejeitada para todos os valores de k, exceto para $k = 8$ (valor de prova de 0,0581),

$k = 9$ (valor de prova de 0,1447) e $k = 17$ (valor de prova de 0,0833). De facto, os resultados do teste Ljung-Box estão de acordo com a análise da FAC e a FACP dos resíduos (Figura 7.22), uma vez que, para o modelo a 1-passo (esquerda), não apresentam correlações significativas, contrariamente ao que se sucede no modelo a 6-passos (direita), na qual se destaca uma correlação bastante significativa para ambos os gráficos.

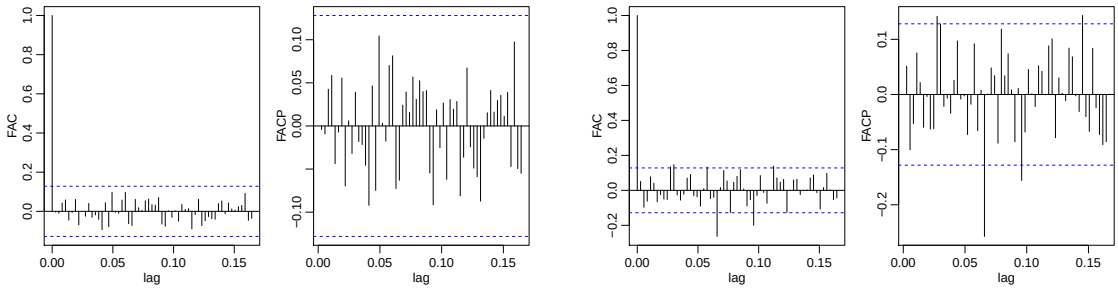


Figura 7.26: FAC e FACP dos resíduos da temperatura máxima; lado esquerdo - MEEC$_1^*$, lado direito - MEEC$_6^*$.

Considerações adicionais

Num ponto de vista mais amplo, também é possível obter mais conclusões através da comparação das Tabelas 7.3 e 7.5

- O modelo MEEC$_h$ obteve menores estimativas de ϕ do que o modelo MEES$_h$, assim como o modelo MEEC$_h^*$ produziu menores estimativas para o mesmo parâmetro do que o modelo MEES$_h^*$;
- O modelo de regressão linear RLSC$_h$ obteve menores estimativas para σ_e do que o modelo RLSS$_h$. Analogamente, o modelo RLSC$_h^*$ obteve menores estimativas para σ_e do que o modelo RLSS$_h^*$.

Comparando as Tabelas 7.4 e 7.6, conclui-se que

- De todos os modelos cuja variável independente é $W_{t,(h)}$, o que obteve menores valores de AIC e BIC, assim como produziu maioritariamente menores valores para as medidas REQM, EAM, EEAM e U-Theil e maior r_a^2 foi o modelo MEEC$_h$;
- De todos os modelos cuja variável independente é $W_{t,(h)}^*$, o que obteve menores valores de AIC e BIC, assim como produziu maioritariamente menores valores para as medidas REQM, EAM, EEAM e U-Theil e maior r_a^2 foi o modelo MEEC$_h^*$;

– Por um lado, verificou-se que o modelo RLSC_h gerou menores valores de REQM, EAM, EEAM e obteve maiores valores de r_a^2 do que o modelo MEES_h. No entanto, o modelo MEES_h obteve menores valores de U-Theil e de AIC e BIC. Esta situação é parecida nos modelos RLSC_h^* e MEES_h^* (exceto no caso do r_a^2);

– Todos os modelos obtiveram valores elevados de r_a^2, dado que todos são superiores a 0,70.

Relativamente ao cumprimento dos pressupostos associados aos modelos (normalidade, média nula, variância constante e a não correlação dos erros), verificou-se que

– Os modelos MEES_h e MEEC_h falham no pressuposto da independência dos erros para $h = 1, 6$ dias; além disso, não cumprem o pressuposto da normalidade para $h = 1$.

  Atendendo aos restantes horizontes temporais ($h = 2, 3, 4, 5$ dias), foram efetuados os testes de normalidade Kolmogorov-Smirnov e verificou-se que a condição da normalidade dos erros não é válida em ambos os modelos com $h = 2, 3, 4$ dias. Já a condição da não correlação, recorreu-se ao teste Ljung-Box, onde se fez variar o número de autocorrelações de 7 a 17, inclusive, e chegou-se à conclusão de que nenhum dos modelos compre o referido pressuposto;

– Ambos os modelos MEES_h^* e MEEC_h^* cumprem os pressupostos da normalidade para $h = 1, 6$ dias; o pressuposto da independência dos erros foi verificada para ambos os modelos com $h = 1$; no entanto, não foi verificada para $h = 6$.

  Em relação aos restantes horizontes temporais ($h = 2, 3, 4, 5$ dias), foram efetuados os testes de normalidade Kolmogorov-Smirnov e verificou-se que a condição da normalidade dos erros não é válida para o modelo MEES_h^* com $h = 4$ dias (valor de prova de 0,0112) e não é válida para o modelo MEEC_h^* com $h = 2$ dias (valor de prova de 0,0457) e $h = 4$ dias (valor de prova de 0,0306). Já a condição da não correlação, recorreu-se ao teste Ljung-Box, onde se fez variar o número de autocorrelações de 7 a 17, inclusive. O modelo MEES_h^* não rejeitou a hipótese da independência dos erros apenas para $h = 3$ dias (cujos valores de prova variam entre 0,0753 ($k = 7$) e 0,4776 ($k = 15$)). O modelo MEEC_h^* não rejeitou a hipótese da independência dos erros para $h = 2$ dias (valores de prova variam entre 0,0937 ($k = 8$) e 0,4993 ($k = 17$)), $h = 3$ (valores de prova variam entre 0,0978 ($k = 7$) e 0,6483 ($k = 15$)) e $h = 5$ (valores de prova variam entre 0,1150 ($k = 8$) e 0,3713 ($k = 14$)).

7.3.2 Temperatura Mínima

Na Figura 7.27 estão representadas as séries temporais da temperatura mínima observada (a preto) com as respetivas previsões a h-passos, $h = 1, \ldots, 6$ dias, provenientes do *website weatherstack* (a vermelho). Diferentemente ao que acontece na temperatura máxima, na qual as previsões subestimavam o valor observado, através da análise gráfica verifica-se que as previsões da temperatura mínima parecem acompanhar a série observada, existindo previsões que se afastam do valor da temperatura mínima observada.

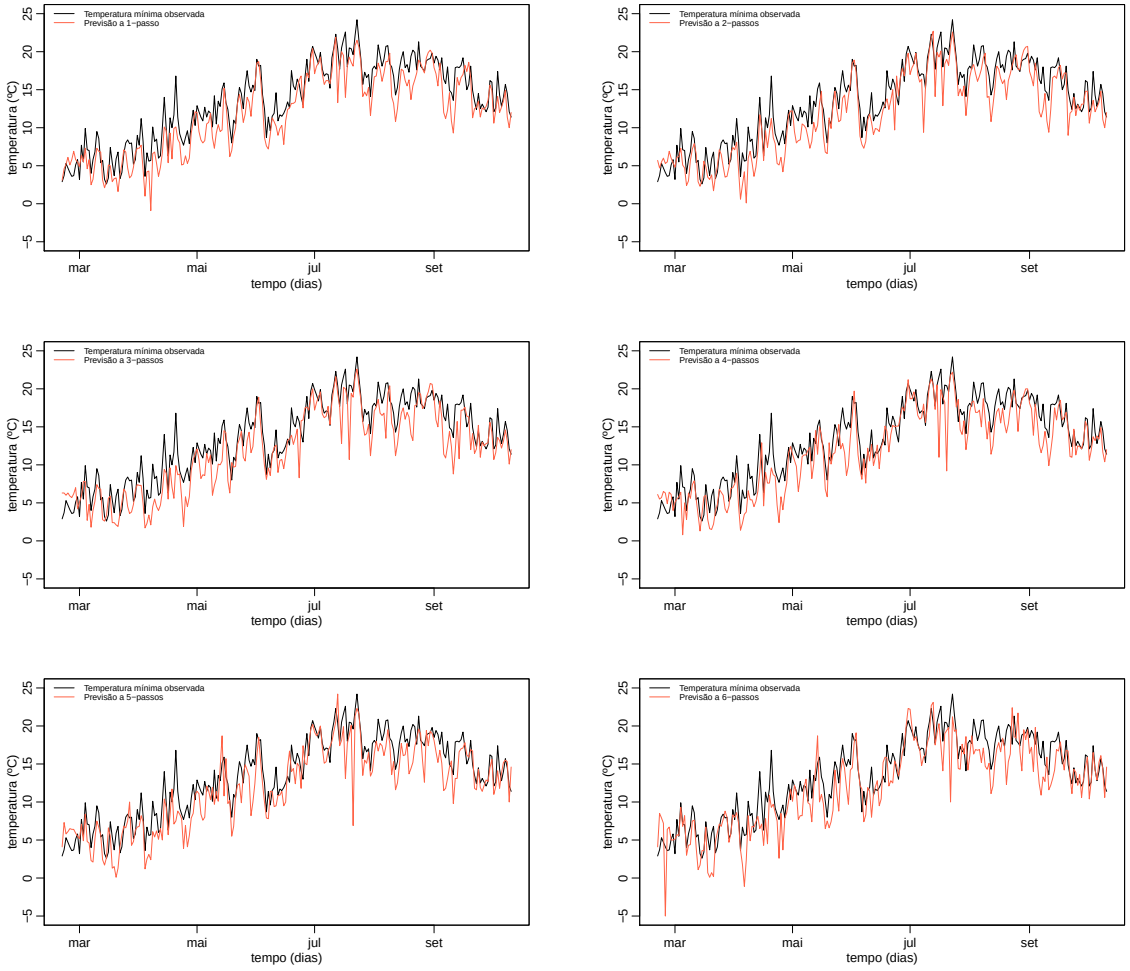


Figura 7.27: Séries da temperatura mínima observada (a preto) e das respetivas previsões a h-passos (a vermelho), $h = 1, \ldots, 6$ dias.

Analisando a Tabela 7.7, é possível tirar algumas conclusões

- Nos quatro modelos, as estimativas e os erros padrão de σ_e tendem a aumentar com o aumento do horizonte temporal. Por outro lado, a log-verosimilhança tende a diminuir com o aumento do horizonte temporal;
- No modelo MEES_h, o valor estimado de ϕ está muito próximo de 0,90 (um processo $AR(1)$ é estacionário se $|\phi| < 1$);
- Comparando os modelos RLSS_h e MEES_h, verifica-se que as estimativas do valor médio do processo $\{\beta_t\}_{t=1,\ldots,n}$, μ, variam entre os 1,09-1,14°C, que é ligeiramente superior à estimativa de β no modelo de regressão linear, que varia entre os 1,07-1,10°C. Isto indica que, em média, a temperatura mínima observada é cerca de 9-14% superior em relação às previsões, no modelo de calibração, e cerca de 7-10% superior em relação às previsões, no modelo de regressão linear simples. Por outro lado, as estimativas e os erros padrão de σ_e são mais elevados no modelo de regressão linear. As mesmas conclusões podem ser tiradas comparando os modelos RLSS_h^* e MEES_h^*.

Na Tabela 7.8 estão exibidas algumas medidas de avaliação e critérios de seleção referentes aos quatro modelos cujas estimativas dos parâmetros encontram-se na Tabela 7.7. No geral, constata-se que

- Nos quatro modelos, os critérios AIC e BIC tendem a aumentar com o aumento do horizonte temporal, assim como as medidas REQM, EAM, EEAM e U-Theil. Por outro lado, r^2 e r_a^2, tendem a diminuir com o aumento do horizonte temporal;
- O modelo MEES_h obteve menores valores de AIC e BIC em comparação com o modelo RLSS_h, assim como obteve menores valores de REQM, EAM, EEAM e U-Theil e menor r_a^2. O mesmo acontece com os modelos RLSS_h^* e MEES_h^*;
- Nos modelos MEES_h e RLSS_h, verifica-se que, a partir de $h = 5$, a estatística U-Theil assume valores muito elevados.

Tabela 7.7: Estimativas dos parâmetros e respetivos erros padrão dos quatro modelos para a série da temperatura mínima, onde Y_t^m representa a temperatura mínima observada, $W_{t,(h)}^m$ corresponde às previsões a h-passos do *website* e $W_{t,(h)}^{m*}$ corresponde às previsões com a substituição dos *outliers* do rácio $Y_t^m/W_{t,(h)}^m$, $h = 1, \ldots, 6$ dias.

		h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^m = \beta W_{t,(h)}^m + e_t$							
β	estimativa	1,1028	1,0971	1,1003	1,0912	1,0880	1,0733
	erro padrão	0,0106	0,0107	0,0123	0,0119	0,0127	0,0135
σ_e	estimativa	2,0906	2,1134	2,4185	2,3681	2,5195	2,7176
	erro padrão	0,1341	0,1354	0,1551	0,1523	0,1622	0,1743
	logL	-504,0925	-506,6304	-538,1841	-533,2560	-547,7639	-565,4683
$Y_t^m = \beta_t W_{t,(h)}^m + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,8943	0,8797	0,9207	0,8994	0,9059	0,9052
	erro padrão	0,0541	0,0574	0,0397	0,0562	0,0536	0,0440
μ	estimativa	1,1348	1,1274	1,1397	1,1079	1,0975	1,0865
	erro padrão	0,0268	0,0245	0,0331	0,0222	0,0196	0,0228
σ_ε	estimativa	0,0356	0,0354	0,0322	0,0227	0,0181	0,0238
	erro padrão	0,0108	0,0112	0,0100	0,0102	0,0101	0,0117
σ_e	estimativa	1,8732	1,9036	2,2052	2,2556	2,4347	2,5640
	erro padrão	0,1030	0,1056	0,1163	0,1161	0,1217	0,1320
	logL	-282,8734	-285,7735	-317,2342	-315,6864	-330,8449	-346,0075
$Y_t^m = \beta W_{t,(h)}^{m*} + e_t$							
β	estimativa	1,1018	1,0945	1,0943	1,0834	1,0792	1,0690
	erro padrão	0,0098	0,0095	0,0108	0,0100	0,0104	0,0110
σ_e	estimativa	1,9278	1,8889	2,1435	2,0000	2,1022	2,2298
	erro padrão	0,1246	0,1219	0,1386	0,1298	0,1368	0,1451
	logL	-485,1175	-480,3539	-509,9389	-493,7298	-505,3940	-519,1760
$Y_t^m = \beta_t W_{t,(h)}^{m*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,8406	0,8208	0,9039	0,8754	0,8891	0,8392
	erro padrão	0,0733	0,0703	0,0406	0,0560	0,0550	0,0821
μ	estimativa	1,1313	1,1257	1,1349	1,1033	1,0891	1,0811
	erro padrão	0,0229	0,0217	0,0310	0,0204	0,0184	0,0218
σ_ε	estimativa	0,0474	0,0504	0,0396	0,0308	0,0238	0,0449
	erro padrão	0,0142	0,0136	0,0109	0,0109	0,0099	0,0189
σ_e	estimativa	1,6253	1,5539	1,8290	1,8028	1,9684	1,9457
	erro padrão	0,1066	0,1058	0,1052	0,1006	0,1025	0,1326
	logL	-260,3443	-253,7488	-282,5510	-271,4519	-286,0662	-296,0774

Tabela 7.8: Comparação das medidas e de critérios de seleção dos quatro modelos sem constante aditiva da temperatura mínima diária.

	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^m = \beta W_{t,(h)}^m + e_t$						
REQM	2,0861	2,1089	2,4133	2,3630	2,5141	2,7117
EAM	1,6098	1,6694	1,8798	1,8224	1,9036	2,0441
EEAM	0,9084	0,9420	1,0608	1,0284	1,0742	1,1535
U-Theil	0,5364	0,5548	1,2234	0,8823	3,7574	**2,0140**
r^2	0,8726	0,8718	0,8301	0,8316	0,8066	0,7844
r_a^2	0,8726	0,8718	0,8301	0,8316	0,8066	0,7844
AIC	1012,1850	1017,2610	1080,3680	1070,5120	1099,5280	1134,9370
BIC	1019,0960	1024,1710	1087,2790	1077,4230	1106,4380	1141,8470
$Y_t^m = \beta_t W_{t,(h)}^m + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**2,0074**	**2,0339**	**2,3293**	**2,3236**	**2,4853**	**2,6540**
EAM	**1,5499**	**1,5874**	**1,8122**	**1,7610**	**1,8881**	**2,0015**
EEAM	**0,8746**	**0,8958**	**1,0226**	**0,9937**	**1,0655**	**1,1294**
U-Theil	**0,4717**	**0,4644**	**1,0817**	**0,8587**	**3,7498**	2,0199
r^2	**0,8767**	**0,8771**	**0,8391**	**0,8375**	**0,8132**	**0,7978**
r_a^2	**0,8767**	**0,8771**	**0,8391**	**0,8375**	**0,8132**	**0,7978**
AIC	**573,7468**	**579,5470**	**642,4684**	**639,3728**	**669,6898**	**700,0150**
BIC	**587,5681**	**593,3683**	**656,2897**	**653,1941**	**683,5111**	**713,8363**
$Y_t^m = \beta W_{t,(h)}^{m*} + e_t$						
REQM	1,9236	1,8849	2,1389	1,9957	2,0977	2,2250
EAM	1,5019	1,5161	1,6828	1,5894	1,6378	1,7467
EEAM	0,8475	0,8555	0,9496	0,8969	0,9242	0,9857
U-Theil	0,9550	0,9309	1,2280	0,9327	1,0275	0,9087
r^2	0,8813	0,8879	0,8526	0,8651	0,8454	0,8252
r_a^2	0,8813	0,8879	0,8526	0,8651	0,8454	0,8252
AIC	974,2350	964,7078	1023,8778	991,4596	1014,7880	1042,3520
BIC	981,1456	971,6184	1030,7884	998,3702	1021,6986	1049,2626
$Y_t^m = \beta_t W_{t,(h)}^{m*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,8407**	**1,7881**	**2,0219**	**1,9331**	**2,0607**	**2,1735**
EAM	**1,4376**	**1,4064**	**1,5815**	**1,5343**	**1,6218**	**1,7056**
EEAM	**0,8112**	**0,7936**	**0,8925**	**0,8658**	**0,9152**	**0,9625**
U-Theil	**0,8476**	**0,8049**	**1,0390**	**0,8557**	**0,9776**	**0,8312**
r^2	**0,8863**	**0,8942**	**0,8650**	**0,8723**	**0,8515**	**0,8346**
r_a^2	**0,8863**	**0,8942**	**0,8650**	**0,8723**	**0,8515**	**0,8346**
AIC	**528,6886**	**515,4976**	**573,1020**	**550,9038**	**580,1324**	**600,1548**
BIC	**542,5099**	**529,3189**	**586,9233**	**564,7251**	**593,9537**	**613,9761**

Por observação gráfica da Figura 7.28, onde estão representadas as séries da temperatura mínima observada com as previsões a h-passos do *website weatherstack*

e respetivas previsões a h-passos calibrada, correspondentes aos modelos MEES_1 (parte superior) e MEES_6 (parte inferior), percebe-se que nas duas representações gráficas, tanto a previsão como a previsão calibrada acompanham o comportamento da série da temperatura mínima observada. No gráfico inferior, verifica-se que o modelo não conseguiu calibrar a previsão relativa ao mês de fevereiro, tendo em conta que a série correspondente às previsões a 6-passos calibrada está sobreposta à série das previsões a 6-passos.

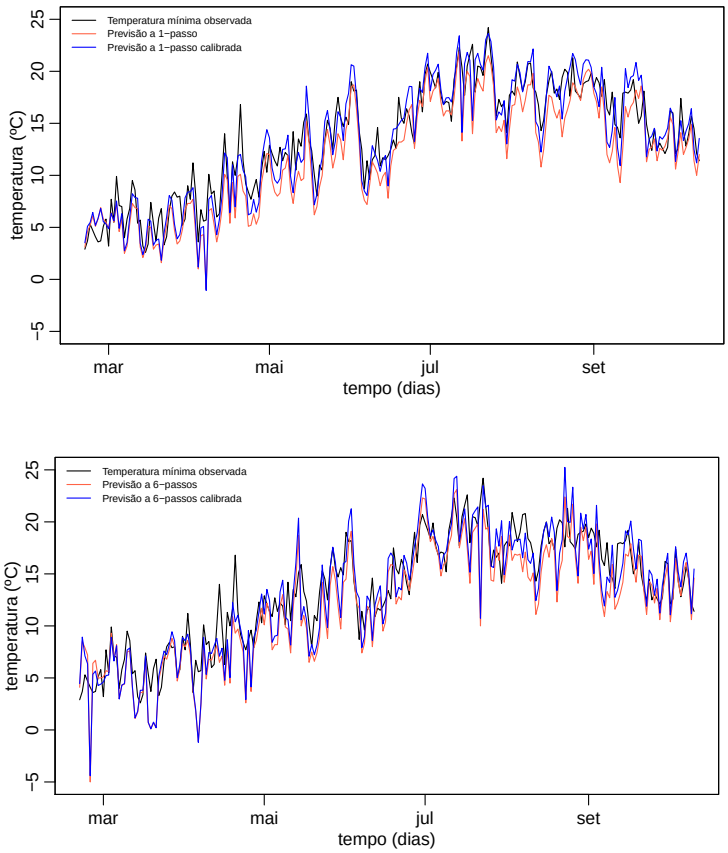


Figura 7.28: Séries da temperatura mínima observada (a preto), previsões do *website* (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEES_1, gráfico inferior - MEES_6.

Na parte superior da Figura 7.29 estão representadas as previsões do fator de calibração, $\beta_{t|t-1}$, para $t = 1, \ldots, 234$ dias, como uma linha contínua, obtidas através do FK, com os limites do erro, $\beta_{t|t-1} \pm 1,96\sqrt{P_{t|t-1}}$, como linhas tracejadas. No gráfico do meio estão representados os valores filtrados, $\beta_{t|t}$, para $t = 1, \ldots, 234$ dias, como uma linha e os limites do erro, $\beta_{t|t} \pm 1,96\sqrt{P_{t|t}}$ como linhas tracejadas. No gráfico inferior estão representados, como uma linha, o fator de calibração alisado, $\beta_{t|n}$ para $t = 1, \ldots, 234$ dias e $n = 234$, obtidos através do alisamento de Kalman e os limites do erro, $\beta_{t|n} \pm 1,96\sqrt{P_{t|n}}$. Através da análise gráfica, verifica-se que

os comportamentos de $\beta_{t|t-1}$ (previsão), $\beta_{t|t}$ (filtragem) e $\beta_{t|n}$ (alisamento) correspondentes ao modelo de calibração a 6-passos (direita), aparentam ser ligeiramente mais suaves do que os do modelo a 1-passo (esquerda). Em relação aos limites do erro, percebe-se que não existe uma diferença relevante entre os respetivos gráficos. Além disso, também é possível notar que, para ambos os modelos, as estimativas de β_t parecem oscilar entre 0,8 e 1,2.

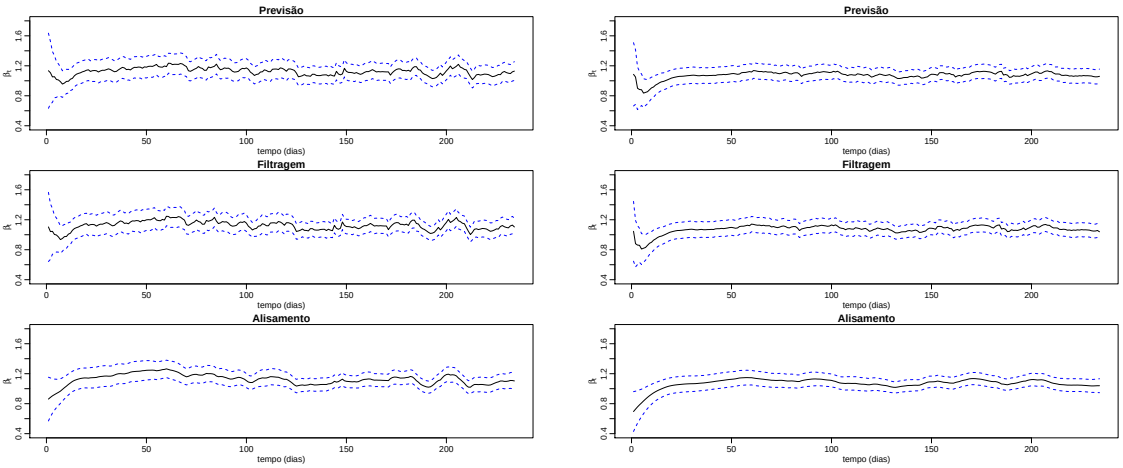


Figura 7.29: Na parte superior - previsões $\beta_{t|t-1}$; no centro - filtragem $\beta_{t|t}$; na parte inferior - alisamento $\beta_{t|n}$ da temperatura mínima. Lado esquerdo - MEES_1, lado direito - MEES_6.

A fim de validar um modelo de previsão, é necessário proceder-se com a análise das inovações a fim de verificar se têm um comportamento similar ao de um ruído branco.

Analisando o histograma dos resíduos (Figura 7.30) correspondente ao modelo MEES_1 (lado esquerdo), sugere que os resíduos apresentam uma assimetria à esquerda. Além disso, o respetivo *QQ-plot* apresenta pontos que estão ligeiramente afastados de uma reta; no entanto, o teste Kolmogorov-Smirnov não rejeita a hipótese nula da normalidade dos erros (valor de prova de 0,1627). O mesmo acontece no modelo de calibração para $h = 6$ dias (lado direito), cujo valor de prova do teste Kolmogorov-Smirnov é de 0,2943.

Quanto ao pressuposto da média nula dos erros, efetuou-se o teste t para o valor esperado, que não rejeitou a hipótese da média nula dos erros referente ao modelo de calibração a 1-passo (valor de prova de 0,0636); no entanto, a hipótese é rejeitada para o modelo a 6-passos (valor de prova de 0,0062). Relativamente ao pressuposto da homocedasticidade dos erros, a análise gráfica sugere que, para ambos os modelos, a variância não é constante.

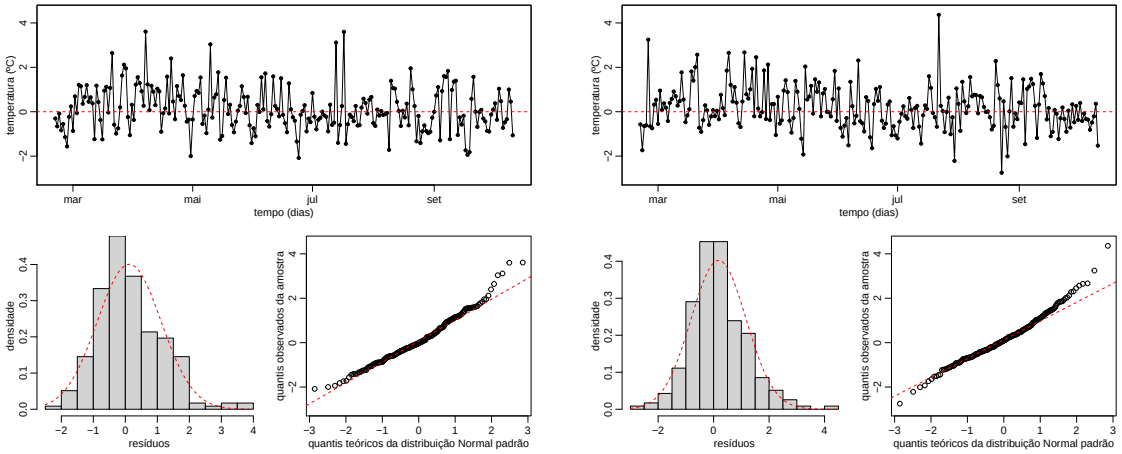


Figura 7.30: Série, histograma e *QQ-plot* dos resíduos da temperatura mínima; lado esquerdo - MEES_1, lado direito - MEES_6.

Relativamente ao pressuposto da independência dos erros, efetuou-se o teste Ljung-Box à série dos resíduos (inovações), fazendo variar k (número de autocorrelações) entre 7 e 17, com o objetivo de testar se as primeiras k autocorrelações são conjuntamente nulas. Segundo os resultados deste teste, a hipótese de independência dos erros relativos ao modelo a 1-passo é rejeitada para todos os valores de k, exceto para $k = 15$ (valor de prova de 0,0543) e $k = 17$ (valor de prova de 0,0590) e também é rejeitada para o modelo a 6-passos apenas para $k = 7$ (valor de prova de 0,0250).

A FAC e a FACP da série dos resíduos (Figura 7.31) apresentam correlações significativas para ambos os modelos de calibração. Portanto, o pressuposto da independência não é verificado.

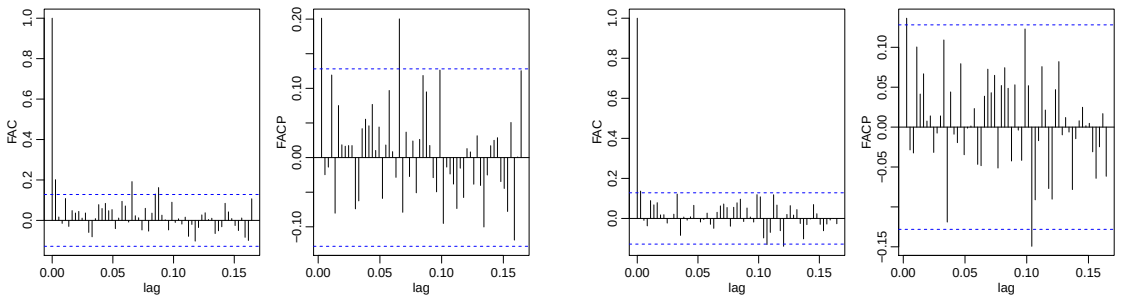


Figura 7.31: FAC e FACP dos resíduos da temperatura mínima; lado esquerdo - MEES_1, lado direito - MEES_6.

Tendo em conta à Figura 7.32, verifica-se que, no geral, ambos os modelos melhoraram as previsões do *website*, dado que ambas as séries das previsões calibradas

parecem estar mais próximas do comportamento da temperatura mínima observada.

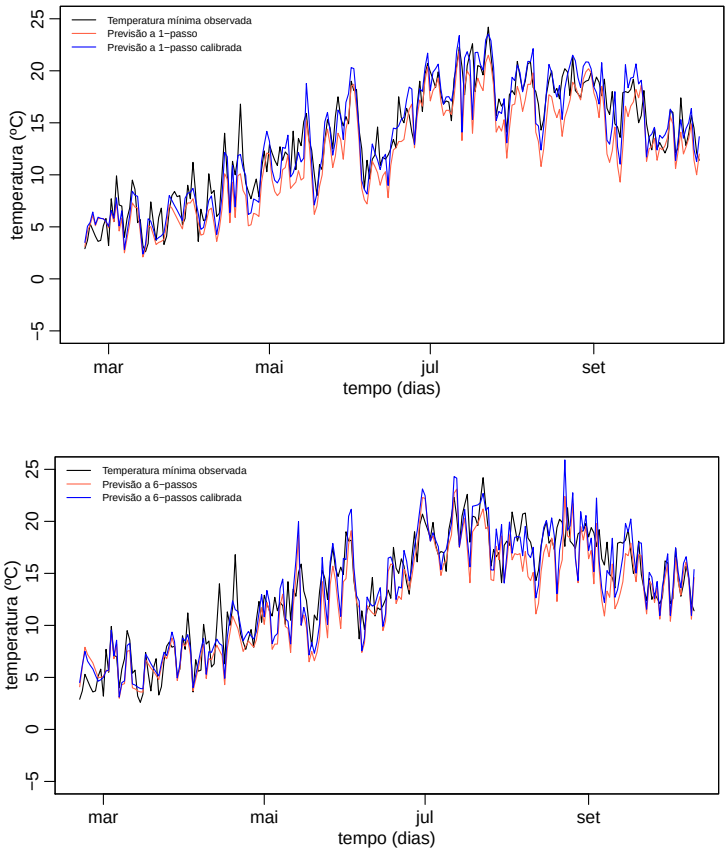


Figura 7.32: Séries da temperatura mínima observada (a preto), previsões do *website* a h-passos (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEES_1^*, gráfico inferior - MEES_6^*.

Comparando os gráficos da Figura 7.33, verifica-se que não existe uma característica relevante que os permita discernir, nomeadamente em relação tanto à amplitude das margens de erro como do seu próprio comportamento ao longo do tempo. Verifica-se que tendem a ficar mais irregulares ao longo do tempo. Também percebe-se que os gráficos relativos às previsões e às filtragens possuem maior variabilidade no início em comparação com o gráfico do alisamento. Isto deve-se à quantidade de informação disponível, uma vez que as previsões ($\beta_{t|t-1}$) utilizam toda a informação passada e a filtragem ($\beta_{t|t}$) utiliza toda a informação até ao momento, inclusive.

De acordo com a análise da Figura 7.34, verifica-se que o histograma associado aos resíduos do modelo a 1-passo (esquerda) apresenta uma assimetria à esquerda, não sendo o suficiente para rejeitar a hipótese da normalidade dos erros do teste Kolmogorov-Smirnov (valor de prova de 0,4686); o teste também não rejeita a hipótese nula da distribuição Normal dos erros para o modelos a 6-passos (valor de prova

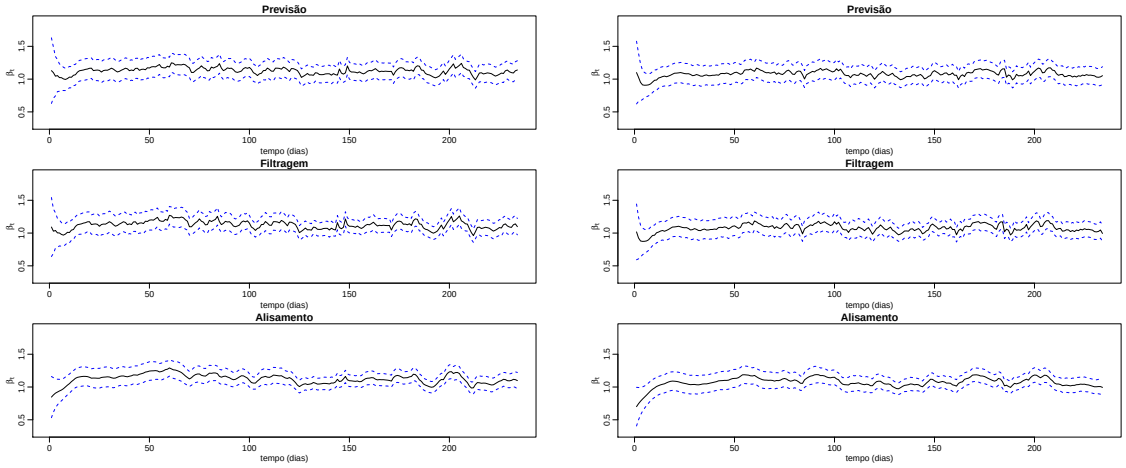


Figura 7.33: Na parte superior - previsões $\beta_{t|t-1}$; no centro - filtragem $\beta_{t|t}$; na parte inferior - alisamento $\beta_{t|n}$ da temperatura mínima; lado esquerdo -MEES$_1^*$, lado direito - MEES$_6^*$.

de 0,7959). Analisando o *QQ-plot* para ambos os modelos, verifica-se que os pontos encontram-se alinhados, havendo alguns pontos no extremo que se não se encontram alinhados, principalmente no modelo a 1-passo.

De acordo com as representações gráficas da série dos resíduos, estas parecem oscilar em torno de zero, sendo comprovada pelo teste t para o valor esperado (valor de prova de 0,5125 para o modelo a 1-passo e 0,4886 para o modelo a 6-passos). Relativamente à variância dos erros, a análise gráfica sugere que os resíduos não parecem oscilar uniformemente em torno de zero.

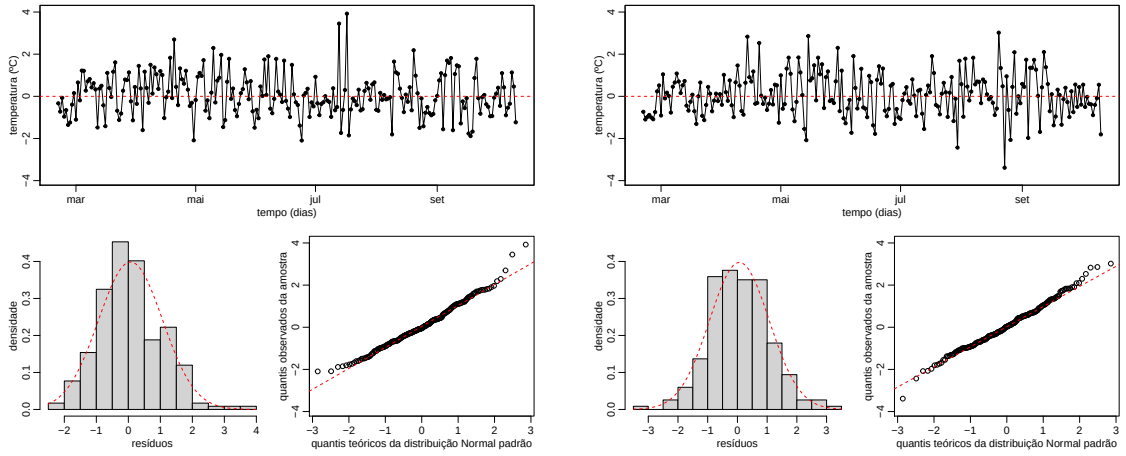


Figura 7.34: Série, histograma e *QQ-plot* dos resíduos da temperatura mínima; lado esquerdo - MEES$_1^*$, lado direito - MEES$_6^*$.

Para verificar o pressuposto da independência dos erros, recorreu-se ao teste Ljung-Box, para testar se as primeiras k autocorrelações, onde k varia entre 7 a 17,

eram conjuntamente nulas. Segundo os resultados do teste, a hipótese da independência para o modelo a 1-passo não é rejeitada para nenhum valor de k, apresentando valores de prova entre 0,0615 ($k = 7$) e 0,3752 ($k = 17$). No entanto, para o modelo a 6-passos, a hipótese da independência dos erros é rejeitada para todos os valores de k. No entanto, a FAC e a FACP dos resíduos (Figura 7.14) apresentam correlações significativas para alguns lags para ambos os modelos a 1-passo (esquerda) e a 6-passos (direita).

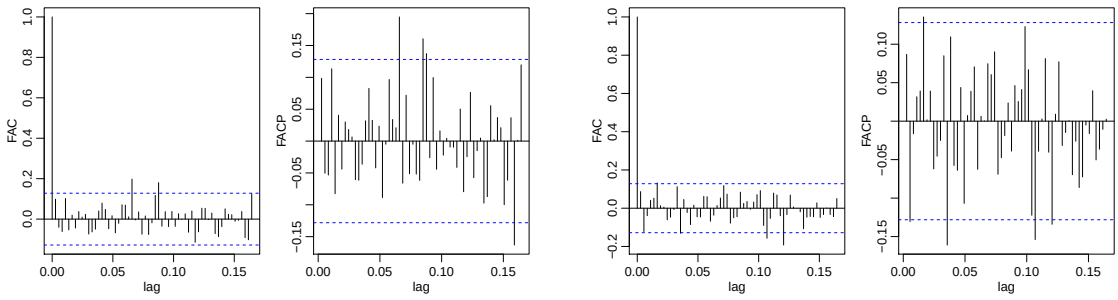


Figura 7.35: FAC e FACP dos resíduos da temperatura mínima. Lado esquerdo - MEES_1^*, lado direito - MEES_6^*.

Explorando a Tabela 7.9, percebe-se que

- Nos quatro modelos, a log-verosimilhança tende a diminuir com o aumento do horizonte temporal. Por outro lado, as estimativas e os erros padrão de σ_e tendem a aumentar;

- O parâmetro α é a componente aditiva constante e representa o erro sistemático, na qual verifica-se que não sofre alterações muito significativas com o aumento do horizonte temporal. Além disso, as estimativas de β, no caso dos modelos de regressão linear, e as estimativas de μ, no caso dos modelos de calibração, tendem a diminuir com o aumento do horizonte temporal, exceto no modelo RLSC_h^*, onde as estimativas de β pareces não ser afetadas pelo aumento do horizonte temporal;

- Relativamente aos modelos RLSC_h e MEEC_h, é possível ver que as estimativas do valor médio do processo $\{\beta_t\}_{t=1,\ldots,n}$, μ variam entre os 0,71-0,82°C, enquanto que no modelo de regressão linear simples, as estimativas de β variam entre os 0,85-0,93°C. Neste caso, ambos representam a mudança na média de Y_t por cada aumento unitário nas previsões. Por outro lado, as estimativas e os erros padrão de σ_e são mais elevados no modelo de regressão linear simples;

– Já nos modelos MEEC$_h^*$ e RLSC$_h^*$, as estimativas do valor médio do processo $\{\beta_t\}_{t=1,\ldots,n}$, μ variam entre os 0,78-0,87°C, enquanto que no modelo de regressão linear, as estimativas de β variam entre os 0,94-0,97°C. Ambas as estimativas representam a mudança na média de Y_t por cada aumento unitário nas previsões. Além disso, as estimativas e os erros padrão de σ_e são mais elevados no modelo de regressão linear.

Na Tabela 7.10 estão apresentadas algumas medidas e critérios de seleção referentes aos quatro modelos cujas estimativas dos parâmetros encontram-se na Tabela 7.9. De um modelo geral, conclui-se que

– Nos quatro modelos, à medida que o horizonte temporal aumenta, os critérios AIC e BIC tendem a aumentar, assim como as medidas REQM, EAM e EEAM. O coeficiente de determinação, r^2, e o coeficiente de determinação ajustado, r_a^2, tendem a diminuir. Já a medida de precisão U-Theil apresenta um comportamento oscilatório;

– O modelo MEEC$_h$ obteve menores valores de AIC e BIC comparativamente ao modelo RLSC$_h$, assim como obteve maioritariamente os menores valores de REQM, EAM e EEAM. Também obteve melhor ajustamento, visto que produziu maior r_a^2. No entanto, o modelo de regressão linear simples obteve os menores valores para U-Theil;

– O modelo MEEC$_h^*$ produziu menores valores de AIC e BIC, assim como obteve os menores valores de REQM, EAM e EEAM do que o modelo RLSC$_h^*$. Também obteve maior r_a^2. No entanto, o modelo de regressão linear simples obteve menores valores para a estatística U-Theil;

– Relativamente aos modelos MEEC$_h$ e MEEC$_h^*$, verifica-se que U-Theil$>$ 1 e, portanto, o método *naïve* (onde a última observação é utilizada como previsão) é mais eficiente do que o método em avaliação.

Tabela 7.9: Estimativas dos parâmetros e respetivos erros padrão dos quatro modelos com a componente aditiva determinística α para a série da temperatura mínima, onde Y_t^m representa a temperatura mínima observada, $W_{t,(h)}$ corresponde às previsões a h-passos do *website* e $W_{t,(h)}^{m*}$ corresponde às previsões com a substituição dos *outliers* do rácio $Y_t^m/W_{t,(h)}^m$, $h = 1, \ldots, 6$ dias.

		h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^m = \alpha + \beta W_{t,(h)}^m + e_t$							
α	estimativa	2,4671	2,5456	2,8475	2,6601	2,7610	3,1641
	erro padrão	0,2997	0,2989	0,3430	0,3464	0,3731	0,3859
β	estimativa	0,9274	0,9172	0,8979	0,9033	0,8931	0,8532
	erro padrão	0,0233	0,0231	0,0267	0,0267	0,0287	0,0294
σ_e	estimativa	1,8431	1,8486	2,1282	2,1191	2,2711	2,3980
	erro padrão	0,1202	0,1206	0,1388	0,1382	0,1481	0,1564
	logL	-474,1104	-474,8003	-507,7579	-506,7541	-522,9656	-535,6889
$Y_t^m = \alpha + \beta_t W_{t,(h)}^m + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,9675	0,9690	0,9700	0,9724	0,9751	0,9739
	erro padrão	0,0090	0,0075	0,0079	0,0074	0,0070	0,0088
α	estimativa	4,1287	4,2973	4,7340	4,8140	5,2705	5,4145
	erro padrão	0,4140	0,3999	0,4585	0,4573	0,4647	0,4211
μ	estimativa	0,8222	0,8118	0,7847	0,7727	0,7375	0,7097
	erro padrão	0,0396	0,0361	0,0427	0,0476	0,0566	0,0677
σ_ε	estimativa	0,0119	0,0096	0,0112	0,0127	0,0150	0,0208
	erro padrão	0,0054	0,0361	0,0062	0,0056	0,0053	0,0075
σ_e	estimativa	1,5809	1,5634	1,7964	1,7868	1,8703	1,8856
	erro padrão	0,0807	0,0795	0,0922	0,0906	0,0933	0,0993
	logL	-237,5923	-234,4363	-267,2617	-266,5849	-278,1084	-283,2522
$Y_t^m = \alpha + \beta W_{t,(h)}^{m*} + e_t$							
α	estimativa	1,9337	1,9661	2,1240	1,7043	1,5170	1,5654
	erro padrão	0,2999	0,2899	0,3341	0,3278	0,3591	0,3846
β	estimativa	0,9638	0,9553	0,9434	0,9632	0,9721	0,9592
	erro padrão	0,0232	0,0223	0,0258	0,0250	0,0273	0,0290
σ_e	estimativa	1,7790	1,7293	1,9823	1,8969	2,0302	2,1588
	erro padrão	0,1160	0,1128	0,1293	0,1237	0,1324	0,1408
	logL	-465,8288	-459,1902	-491,1491	-480,8334	-496,7269	-511,1053
$Y_t^m = \alpha + \beta_t W_{t,(h)}^{m*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$							
ϕ	estimativa	0,9647	0,9650	0,9643	0,9696	0,9753	0,9732
	erro padrão	0,0114	0,0102	0,0105	0,0093	0,0079	0,0113
α	estimativa	3,4283	3,5093	3,8176	3,6473	3,9214	4,1443
	erro padrão	0,4694	0,4461	0,5011	0,5182	0,5628	0,5754
μ	estimativa	0,8672	0,8592	0,8389	0,8431	0,8211	0,7830
	erro padrão	0,0409	0,0373	0,0424	0,0457	0,0560	0,0747
σ_ε	estimativa	0,0121	0,0104	0,0120	0,0122	0,0134	0,0230
	erro padrão	0,0063	0,0066	0,0070	0,0060	0,0052	0,0086
σ_e	estimativa	1,5732	1,5183	1,6955	1,6486	1,7606	1,7785
	erro padrão	0,0822	0,0809	0,0892	0,0854	0,0888	0,0996
	logL	-235,1674	-226,2863	-253,4555	-246,8862	-263,1579	-273,4144

Tabela 7.10: Comparação das medidas e dos critérios de seleção dos quatro modelos com constante aditiva para a série da temperatura mínima.

	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^m = \alpha + \beta W_{t,(h)}^m + e_t$						
REQM	1,8352	1,8407	2,1190	2,1100	2,2613	2,3877
EAM	1,4030	1,4205	1,5760	1,5732	1,6969	1,8765
EEAM	0,7917	0,8016	0,8893	0,8878	0,9576	1,0589
U-Theil	**0,5955**	**0,8592**	**1,1117**	**0,9649**	**1,0375**	**0,6035**
r^2	0,8726	0,8718	0,8301	0,8316	0,8066	0,7844
r_a^2	0,8721	0,8712	0,8294	0,8309	0,8058	0,7835
AIC	954,2208	955,6006	1021,5158	1019,5082	1051,9312	1077,3778
BIC	964,5868	965,9666	1031,8818	1029,8742	1062,2972	1087,7438
$Y_t^m = \alpha + \beta_t W_{t,(h)}^m + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,6738**	**1,6610**	**1,9091**	**1,9025**	**1,9919**	**2,0424**
EAM	**1,2841**	**1,2992**	**1,4430**	**1,4352**	**1,5166**	**1,5425**
EEAM	**0,7246**	**0,7331**	**0,8143**	**0,8099**	**0,8558**	**0,8704**
U-Theil	1,0599	1,1085	1,2822	1,2189	1,2703	1,2426
r^2	**0,8941**	**0,8958**	**0,8623**	**0,8633**	**0,8501**	**0,8423**
r_a^2	**0,8936**	**0,8954**	**0,8617**	**0,8627**	**0,8495**	**0,8416**
AIC	**485,1846**	**478,8726**	**544,5234**	**543,1698**	**566,2168**	**576,5044**
BIC	**502,4612**	**496,1492**	**561,8000**	**560,4464**	**583,4934**	**593,7810**
$Y_t^m = \alpha + \beta W_{t,(h)}^{m*} + e_t$						
REQM	1,7714	1,7219	1,9739	1,8887	2,0215	2,1496
EAM	1,3548	1,3376	1,4711	1,4694	1,6036	1,7200
EEAM	0,7645	0,7548	0,8302	0,8292	0,9049	0,9706
U-Theil	**1,0624**	**1,0055**	**1,2367**	**1,0490**	**1,1426**	**1,0324**
r^2	0,8813	0,8879	0,8526	0,8651	0,8454	0,8252
r_a^2	0,8808	0,8874	0,8520	0,8645	0,8447	0,8244
AIC	937,6576	924,3804	988,2982	967,6668	999,4538	1028,2106
BIC	948,0236	934,7464	998,6642	978,0328	1009,8198	1038,5766
$Y_t^m = \alpha + \beta_t W_{t,(h)}^{m*} + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$						
REQM	**1,6565**	**1,5996**	**1,8007**	**1,7486**	**1,8696**	**1,9606**
EAM	**1,2586**	**1,2495**	**1,3653**	**1,3688**	**1,4912**	**1,5334**
EEAM	**0,7102**	**0,7051**	**0,7704**	**0,7724**	**0,8415**	**0,8653**
U-Theil	1,2306	1,1423	1,3212	1,2534	1,4343	1,3065
r^2	**0,8963**	**0,9033**	**0,8775**	**0,8845**	**0,8679**	**0,8547**
r_a^2	**0,8959**	**0,9029**	**0,8770**	**0,8840**	**0,8673**	**0,8541**
AIC	**480,3348**	**462,5726**	**516,9110**	**503,7724**	**536,3158**	**556,8288**
BIC	**497,6114**	**479,8492**	**534,1876**	**521,0490**	**553,5924**	**574,1054**

Tendo em conta a Figura 7.36, parece verificar-se uma melhoria das previsões calibradas em relação às previsões dadas pelo *website*. É de notar que o modelo de calibração a 6-passos conseguiu melhorar a previsão no mês de fevereiro.

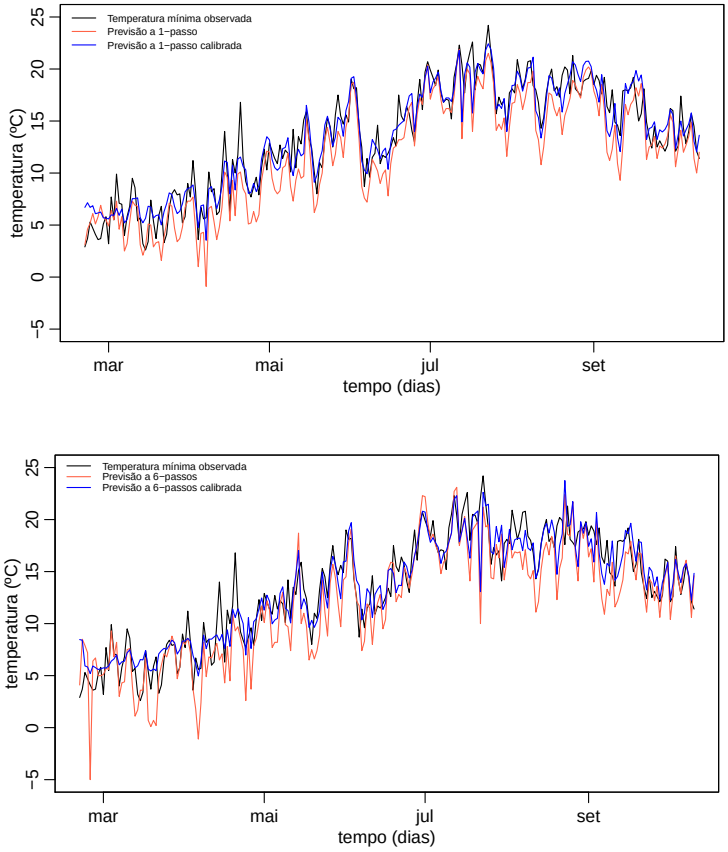


Figura 7.36: Séries da temperatura mínima observada (a preto), previsões do *website* (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEEC_1, gráfico inferior - MEEC_6.

Através da análise comparativa dos gráficos da Figura 7.37, verifica-se que os intervalos (limites do erro) para o modelo de calibração com $h = 6$ dias (lado direito) apresentam margens de erro ligeiramente superiores comparativamente aos intervalos para o modelo de calibração a 1-passo (lado esquerdo). Além disso, o comportamento dos três gráficos relativos ao modelo MEEC_1 apresentam um comportamento mais suave.

Relativamente à análise dos resíduos, o histograma dos resíduos (Figura 7.30) correspondente ao modelo MEEC_1 (lado esquerdo), sugere que os resíduos apresentam uma distribuição aproximadamente assimétrica à esquerda. No *QQ-plot* existem pontos que estão afastados de uma reta. No entanto, o teste Kolmogorov-Smirnov não rejeita a hipótese da normalidade dos erros (valor de prova de 0,7497). No caso do *QQ-plot* relativo ao modelo MEEC_1 (lado direito), também possui pontos que

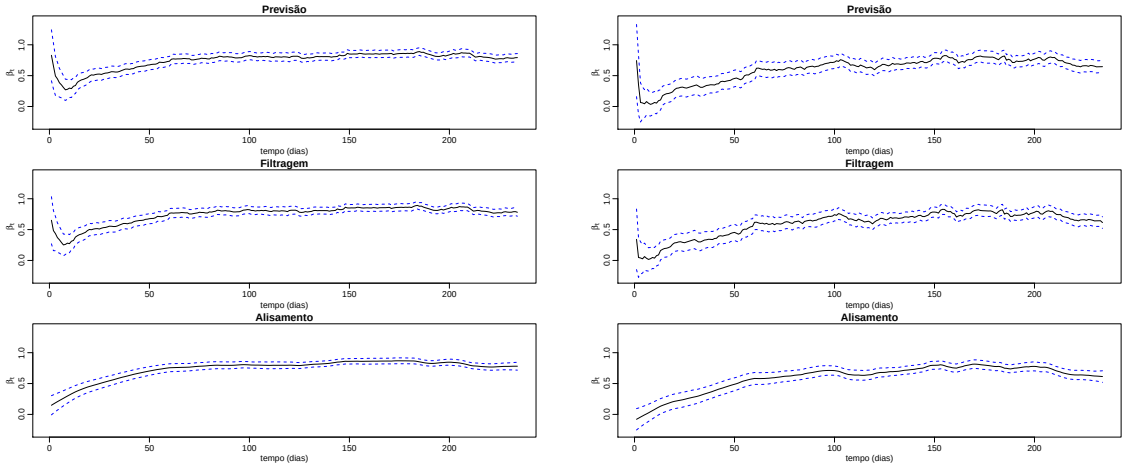


Figura 7.37: Na parte superior - previsões $\beta_{t|t-1}$; no centro - filtragem $\beta_{t|t}$; na parte inferior - alisamento $\beta_{t|n}$ da temperatura mínima; lado esquerdo - MEEC_1, lado direito - MEEC_6.

se afastam da reta, porém, o teste de Kolmogorov-Smirnov também não rejeita a hipótese da distribuição Normal dos erros (valor de prova de 0,6514). Analisando o gráfico da série dos resíduos de ambos os modelos, parecem apresentar uma distribuição em torno de zero, não apontando para a rejeição da média nula dos erros. O teste t ao valor esperado apresenta um valor de prova de 0,7380 para o modelo a 1-passo e 0,8145 para o modelo a 6-passos, pelo que resulta na não rejeição da hipótese de média nula. Relativamente ao pressuposto da homocedasticidade dos erros, a análise gráfica das séries dos resíduos sugerem que, para ambos os modelos, a variância não é constante.

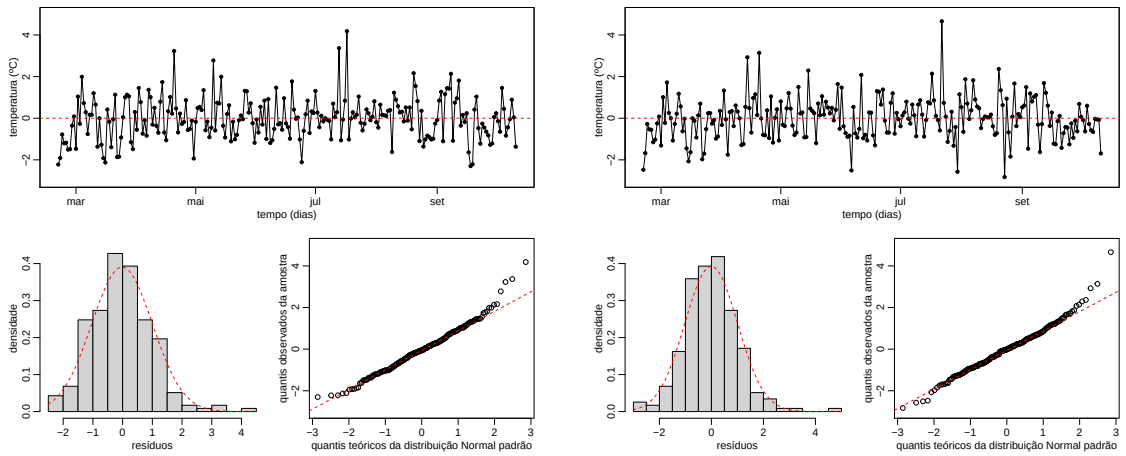


Figura 7.38: Série, histograma e *QQ-plot* dos resíduos da temperatura mínima; lado esquerdo - MEEC_1, lado direito - MEEC_6.

O teste Ljung-Box é aplicado à série dos resíduos (inovações) para verificar o pressuposto da independência dos erros, onde k, que representa o número de auto-correlações a serem testadas, varia entre 7 e 17. Conforme os resultados do teste, a hipótese da independência dos erros para o modelo MEEC_1 é rejeitada para todos os valores de k, exceto para $k = 17$ (valor de prova de 0,0511). Para o modelo MEEC_6, a hipótese da independência dos erros é rejeitada para todos os valores de k. De facto, a FAC e a FACP dos resíduos (Figura 7.39) apresentam correlações significativas para alguns lags em ambos os modelos de calibração. Portanto, o pressuposto da independência não é verificado.

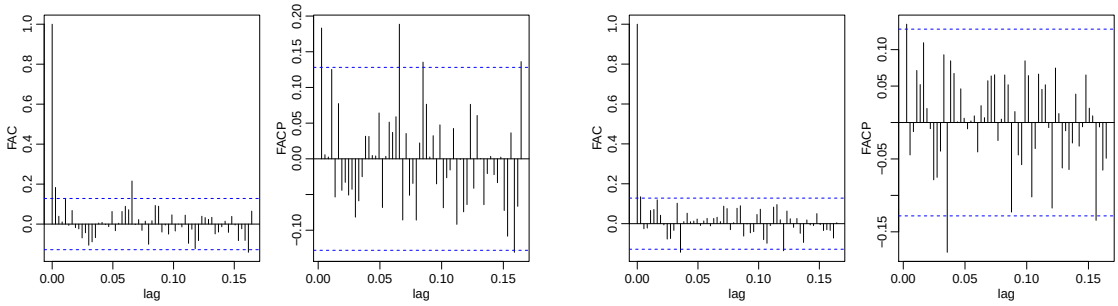


Figura 7.39: FAC e FACP dos resíduos da temperatura mínima; lado esquerdo - MEEC_1, lado direito - MEEC_6.

Na Figura 7.40 estão representadas as séries da temperatura mínima observada, das previsões a h-passos, obtidas no *website weatherstack* com a substituição dos *outliers* do rácio $Y_t^m/W_{t,(h)}^m$ por estimativas através da interpolação linear, e das respetivas previsões a h-passos calibrada, cujo gráfico superior corresponde ao modelo MEEC_1^* e o gráfico inferior ao modelo MEEC_6^*.

Através da análise gráfica, verifica-se uma melhoria das previsões calibradas em relação às previsões dadas pelo *website* com a substituição de *outliers* do rácio $Y_t^m/W_{t,(h)}^m$, uma vez que ambas estão mais próximas do comportamento da série da temperatura mínima observada.

Comparando os gráficos da Figura 7.41, verifica-se que o comportamento dos gráficos da previsão ($\beta_{t|t-1}$), filtragem ($\beta_{t|t}$) e alisamento ($\beta_{t|n}$) correspondentes ao modelo a 1-passo (esquerda) parecem ser mais suaves do que os do modelo a 6-passos (direita). Além disso, os três gráficos correspondentes ao modelo a 6-passos parecem apresentar amplitudes de erro maiores comparativamente ao modelo a 1-passo.

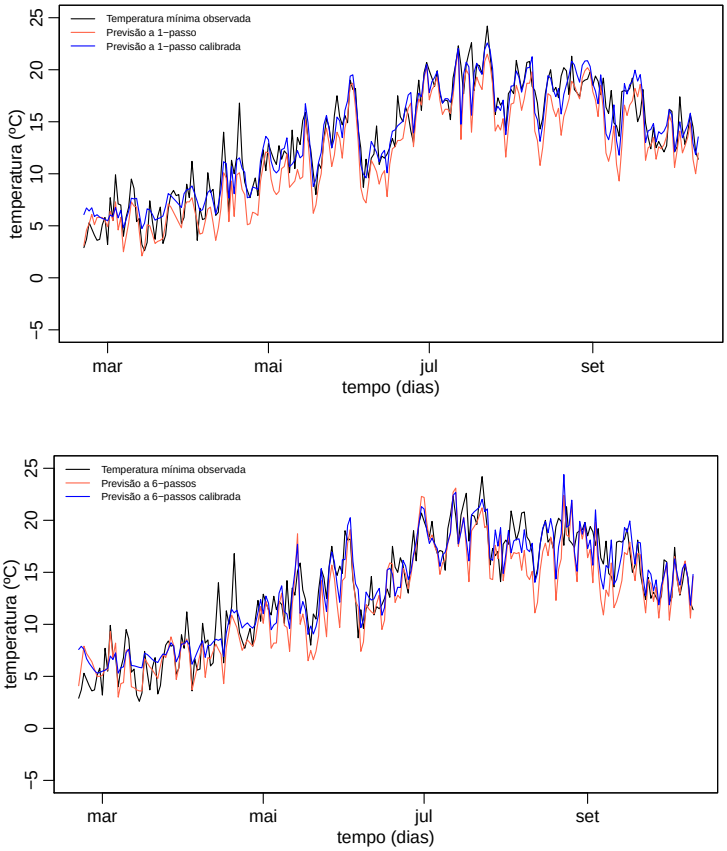


Figura 7.40: Séries da temperatura mínima observada (a preto), previsões do *website* com a substituição dos *outliers* do rácio $Y_t^m/W_{t,(h)}^m$ (a vermelho) e das previsões calibrada (a azul); gráfico superior - MEEC_1^*, gráfico inferior - MEEC_6^*.

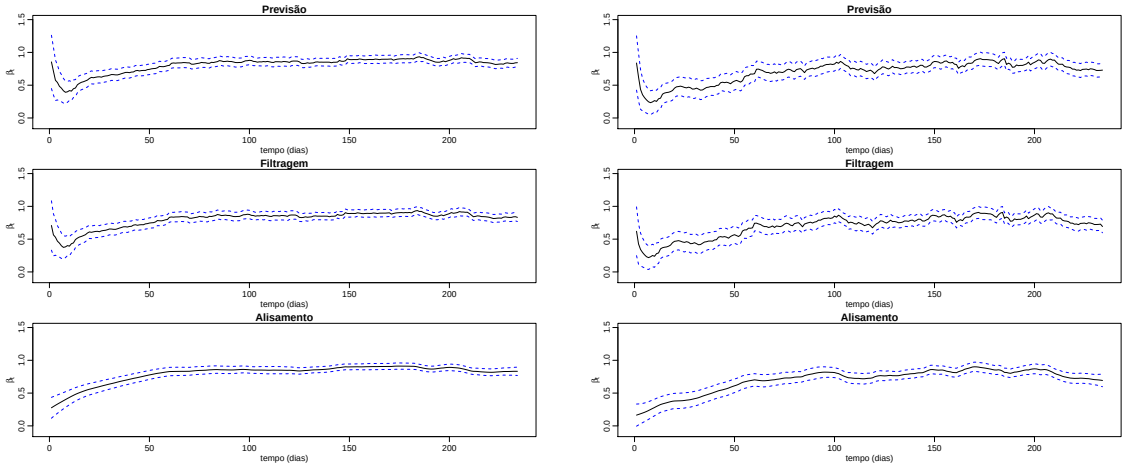


Figura 7.41: Na parte superior - previsões $\beta_{t|t-1}$; no centro - filtragem $\beta_{t|t}$; na parte inferior - alisamento $\beta_{t|n}$ da temperatura mínima; lado esquerdo - MEEC_1^*, lado direito - MEEC_6^*.

De acordo com a análise da Figura 7.42, verifica-se que o histograma associado aos resíduos do modelo a 1-passo (esquerda) apresenta uma cauda evidente à direita, que, no entanto, não foi suficiente para rejeitar a hipótese da normalidade dos erros do teste Kolmogorov-Smirnov (valor de prova de 0,5435); o histograma dos resíduos associado ao modelo a 6-passos (direita) sugere que os resíduos têm uma distribuição simétrica, sendo comprovado pelo teste Kolmogorov-Smirnov (valor de prova de 0,9815). Analisando o *QQ-plot* para ambos os modelos, verifica-se que nem todos os pontos encontram-se alinhados, principalmente no modelo MEEC_1^*, havendo pontos no extremo que se distanciam da reta.

De acordo com a representação gráfica da série de ambos os resíduos (Figura 7.42), estes parecem oscilar em torno de zero, sendo pelo teste t para o valor esperado (valor de prova de 0,7615 para o modelo a 1-passo e 0,7649 para o modelo a 6-passos). Relativamente à variância dos erros, a análise gráfica sugere que a variabilidade dos resíduos não aparentam apresentar um comportamento constante ao longo do tempo.

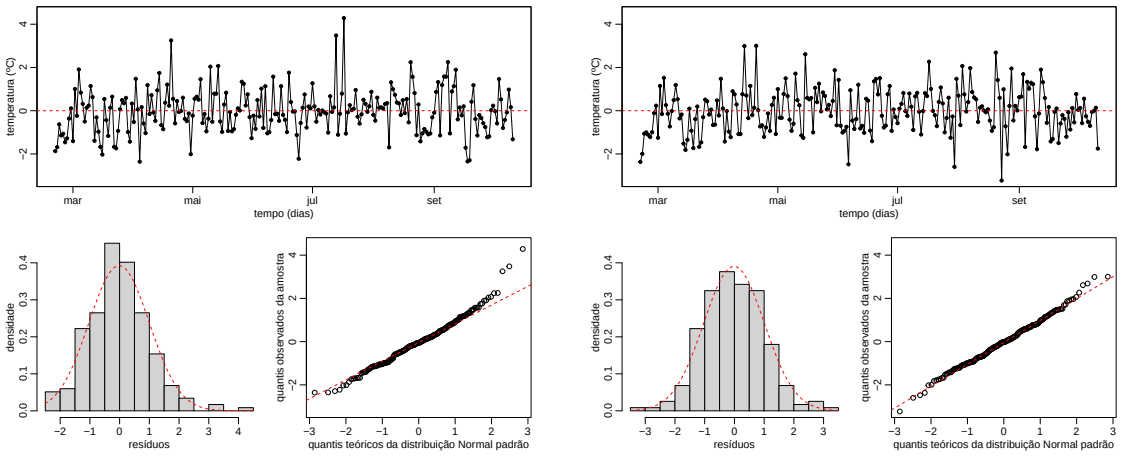


Figura 7.42: Série, histograma e *QQ-plot* dos resíduos da temperatura mínima; lado esquerdo - MEEC_1^*, lado direito - MEEC_6^*.

Segundo os resultados do teste Ljung-Box (cujo número de autocorrelações k testadas variam entre 7 e 17), a hipótese da independência dos erros é rejeitada para ambos os modelos, para quaisquer valores de k.

A FAC e a FACP dos resíduos de ambos os modelos (MEEC_1^* e MEEC_6^*) estão representadas na Figura 7.22. Verifica-se que ambos apresentam correlações significativas, uma vez que algumas correlações não de encontram dentro dos intervalos de confiança de 95%.

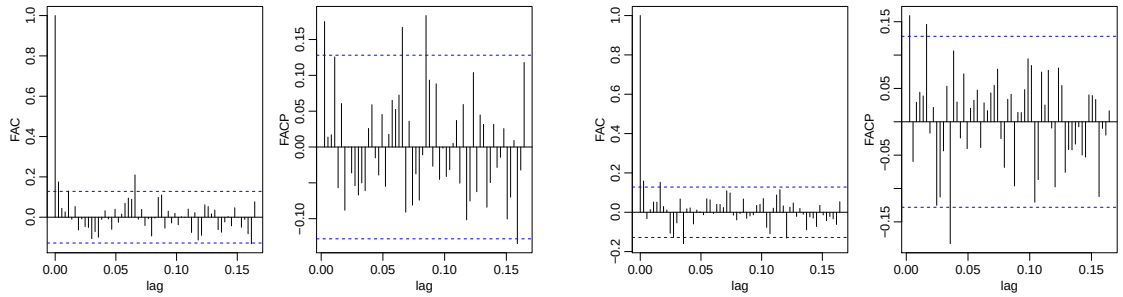


Figura 7.43: FAC e FACP dos resíduos da temperatura mínima; lado esquerdo - MEEC$_1^*$, lado direito - MEEC$_6^*$.

Considerações adicionais

Através da análise comparativa das Tabelas 7.7 e 7.9, conclui-se que

- O modelo MEES$_h$ obteve menores estimativas para ϕ do que o modelo MEEC$_h$, assim como o modelo MEES$_h^*$ obteve menores estimativas para o mesmo parâmetro do que o modelo MEEC$_h^*$;
- O modelo de regressão linear RLSC$_h$ obteve menores estimativas para σ_e do que o modelo RLSS$_h$. Analogamente, o modelo RLSC$_h^*$ obteve menores estimativas para σ_e do que o modelo RLSS$_h^*$.

Comparando as Tabelas 7.8 e 7.10, conclui-se que

- Tendo em conta aos modelos RLSS$_h$, RLSC$_h$, MEES$_h$ e MEEC$_h$, o que obteve menores valores de AIC e BIC, assim como produziu maioritariamente menores valores para as medidas REQM, EAM, EEAM e U-Theil e maior r_a^2 foi o modelo MEEC$_h$;
- Relativamente aos modelos RLSS$_h^*$, RLSC$_h^*$, MEES$_h^*$ e MEEC$_h^*$, o que obteve menores valores de AIC e BIC, assim como produziu maioritariamente menores valores para as medidas REQM, EAM, EEAM e U-Theil e maior r_a^2 foi o modelo MEEC$_h^*$;
- Por um lado, verificou-se que o modelo RLSC$_h$ gerou menores valores de REQM, EAM, EEAM do que o modelo MEES$_h$. No entanto, o modelo MEES$_h$ obteve maiores valores de r_a^2 e menores valores de AIC e BIC. Esta situação é idêntica para os modelos RLSC$_h^*$ e MEES$_h^*$;
- Todos os modelos obtiveram valores elevados de r_a^2, uma vez que todos são superiores a 0,75.

Relativamente ao cumprimento dos pressupostos sobre os erros associados aos modelos, nomeadamente a normalidade, média nula, variância constante e a não correlação dos erros, verificou-se que

- Os modelos MEES_h e MEEC_h falharam no pressuposto da independência dos erros para $h = 1$ e $h = 6$ dias; no entanto, cumprem o pressuposto da normalidade.

  No que diz respeito aos restantes horizontes temporais ($h = 2, 3, 4, 5$ dias), foram efetuados os testes de normalidade Kolmogorov-Smirnov e verificou-se que a condição da normalidade dos erros também é válida em ambos os modelos para $h = 2, 3, 4$ dias. Já a condição da não correlação, recorreu-se ao teste Ljung-Box, onde se fez variar o número de autocorrelações de 7 a 17, inclusive, e chegou-se à conclusão de que nenhum dos modelos compre o referido pressuposto;

- Ambos os modelos MEES_h^* e MEEC_h^* cumprem os pressupostos da normalidade para $h = 1$ e $h = 6$ dias; o pressuposto da independência dos erros foi apenas verificada pelo modelo MEES_h^* com $h = 1$ dia.

  Em relação aos restantes horizontes temporais ($h = 2, 3, 4, 5$ dias), foram efetuados os testes de normalidade Kolmogorov-Smirnov e verificou-se que a condição da normalidade dos erros é válida para ambos os modelos. Já a condição da não correlação, recorreu-se ao teste Ljung-Box, onde se fez variar o número de autocorrelações de 7 a 17, inclusive e verificou-se que os modelos MEES_h^* e MEEC_h^* não verificaram o pressuposto da independência dos erros.

Capítulo 8

Conclusão

A motivação deste trabalho fundamentou-se na gestão de forma eficiente dos recursos hídricos nos sistemas de irrigação. Nesse sentido, foram estudadas variáveis meteorológicas que têm impacto no processo de evapotranspiração, nomeadamente as temperaturas máxima e mínima do ar, no período de 20 de fevereiro a 11 de outubro de 2019, registadas numa quinta em Carrazeda de Ansiães, situada no distrito de Bragança. Foram propostos modelos de calibração que admitem uma representação de espaço de estados na qual estão associados ao filtro de Kalman, que é um algoritmo recursivo que permite atualizar o vetor de estados, em tempo real, cada vez que uma nova observação fica disponível. Além disso, permite melhorar a qualidade das previsões obtidas, na qual se pretendia calibrar as previsões dadas pelo *website weatherstack.com* para a mesma região onde se localiza a quinta, para cada horizonte temporal de 1 até 6 dias.

Nesta dissertação foram comparados modelos de calibração com modelos de regressão linear simples que, particularmente podem ser vistos como modelos mais simplistas do que os modelos de espaço de estados, a fim de identificar qual o modelo de previsão que seria mais apropriado em termos da qualidade de ajustamento e capacidade preditiva. Para proceder com as comparações, foram utilizadas algumas medidas de avaliação, nomeadamente a REQM, EAM, EEAM, U-Theil, foram calculados os coeficientes de determinação ajustados e também foram determinados os critérios de seleção AIC e BIC.

De um modo geral, verificou-se que os MEE obtiveram melhor desempenho do que os modelos de RLS. Com a adição da componente constante no modelo de calibração na equação de observação, obteve-se melhores resultados no sentido em que possuem menores valores de REQM, EAM, EEAM, U-Theil, melhor ajustamento dado pelo r_a^2 e menores valores para os critérios de seleção AIC e BIC, tanto para a

temperatura mínima como para a temperatura máxima.

No que se refere ao cumprimento dos pressupostos dos resíduos dos modelos relativos à temperatura máxima, verificou-se que tanto nos modelos de calibração com a constante aditiva como nos modelos de calibração sem a constante aditiva, a análise dos resíduos foi idêntica, destacando-se, contudo, que a hipótese da independência dos erros foi rejeitada para a maioria dos modelos. Além disso, o pressuposto da normalidade dos resíduos também falhou para vários modelos. Já para os modelos de calibração (com e sem a constante aditiva) referentes à temperatura mínima, verificou-se que a hipótese da normalidade dos erros não foi rejeitada em nenhum dos modelos. No entanto, o pressuposto da independência dos resíduos não foi cumprido para a maioria.

Mesmo assim, pode-se concluir que, com o acréscimo da componente constante nos modelos de calibração na equação de observação, o ganho foi satisfatório.

Uma das principais dificuldades encontradas no desenvolvimento desta dissertação esteve relacionado com a qualidade das previsões provenientes do *website*, dada a existência de previsões bastante diferentes das que realmente se observou. Por outro lado, o facto de modelar séries temporais de variáveis meteorológicas de extremos (temperaturas máxima e mínima) também foi uma das dificuldades encontradas pois este tipo de séries tendem a ter um comportamento mais instável e, portanto, mais difícil de se prever. Talvez, se se modelasse séries temporais da temperatura média do ar, o desempenho dos modelos propostos teria sido melhor.

De uma forma geral, apesar dos modelos propostos não terem cumprido alguns dos pressupostos relativos aos resíduos dos modelos, verificou-se que os modelos de calibração baseados no filtro de Kalman melhoraram a qualidade das previsões obtidas.

8.1 Trabalho futuro

Dada as dificuldades encontradas no desenvolvimento desta dissertação, assim como questões que foram surgindo, ficam alguns desenvolvimentos e possíveis investigações interessantes por fazer.

Assim, como trabalho futuro propõe-se:

– Implementar e comparar os modelos propostos através da validação cruzada de séries temporais proposto por Hyndman (2014) e Hyndman e Athanasopoulos (2018). Neste trabalho, não foi possível aplicar esta abordagem aos dados uma vez que, sendo um método iterativo com a necessidade de colocar valores inici-

ais, a dificuldade consistiu na convergência em todos os passos deste método. Seria, portanto, interessante trabalhar com outras funções do *software* R que permitisse aplicar esta abordagem;

– Ajustar o modelo de calibração definido pelas equações (7.3) e (7.4), considerando a componente aditiva α estocástica e variando ao longo do tempo (processo estocástico);

– Considerar os modelos com a formulação de espaço de estados num contexto espaço-temporal. O estudo da aplicação dos MEE a séries temporais, tendo em conta a vários pontos geográficos, ou seja, observações noutros pontos no espaço poderá trazer vantagens (mais informação) no processo de modelação e, consequentemente melhorar a qualidade das previsões obtidas;

– Modelar separadamente as séries, considerando a estação fria e quente. Esta divisão pode trazer vantagens tanto na convergência do método de estimação dos parâmetros como na qualidade das previsões, visto que o comportamento das séries, sendo de natureza meteorológica, tendem a ter comportamentos diferentes entre estações do ano devido à existência de sazonalidade. Parte desta análise foi efetuada para a temperatura máxima, mas não foi possível concluí-la. As estimativas dos parâmetros dos modelos, assim como os valores das medidas de avaliação encontram-se nas Tabelas B.1, B.2 e B.3, no apêndice B.

Bibliografia

[1] Achar, A., Bharathi, D., Kumar, B. A. e Vanajakshi, L. (2020). Bus Arrival Time Prediction: A Spatial Kalman Filter Approach. *IEEE Transactions on Intelligent Transportation Systems*, 21(3): 1298-1307.

[2] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC- 19(6): 716-723.

[3] Álvarez, B., Padilla, L., Mateu, J. e Ferreira, G. (2019). A Kalman Filter Method for Estimation and Prediction of Space–Time Data with an Autoregressive Structure. *Journal of Statistical Planning and Inference*, 203: 117-130.

[4] ANP/WWF (2019). Relatório da Associação Natureza Portugal em Associação com a World Wide Fund for Nature: Vulnerabilidade de Portugal à Seca e Escassez de Água.

[5] Arlot, S. e Celisse, A. (2010). A Survey of Cross Validation Procedures for Model Selection. *Statistics Surveys, Institute of Mathematical Statistics (IMS)*, 4: 40-79.

[6] Baturin, O. (2016). Modelos Estruturais na Análise de Séries Temporais de Dados Ambientais. Tese de Mestrado, Escola de Ciências da Universidade do Minho.

[7] Brockwell, P. J. e Davis, R. A. (1998). *Introduction to Time Series and Forecasting*. Springer.

[8] Bergmeir, C. e Benítez, J. M. (2012). On the Use of Cross-Validation for Time Series Predictor Evaluation. *Information Sciences*, 191: 192-213.

[9] Bergmeir, C., Costantini. M e Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, 76: 132-143.

[10] Bergmeir, C., Hyndman, R. J. e Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120: 70-83.

[11] Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. Biometrika, 76(3), 503-514.

[12] Burman, P., Chow, E. e Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 84: 351-358.

[13] Cerqueira, V., Torgo, L. e Mozetic, I. (2019). *Evaluating time series forecasting models: An empirical study on performance estimation methods*. arXiv: 1905.11744.

[14] Chai, T. e Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3): 1247-1250.

[15] Commandeur, J. J. F. e Koopman, S. J. (2007). *An Introduction to State Space Time Series Analysis*. Oxford University Press.

[16] Cordeiro, C. (2003). *Modelos de Previsão em Séries Temporais. Aplicação da Metodologia Bootstrap*. Tese de Mestrado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[17] Costa, C. (2019). *Desenvolvimento de Modelos de Previsão de Variáveis Climáticas*. Tese de Mestrado, Escola de Ciências da Universidade do Minho.

[18] Costa, M. (2006). *Estimação dos Parâmetros de Modelos em Espaço de Estados. Uma aplicação à calibração do radar meteorológico*. Tese de Doutoramento, Faculdade de Ciências da Universidade de Lisboa.

[19] Costa, M. e Alpuim, T. (2010). Parameter Estimation of State Space Models for Univariate Observations. *Journal of Statistical Planning and Inference*, 140(7): 1889-1902.

[20] Costa, M. e Alpuim, T. (2011). Adjustment of State Space Models in View of Area Rainfall Estimation. *Environmetrics*, 22(4): 530-540.

[21] Costa, M., Gonçalves, A. M. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*, 25: 151–163.

[22] Dabrowski, J. J., Rahman, A., George, A., Arnold, S. e McCulloch, J. (2018). State Space Models for Forecasting Water Quality Variables: An Application in Aquaculture Prawn Farming. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA: 177–185.

[23] Dempster, A. P., Laird, N. M. e Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1): 1–38.

[24] Durbin, J. e Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

[25] Gonçalves, A. M., Baturin, O. e Costa, M. (2018). Time Series Analysis by State Space Models Applied to a Water Quality Data in Portugal. *AIP Conference Proceedings* Vol. 1978: 470101-1 - 470101-4.

[26] Gonçalves, A. M. e Costa, M. (2013). Predicting Seasonal and Hydro-meteorological Impact in Environmental Variables Modelling via Kalman Filtering. *Stochastic Environmental Research and Risk Assessment*, 27(5): 1021-1038.

[27] Harvey, A. C. (1992).*Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

[28] Harvey, A. C., Koopman, S. J. e Shephard, N. (2004). State Space and Unobserved Component Models: Theory and Applications. Cambridge University Press.

[29] Hyndman, R. J. (2014). Measuring forecast accuracy. *Citeseer*.

[30] Hyndman, R. J. e Athanasopoulos, G. (2018). *Forecasting: principles and practice*, 2.ª edição. OTexts: Melbourne, Australia. `https://otexts.com/fpp2/`.

[31] Hyndman, R. J. e Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International journal of forecasting, 22(4)*: 679-688.

[32] Inoue, A. e Kilian, L. (2006). On the selection of forecasting models.*Journal of Econometrics*, 130(2): 273-306

[33] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME. *Journal of Basic Engineering*, 82(Series D): 35–45.

[34] Kalman, R. E. e Bucy, R. S. (1961). New Results in Linear Filtering and Prediction Theory. Transactions of the ASME. *Journal of Basic Engineering*, 83(Series D) : 95-108.

[35] Linroth, C. (2014). Statistical analysis of wave heights using Kalman Filtering methods. Relatório de Projeto, Universidade de Uppsala.

[36] Lopes, S. O., Fontes, F., Pereira, R., Pinho, M. e Gonçalves, A. M. (2016). Optimal Control Applied to an Irrigation Planning Problem. *Mathematical Problems in Engineering*, 2016: 1-10.

[37] Makridakis, S., Wheelwright, S. C., e Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. John wiley and Sons, 3.ª edição.

[38] Moreira, R. (2019). Principais métodos de rega utilizados na agricultura. `https://acientistaagricola.pt/principais-metodos-de-rega/`. 27/08/2020.

[39] Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M. e Stork, J. (2015). Comparison of Different Methods for Univariate Time Series Imputation in R. *arXiv preprint arXiv:1510.03924*.

[40] Murteira, B., Muller, D. e Turkman, K. F. (1993). *Análise de Sucessões Cronológicas*. McGraw-Hill.

[41] Murteira, B., Ribeiro, C. S., Silva, J. A., Pimenta, C. e Pimenta, F. (2015). *Introdução à Estatística*. Escolar editora, 3.ª edição.

[42] National Geographic (2020). Freshwater Crisis. `https://www.nationalgeographic.com/environment/freshwater/freshwater-crisis/`. 29/08/20.

[43] Petris, G., Petrone, S. e Campagnoli, P. (2009). *Dynamic Linear Models with* R. Springer.

[44] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[45] Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1): 39-61.

[46] Racine, J. (1997). Feasible cross-validatory model selection for general stationary processes. *Journal of Applied Econometrics*, 12: 169-179.

[47] Reis, Elizabeth (2008). *Estatística Descritiva*. Sílabo, 7.ª edição.

[48] Samarin, M. K. (2012). Linear Interpolation, Encyclopedia of Mathematics. `https://encyclopediaofmath.org/wiki/Linear_interpolation`. 26/08/20.

[49] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6: 461-464.

[50] Shumway, R. H. e Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer, 4th edition.

[51] Shumway, R. H. e Stoffer, D. S. (2019). *Time Series: A Data Analysis Approach using R*. CRC Press.

[52] Vicente, M. (1997). *Métodos de Determinação do Azimute por Observações Astronómicas*. Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

[53] WWAP (2019). Relatório Mundial das Nações Unidas sobre o Desenvolvimento dos Recursos Hídricos 2019: Não Deixar Ninguém para Trás, fatos e dados. UNESCO.

[54] WWAP (2020). Relatório Mundial das Nações Unidas sobre o Desenvolvimento dos Recursos Hídricos 2020: Água e Mudança Climática, resumo executivo. UNESCO.

Apêndice A

Tabelas

Tabela A.1: Teste da Normalidade Kolmogorov-Smirnov para a série da temperatura máxima, Y_t^M e das respetivas previsões do *website weatherstack*, $W_{t,(h)}^M$, $h = 1, \ldots, 6$ dias.

	Y_t^M	$W_{t,(1)}^M$	$W_{t,(2)}^M$	$W_{t,(3)}^M$	$W_{t,(4)}^M$	$W_{t,(5)}^M$	$W_{t,(6)}^M$
estatística de teste	0,0541	0,0608	0,0746	0,0672	0,0869	0,0647	0,0877
valor de prova	0,5274	0,3526	0,1474	0,2417	0,0586	0,2817	0,0547

Tabela A.2: Teste da Normalidade Kolmogorov-Smirnov para a série da temperatura mínima, Y_t^m e das respetivas previsões do *website weatherstack*, $W_{t,(h)}^m$, $h = 1, \ldots, 6$ dias.

	Y_t^m	$W_{t,(1)}^m$	$W_{t,(2)}^m$	$W_{t,(3)}^m$	$W_{t,(4)}^m$	$W_{t,(5)}^m$	$W_{t,(6)}^m$
estatística de teste	0,0882	0,0760	0,0726	0,0627	0,0767	0,0728	0,0798
valor de prova	0,0612	0,1338	0,1697	0,3162	0,1274	0,1673	0,1016

Tabela A.3: *Outliers* do rácio, $Y_t^M/W_{t,(h)}^M$, temperatura máxima observada, Y_t^M, previsões do *website weatherstack*, $W_{t,(h)}^M$ e previsões do *website* com a substituição dos *outliers* do rácio por interpolação linear, $W_{t,(h)}^{M*}$, $h = 1, \ldots, 6$ dias.

horizonte temporal	**data**	***outliers* de $Y_t^M/W_{t,(h)}^M$**	Y_t^M	$W_{t,(h)}^M$	$W_{t,(h)}^{M*}$
h=1	04/03/2019	1,3701	17,4	12,7	15,7
	06/03/2019	1,6436	16,6	10,1	12,8
	03/04/2019	1,4508	17,7	12,2	15,4
	05/04/2019	1,4851	15,0	10,1	10,8
	06/04/2019	1,4375	13,8	9,6	10,0
	09/05/2019	1,4867	22,3	15,0	20,1
	10/05/2019	1,6732	25,6	15,3	22,2
	16/05/2019	0,9454	22,5	23,8	23,9
	13/07/2019	1,5388	33,7	21,9	34,9

Continua na próxima página

Tabela A.3 – continuação da página anterior

horizonte temporal	data	*outliers* de $Y_t^M/W_{t,(h)}^M$	Y_t^M	$W_{t,(h)}^M$	$W_{t,(h)}^{M*}$
	17/07/2019	1,8551	38,4	20,7	33,1
	25/08/2019	0,8589	27,4	31,9	32,2
	16/09/2019	0,9441	27,0	28,6	29,4
	04/03/2019	1,4622	17,4	11,9	15,2
	06/03/2019	1,6768	16,6	9,9	11,9
	03/04/2019	1,3828	17,7	12,8	15,6
	05/04/2019	1,5306	15,0	9,8	10,8
	06/04/2019	1,3939	13,8	9,9	9,9
	25/04/2019	1,4216	14,5	10,2	14,3
h=2	09/05/2019	1,4867	22,3	15,0	19,5
	10/05/2019	1,6732	25,6	15,3	21,7
	23/06/2019	1,5529	26,4	17,0	29,2
	24/06/2019	0,8084	21,1	26,1	27,9
	14/07/2019	1,7136	37,7	22,0	32,0
	18/07/2019	1,8454	38,2	20,7	34,8
	27/02/2019	1,4774	22,9	15,5	18,3
	04/03/2019	1,5398	17,4	11,3	15,8
	05/03/2019	1,9770	17,2	8,7	14,2
	06/03/2019	1,8652	16,6	8,9	12,6
	09/05/2019	1,6519	22,3	13,5	18,4
h=3	10/05/2019	1,6732	25,6	15,3	21,1
	16/05/2019	0,8929	22,5	25,2	21,9
	24/06/2019	0,8474	21,1	24,9	21,6
	15/07/2019	1,8155	37,4	20,6	33,2
	19/07/2019	2,2629	39,6	17,5	34,2
	25/08/2019	0,8457	27,4	32,4	32,9
	04/03/2019	1,5398	17,4	11,3	16,1
	05/03/2019	1,7551	17,2	9,8	14,4
	06/03/2019	1,6275	16,6	10,2	12,6
	14/03/2019	1,4690	21,3	14,5	15,8
	17/03/2019	1,4773	19,5	13,2	16,9
	19/03/2019	1,4507	20,6	14,2	14,2
	05/04/2019	1,4423	15,0	10,4	12,4
h=4	15/04/2019	0,8690	19 9	22,9	19,6
	17/04/2019	0,8128	19,1	23,5	17,9
	16/05/2019	0,8824	22,5	25,5	24,5
	17/05/2019	1,5581	20,1	12,9	19,9
	24/06/2019	0,7563	21,1	27,9	27,1
	16/07/2019	2,0109	37,0	18,4	33,9
	20/07/2019	2,0508	36,3	17,7	35,7

Continua na próxima página

Tabela A.3 – continuação da página anterior

horizonte temporal	data	*outliers* de $Y_t^M/W_{t,(h)}^M$	Y_t^M	$W_{t,(h)}^M$	$W_{t,(h)}^{M*}$
	25/08/2019	0,8354	27,4	32,8	31,8
h=5	06/03/2019	1,5370	16,6	10,8	12,6
	12/03/2019	1,7583	21,1	12,0	15,2
	17/03/2019	1,5600	19,5	12,5	15,7
	05/04/2019	1,5789	15,0	9,5	12,9
	17/07/2019	1,8916	38,4	20,3	32,9
	21/07/2019	2,3735	39,4	16,6	36,5
	25/08/2019	0,7405	27,4	37,0	34,7
h=6	27/02/2019	2,3608	22,9	9,7	18,2
	12/03/2019	2,0095	21,1	10,5	16,1
	31/03/2019	1,6497	25,9	15,7	20,3
	14/04/2019	1,5921	24,2	15,2	18,6
	18/04/2019	1,6111	20,3	12,6	18,2
	24/05/2019	1,7640	28,4	16,1	26,4
	18/07/2019	1,8019	38,2	21,2	33,1
	22/07/2019	2,1156	42,1	19,9	36,4

Tabela A.4: *Outliers* do rácio, $Y_t^m/W_{t,(h)}^m$, temperatura mínima observada, Y_t^m, previsões do *website weatherstack*, $W_{t,(h)}^m$ e previsões do *website* com a substituição dos *outliers* do rácio por interpolação linear, $W_{t,(h)}^{m*}$, $h = 1, \ldots, 6$ dias.

horizonte temporal	data	*outliers* de $Y_t^m/W_{t,(h)}^m$	Y_t^m	$W_{t,(h)}^m$	$W_{t,(h)}^{m*}$
h=1	26/02/2019	0,5362	3,7	6,9	5,8
	08/03/2019	1,8125	5,8	3,2	3,9
	13/03/2019	1,8387	5,7	3,1	4,5
	18/03/2019	1,8621	5,4	2,9	4,2
	20/03/2019	1,7353	5,9	3,4	3,5
	21/03/2019	4,2500	6,8	1,6	3,6
	26/03/2019	1,7872	8,4	4,7	6,3
	27/03/2019	2,3235	7,9	3,4	5,8
	28/03/2019	2,1622	8,0	3,7	5,3
	03/04/2019	1,8222	8,2	4,5	6,5
	04/04/2019	3,6000	3,6	1,0	5,4
	07/04/2019	-6,3333	5,7	-0,9	5,4
	26/04/2019	1,8113	9,6	5,3	6,2
	10/05/2019	1,9452	14,2	7,3	9,0
	20/02/2019	0,5088	2,9	5,7	4,7
	26/02/2019	0,5362	3,7	6,9	5,8
	08/03/2019	1,9333	5,8	3,0	4,0

Continua na próxima página

Tabela A.4 – continuação da página anterior

horizonte temporal	data	*outliers* de $Y_t^m/W_{t,(h)}^m$	Y_t^m	$W_{t,(h)}^m$	$W_{t,(h)}^{m*}$
	13/03/2019	1,9655	5,7	2,9	4,4
	16/03/2019	0,6071	3,4	5,6	4,0
	21/03/2019	4,0000	6,8	1,7	3,7
	26/03/2019	1,7143	8,4	4,9	6,1
	27/03/2019	2,2571	7,9	3,5	5,7
	28/03/2019	2,2222	8,0	3,6	5,2
h=2	03/04/2019	1,8636	8,2	4,4	7,1
	04/04/2019	6,0000	3,6	0,6	6,2
	05/04/2019	2,6800	6,7	2,5	5,2
	07/04/2019	57,0000	5,7	0,1	5,1
	08/04/2019	1,6833	10,1	6,0	6,0
	26/04/2019	2,2857	9,6	4,2	6,2
	10/05/2019	1,9452	14,2	7,3	8,4
	08/07/2019	1,8085	17,0	9,4	15,9
	20/02/2019	0,4603	2,9	6,3	6,3
	05/03/2019	2,6296	7,1	2,7	6,3
	07/03/2019	2,2222	4,0	1,8	4,4
	13/03/2019	2,0357	5,7	2,8	4,7
	18/03/2019	2,2500	5,4	2,4	4,2
	20/03/2019	2,8095	5,9	2,1	2,8
	21/03/2019	3,5789	6,8	1,9	3,2
	26/03/2019	1,8667	8,4	4,5	6,0
	27/03/2019	2,1944	7,9	3,6	5,6
	28/03/2019	2,1053	8,0	3,8	5,2
h=3	04/04/2019	2,1176	3,6	1,7	4,4
	05/04/2019	2,9130	6,7	2,3	3,9
	07/04/2019	2,7143	5,7	2,1	4,1
	08/04/2019	2,2444	10,1	4,5	4,8
	10/04/2019	1,8889	8,5	4,5	4,8
	19/04/2019	2,0893	11,7	5,6	8,5
	24/04/2019	4,0526	7,7	1,9	5,8
	26/04/2019	2,1333	9,6	4,5	5,8
	10/04/2019	1,9452	14,2	7,3	7,1
	23/06/2019	1,8916	15,7	8,3	15,2
	19/07/2019	1,9159	20,5	10,7	19,1
	20/02/2019	0,4754	2,9	6,1	5,5
	05/03/2019	8,8750	7,1	0,8	5,7
	14/03/2019	2,4615	3,2	1,3	3,5
	18/03/2019	1,9286	5,4	2,8	5,6
	19/03/2019	2,3125	3,7	1,6	5,2
	20/03/2019	3,9333	5,9	1,5	4,7

Continua na próxima página

Tabela A.4 – continuação da página anterior

horizonte temporal	data	*outliers* de $Y_t^m/W_{t,(h)}^m$	Y_t^m	$W_{t,(h)}^m$	$W_{t,(h)}^{m*}$
	21/03/2019	3,0909	6,8	2,2	4,3
	27/03/2019	1,8810	7,9	4,2	5,5
h=4	28/03/2019	2,1622	8,0	3,7	5,1
	03/04/2019	1,7826	8,2	4,6	7,6
	04/04/2019	2,5714	3,6	1,4	6,2
	05/04/2019	2,9130	6,7	2,3	4,8
	23/04/2019	1,7872	8,4	4,7	7,6
	24/04/2019	3,2083	7,7	2,4	6,7
	26/04/2019	2,3415	9,6	4,1	6,0
	16/07/2019	1,9636	21,6	11,0	20,2
	20/07/2019	2,2174	20,4	9,2	19,0
	21/02/2019	0,5068	3,7	7,3	4,9
	25/02/2019	0,5625	3,6	6,4	6,3
	26/02/2019	0,5781	3,7	6,4	6,1
	07/03/2019	1,7391	4,0	2,3	4,8
	08/03/2019	2,7619	5,8	2,1	4,9
	13/03/2019	2,3750	5,7	2,4	4,7
	14/03/2019	1,8824	3,2	1,7	3,7
	16/03/2019	0,5152	3,4	6,6	4,1
	18/03/2019	4,1538	5,4	1,3	5,2
	19/03/2019	2,4667	3,7	1,5	4,9
	20/03/2019	59,0000	5,9	0,1	4,7
	21/03/2019	5,2308	6,8	1,3	4,4
h=5	28/03/2019	1,8605	8,0	4,3	7,4
	04/04/2019	3,0000	3,6	1,2	7,1
	05/04/2019	2,6800	6,7	2,5	6,9
	06/04/2019	1,8065	5,6	3,1	6,6
	07/04/2019	2,3750	5,7	2,4	6,4
	13/04/2019	2,0000	10,0	5,0	8,4
	20/04/2019	2,2400	16,8	7,5	8,0
	24/04/2019	1,9744	7,7	3,9	7,3
	26/04/2019	2,3415	9,6	4,1	6,2
	29/05/2019	1,7927	14,7	8,2	11,7
	17/07/2019	1,7252	22,6	13,1	18,8
	21/07/2019	2,8406	19,6	6,9	20,7
	21/02/2019	0,4353	3,7	8,5	6,0
	24/02/2019	-0,8200	4,1	-5,0	6,8
	26/02/2019	0,5522	3,7	6,7	5,7
	13/03/2019	5,1818	5,7	1,1	3,9
	14/03/2019	1,8824	3,2	1,7	3,7
	18/03/2019	7,7143	5,4	0,7	6,3

Continua na próxima página

Tabela A.4 – continuação da página anterior

horizonte temporal	**data**	***outliers* de $Y_t^m/W_{t,(h)}^m$**	Y_t^m	$W_{t,(h)}^m$	$W_{t,(h)}^{m*}$
	19/03/2019	37,0000	3,7	0,1	5,9
	20/03/2019	8,4286	5,9	0,7	5,6
	21/03/2019	34,0000	6,8	0,2	5,2
h=6	05/04/2019	3,7222	6,7	1,8	4,8
	06/04/2019	-5,0909	5,6	-1,1	5,9
	07/04/2019	2,3750	5,7	2,4	7,1
	14/04/2019	2,1538	14,0	6,5	7,6
	18/04/2019	2,2222	10,0	4,5	9,3
	20/04/2019	1,8065	16,8	9,3	10,3
	24/04/2019	2,9615	7,7	2,6	8,0
	26/04/2019	2,5946	9,6	3,7	8,2
	11/06/2019	1,8250	14,6	8,0	11,2
	22/07/2019	2,2100	22,1	10,0	20,5

Apêndice B

Divisão da Série

Tabela B.1: Estimativas e erros padrão do modelo de regressão linear simples e do modelo de calibração sem constante aditiva para a série da temperatura máxima dividida em duas subséries: primeira subsérie - 20 fevereiro até 20 abril de 2019; segunda subsérie - 21 abril até 11 outubro de 2019.

		1.ª série - 20 fevereiro até 20 abril de 2019						2.ª série - 21 abril até 11 outubro de 2019					
		h=1	h=2	h=3	h=4	h=5	h=6	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \beta W_{t,(h)}^M + e_t$													
β	estimativa	1,2081	1,2203	1,2075	1,1940	1,2105	1,2463	1,1296	1,1229	1,1247	1,1176	1,1060	1,0944
	erro padrão	0,0113	0,0119	0,0169	0,0193	0,0192	0,0261	0,0061	0,0067	0,0083	0,0084	0,0097	0,0109
σ_e	estimativa	1,5197	1,5865	2,2650	2,6114	2,5669	3,3779	2,2184	2,4714	3,0369	3,0825	3,5938	4,0885
	erro padrão	0,1951	0,2034	0,2899	0,3344	0,3298	0,4329	0,1675	0,1866	0,2291	0,2328	0,2716	0,3087
	logL	-109,7445	-112,3230	-133,6869	-142,2261	-141,1936	-157,6669	-385,0359	-403,8241	-439,6791	-442,2733	-468,9758	-476,2951
$Y_t^M = \beta_t W_{t,(h)}^M + e_t,$ $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$													
ϕ	estimativa	0,8277	0,7991	0,7373	0,3764	0,7263		0,9726	0,9375	0,9125	0,9472	0,7606	0,6644
	erro padrão	0,1111	0,1241	0,1824	0,2266	0,3618		0,0227	0,0493	0,0919	0,0447	0,2638	0,2564
μ	estimativa	1,2168	1,2323	1,2198	1,2194	1,2166		1,1232	1,1279	1,1261	1,1117	1,1075	1,0979
	erro padrão	0,0271	0,0258	0,0261	0,0286	0,0240		0,0226	0,0146	0,0120	0,0187	0,0112	0,0134
σ_ε	estimativa	0,0329	0,0363	0,0406	0,1068	0,0258		0,0049	0,0101	0,0088	0,0059	0,0169	0,0349
	erro padrão	0,0106	0,0134	0,0235	0,0376	0,0375		0,0049	0,0043	0,0063	0,0133	0,0180	0,0214
σ_e	estimativa	1,1580	1,2049	1,9834	1,6456	2,4597		2,0764	2,3492	2,9767	2,9885	3,5083	3,8605
	erro padrão	0,1427	0,1630	0,2527	0,6321	0,2942		0,1206	0,1340	0,1700	0,1957	0,2236	0,2811
	logL	-50,6820	-53,2407	-77,5831	-85,1673	-85,9537		-219,6993	-241,9221	-279,697	-280,4175	-308,8848	-331,1200

Nota: O modelo MEES_6 não convergiu.

Tabela B.2: Estimativas e erros padrão do modelo de regressão linear simples e do modelo de calibração com constante aditiva para a série da temperatura máxima dividida em duas subséries: primeira subsérie - 20 fevereiro até 20 abril de 2019; segunda subsérie - 21 abril até 11 outubro de 2019.

		1.ª série - 20 fevereiro até 20 abril de 2019						2.ª série - 21 abril até 11 outubro de 2019					
		h=1	h=2	h=3	h=4	h=5	h=6	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \alpha + \beta W_{t,(h)}^M + e_t$													
α	estimativa	4,5042	4,9453	6,9776	7,8829	7,1521	10,3856	5,2945	5,7778	7,5476	7,2407	8,0799	10,0968
	erro padrão	0,8617	0,8415	1,1039	1,2954	1,5254	1,7580	0,7753	0,8691	1,0136	1,0796	1,3348	1,4385
β	estimativa	0,9538	0,9385	0,8138	0,7534	0,8038	0,6362	0,9423	0,9196	0,8585	0,8636	0,8246	0,7461
	erro padrão	0,0495	0,0489	0,0636	0,0740	0,0883	0,1053	0,0280	0,0312	0,0365	0,0386	0,0473	0,0506
σ_e	estimativa	1,2637	1,2668	1,7579	2,0576	2,2046	2,6919	1,9733	2,2107	2,6486	2,7524	3,2725	3,6152
	erro padrão	0,1618	0,1622	0,2250	0,2634	0,2822	0,3446	0,1492	0,1671	0,2002	0,2081	0,2474	0,2733
	logL	-98,1634	-98,3088	-117,9660	-127,4124	-131,5523	-143,5349	-364,1629	-383,9278	-415,3684	-422,0609	-452,1733	-469,5054
$Y_t^M = \alpha + \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$													
ϕ	estimativa	0,7645	0,7370	0,6560	0,1852	0,4380	0,6386	0,7900	0,7593	0,7720	0,8243	0,8293	0,8068
	erro padrão	0,1615	0,1575	0,2218	0,1773	0,3599	0,1774	0,1382	0,2401	0,2059	0,1002	0,0961	0,0893
α	estimativa	4,1401	4,7056	6,8579	7,2871	7,5738	12,7761	5,5794	5,9267	8,1616	8,6828	10,3791	14,5440
	erro padrão	0,8450	0,8014	1,0701	1,1490	1,7653	2,4290	0,8475	0,9194	1,4573	1,3016	1,7675	1,9703
μ	estimativa	0,9785	0,9571	0,8246	0,7892	0,7800	0,4905	0,9323	0,9147	0,8376	0,8122	0,7438	0,5868
	erro padrão	0,0512	0,0496	0,0649	0,0710	0,1061	0,1536	0,0312	0,0331	0,0524	0,0480	0,0644	0,0742
σ_ε	estimativa	0,0229	0,0259	0,0341	0,0991	0,0720	0,0930	0,0135	0,0091	0,0107	0,0218	0,0266	0,0488
	erro padrão	0,0101	0,0114	0,0223	0,0229	0,0455	0,0340	0,0069	0,0094	0,0123	0,0093	0,0131	0,0191
σ_e	estimativa	1,0565	1,0110	1,5135	0,9358	1,6970	1,9544	1,8653	2,1633	2,5837	2,5147	2,9598	2,8750
	erro padrão	0,1219	0,1286	0,2167	0,5830	0,5518	0,4807	0,1177	0,1377	0,1734	0,1680	0,2203	0,3274
	logL	-40,8815	-39,7844	-61,6455	-69,5453	-75,9899	-87,9603	-203,5249	-224,1219	-255,5233	-259,6524	-289,4737	-302,2500

Tabela B.3: Medidas de avaliação dos modelos referentes às subséries da temperatura máxima.

	1.ª série - 20 fevereiro até 20 abril de 2019						2.ª série - 21 abril até 11 outubro de 2019					
	h=1	h=2	h=3	h=4	h=5	h=6	h=1	h=2	h=3	h=4	h=5	h=6
$Y_t^M = \beta W_{t,(h)}^M + e_t$												
REQM	1,5070	1,5732	2,2461	2,5896	2,5454	3,3496	2,2120	2,4643	3,0282	3,0736	3,5834	4,0768
EAM	1,1645	1,2188	1,6710	1,9312	2,0534	2,4646	1,3895	1,5408	1,9554	2,0576	2,4550	3,0042
EEAM	0,4774	0,4997	0,6851	0,7918	0,8419	1,0105	0,5424	0,6014	0,7633	0,8032	0,9583	1,1727
U-Theil	0,5277	0,5057	0,7726	0,8006	0,7547	0,7036	0,5537	0,5979	0,6500	0,6045	0,6555	0,7522
r_a^2	0,8647	0,8640	0,7382	0,6413	0,5882	0,3861	0,8685	0,8350	0,7632	0,7442	0,6385	0,5588
AIC	223,4890	228,6460	271,3738	288,4522	286,3872	319,3338	774,0718	811,6482	883,3582	888,5466	941,9516	956,5902
$Y_t^M = \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$												
REQM	1,3652	1,4509	2,1766	2,5813	2,5364		2,1302	2,4107	3,0130	3,0189	3,5675	4,0340
EAM	1,0795	1,1053	1,6250	1,8957	2,0170		1,2880	1,4663	1,9380	2,0001	2,4419	2,9778
EEAM	0,4426	0,4532	0,6663	0,7772	0,8270		0,5028	0,5724	0,7565	0,7807	0,9532	1,1623
U-Theil	0,4523	0,4346	0,7097	0,7240	0,7358		0,5120	0,5671	0,6367	0,5796	0,6411	0,7164
r_a^2	0,8694	0,8670	0,7395	0,6316	0,5864		0,8706	0,8354	0,7622	0,7415	0,6412	0,5670
AIC	109,3640	114,4814	163,1662	178,3346	179,9074		447,3986	491,8442	567,3940	568,8350	625,7696	670,2400
$Y_t^M = \alpha + \beta W_{t,(h)}^M + e_t$												
REQM	1,2425	1,2455	1,7283	2,0230	2,1676	2,6467	1,9620	2.1980	2,6333	2,7366	3,2536	3,5944
EAM	0,9830	1,0130	1,4102	1,5741	1,7117	2,0778	1,2454	1,3927	1,7393	1,8922	2,3770	2,7709
EEAM	0,4030	0,4153	0,5782	0,6454	0,7018	0,8519	0,4861	0,5436	0,6789	0,7386	0,9278	1,0816
U-Theil	0,5203	0,4802	0,7779	0,9519	0,9915	1,1771	0,5696	0,6455	0,7255	0,7097	0,8777	1,0470
r_a^2	0,8624	0,8617	0,7337	0,6351	0,5811	0,3755	0,8677	0,8340	0,7618	0,7427	0,6364	0,5562
AIC	202,3268	202,6176	241,9320	260,8248	269,1046	293,0698	734,3258	773,8556	836,7368	850,1218	910,3466	945,0108
$Y_t^M = \alpha + \beta_t W_{t,(h)}^M + e_t$, $\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t$												
REQM	1,1874	1,1798	1,6959	2,0274	2,1579	2,5903	1,9283	2,1867	2,6161	2,6662	3,1650	3,3746
EAM	0,9578	0,9368	1,3821	1,5230	1,6772	2,1481	1,2189	1,3818	1,7374	1,8857	2,3350	2,6275
EEAM	0,3927	0,3841	0,5667	0,6244	0,6877	0,8807	0,4758	0,5394	0,6782	0,7361	0,9114	1,0256
U-Theil	0,4934	0,4507	0,7414	0,8711	0,9665	1,2283	0,5542	0,6409	0,7330	0,7132	0,8963	1,0757
r_a^2	0,8765	0,8781	0,7480	0,6424	0,5919	0,4120	0,8730	0,8367	0,7663	0,7574	0,6581	0,6115
AIC	91,7630	89,5688	133,2910	149,0906	161,9798	185,9206	417,0498	458,2438	521,0466	529,3048	588,9474	614,5000

Nota: O modelo MEES_6 não convergiu.