Universidade do Minho
Escola de Ciências

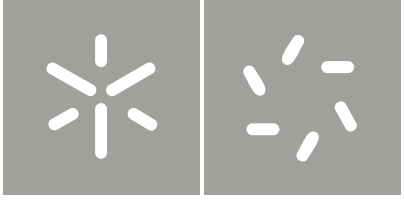
Luís Filipe Ferreira Pires

Modelos Preditivos: Quantos sinistros vão ocorrer?

outubro de 2022

Luís Filipe Ferreira Pires Modelos Preditivos: Quantos sinistros vão ocorrer?

UMinho | 2022



Universidade do Minho
Escola de Ciências

Luís Filipe Ferreira Pires

Modelos Preditivos: Quantos sinistros vão ocorrer?

Relatório de Estágio em Estatística

Trabalho efetuado sob a orientação de:
Professora Doutora Arminda Manuela Andrade Pereira Gonçalves
e de
Dr. Luís Filipe Fonseca da Cunha Ferreira

outubro de 2022

Despacho RT - 31 /2019 - Anexo 3

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Atribuição

CC BY

https://creativecommons.org/licenses/by/4.0/

Agradecimentos

"No man becomes rich without enriching others."
(Andrew Carnegie, s.d.)

Obrigado à Universidade do Minho e a todos os docentes que me forneceram o conhecimento necessário para finalizar esta etapa.

Um especial agradecimento à Professora Doutora Arminda Manuela Gonçalves, não só pela sua constante disponibilidade, mas pelas suas palavras, conselhos e incentivos que me motivaram em grande escala. O meu profundo obrigado por me ter acompanhado neste percurso.

Obrigado ao Dr. Luís Filipe Ferreira e ao Dr. Luís Filipe Maranhão, tanto pelo acompanhamento ao longo desta viagem, como pela disponibilidade e pelo ambiente familiar em que fui recebido.

Um caloroso e amoroso obrigado a toda a minha família que sempre me apoiou e fez todos os possíveis e impossíveis para me ajudar. Obrigado pelo conforto que me proporcionaram sempre que necessitava. Um especial agradecimento ao meu pai e à minha mãe que sempre me proporcionaram a educação e os meios necessários para me tornar na pessoa que hoje sou.

Obrigado aos meus colegas de Mestrado, por me terem ajudado ao longo destes dois anos.

E, finalmente, obrigado aos meus amigos, que de perto ou de longe, nunca me deixaram sentir sozinho.

Despacho RT - 31 /2019 - Anexo 4

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

Modelos Preditivos: Quantos sinistros vão ocorrer?

O mercado das seguradoras está sujeito a vários fatores que influenciam os números de sinistros registados que, consequentemente, têm um enorme peso na gestão de decisões de qualquer empresa. Logo, é de extremo interesse elaborar um estudo temporal para a extrapolação dos números de sinistros que irão ser registados futuramente. Neste âmbito, os modelos de previsão de sinistros têm um grande interesse na gestão e no processo de tomada de decisão das empresas seguradoras.

Nesta dissertação são estudadas séries temporais diárias e mensais relativas aos números de sinistros habitação registados nas categorias: danos por água, riscos elétricos, tempestades, restantes causas e total (junção de todas as categorias). Estes dados foram fornecidos pela empresa de seguros Ageas e apresentam registos desde 1985 até 2021. Assim, são estudadas a série **total** (correspondente ao total de sinistros registados) e as séries marginais **DNA** (correspondente aos registos de sinistros de danos por água), **REL** (correspondente aos registos de sinistros de riscos elétricos), **TMP** (correspondente aos registos de sinistros de tempestades e inundações) e **restantes causas** (correspondente aos registos de sinistros que não se encaixam em nenhuma categoria anterior). Tal como em vários estudos temporais, estas séries apresentam fenómenos sazonais e inúmeros outliers. Por conseguinte, são ajustados modelos SARIMA, TBATS e Holt-Winters aos dados diários e modelos SARIMA e Holt-Winters aos mensais.

O principal objetivo deste trabalho consiste em modelar os números destes sinistros, em prol de uma boa capacidade preditiva, de forma a implementar estes modelos na prática. Desta forma, é efetuado um estudo comparativo entre as três metodologias.

Para avaliar a capacidade preditiva dos modelos, são utilizadas várias medidas de avaliação, nomeadamente EQM, REQM, EPAM, EEAM e a estatística de U de Theil.

Os resultados do estudo comparativo indicam que não há grandes diferenças entre as metodologias, pois, no caso dos dados diários, todos os modelos apresentam bons ajustamentos, mas as suas previsões não são tão precisas. Em contraste, os modelos das séries mensais apresentam melhores previsões. Além disso, a modelação dos números de sinistros é melhor do que a das taxas de frequência.

Palavras-chave: Holt-Winters; SARIMA; Séries temporais; Sinistro; TBATS.

Abstract

Forecasting Models: How many claims will occur?

The insurances' market is subjected to several factors that influence the number of claims recorded which, consequently, have a huge weight in the management of any company's decisions. Therefore, there is a big interest in preparing a time series study to extrapolate the numbers of claims yet to be recorded. In this context, claims forecasting models are of great interest in the management and decision-making process of insurance companies.

In this dissertation daily and monthly time series are studied regarding the number of claims recorded in the categories: water damage, electrical hazards, storms, other and total (grouping of all categories). This data was provided by the insurance company Ageas and shows records from 1985 through 2021. Thus, the studied series are **total** (regarding the total of registered claims) and the marginal series **DNA** (regarding claims to do with water damage), **REL** (regarding claims to do with electrical hazards), **TMP** (regarding claims to do with storms and floods) and **restantes causas** (regarding claims that do not fit in any previous category).As in several time series studies, these series show the presence of seasonal phenomena and numerous outliers. Thus, SARIMA, TBATS and Holt-Winters models will be adjusted to the daily data and only SARIMA and Holt-Winters will be adjusted to the monthly data.

The main objective of this paper is to model thenumber of these claims, favouring a good forecasting capability, in order to implement these models in practice. Therefore, a comparative study between the three methodologies is carried out.

In order to evaluate the models' forecasting capability, several evaluation measures are used, namely: MSE, RMSE, MAPE, MASE and U-Theil's statistic.

The results of the comparative study indicate that there are no major differences between the methodologies, since, in the case of daily data, all models present good adjustments, but their forecasts are not very useful because these series have an extremely strong seasonal component. On the other hand, the monthly models provide better forecasts because of their lower volatilities. In addition, the modeling of claims numbers is better than the modeling of the frequency rates.

Keywords: Holt-Winters; Insurance claim; SARIMA; TBATS; Time series.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

ADF – Augmented Dickey-Fuller (em português, Dickey-Fuller Aumentado)

AIC – Akaike's Information Criterion (em português, Critérios de Informação de Akaike)

AR – Autoregressive (em português, Autorregressivo)

ARFIMA – Autoregressive Fractionally Integrated Moving Average (em português, Autorregressivo e de Média Móveis)

ARIMA – Autoregressive Integrated Moving Average (em português, Autorregressivo e de Médias Móveis Integrado)

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

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

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

CV – Coeficiente de Variação Amostral

DNA – Danos por água

EAM – Erro Absoluto Médio

EEAM – Erro Escalado Absoluto Médio

EPAM – Erro Percentual Absoluto Médio

EPAMS – Erro Percentual Absoluto Médio Simétrico

EQM – Erro Quadrático Médio

ETS – Exponential Trignometric Smoothing (em português, Suavização Exponencial Trignométrica)

FAC – Função de Autocorrelação

FACP – Função de Autocorrelação Parcial

IC – Intervalo de Confiança

IPSS – Instituição Particular de Solideriedade Social

KPSS - Kwiatkowski-Phillips-Schmidt-Shin

MA – Moving Average (em português, Médias Móveis)

MAPE – Mean Absolute Percentage Error (em português, Erro Percentual Absoluto Médio)

MASE – Mean Absolute Scalde Error (em português, Erro Escalado Absoluto Médio)

MSE – Mean Square Error (em português, Erro Quadrático Médio)

NA – Não Aplicável

OLS – Ordinary Least Squares (em português, estimador de mínimos quadrados)

RB – Ruído branco

REL – Riscos elétricos

REQM – Raiz do Erro Quadrático Médio

RMSE – Root Mean Square Error (em português, Raiz do Erro Quadrático Médio)

SAR – Seasonal Autoregressive (em português, Autorregressivo Sazonal)

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

SARMA – Seasonal Autoregressive Moving Average (em português, Autorregressivo e de Médias Móveis Sazonal)

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

TAR – Threshold Autoregressive (em português, Autorregressivo com Limiar)

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

TMP – Tempestades

TSCov – Trigonometric Structural Models with Covariates (em portguês, Suavização Exponencial Trignométrica)

VAR - Vector autoregressive (em português, autorregressivo vetorial)

Capítulo 1

Introdução

Cada vez mais existem registos de sinistros, quer sejam causados pelo ser humano, quer pela natureza. Um crescimento demográfico a nível nacional implica, evidentemente, um aumento destes números (registados pelas seguradoras). Outro fator com enorme peso nestes registos é o das alterações climáticas. Este fenómeno, ao contrário do que se possa assumir em primeiro instante, além de verões mais quentes, contribui também para invernos mais frios, i.e., aumenta a variabilidade da temperatura terrestre. Isto resulta em tempestades mais intensas e com maior frequência, que acabam por se refletir nos mercados das seguradoras.

Nestes mercados, define-se um sinistro como um evento que resulta em prejuízo material para um indivíduo segurado. Existem vários tipos de seguros: de vida, de saúde, financeiros, automóvel, habitação, etc.. Cada uma destas áreas está inserida num mercado distinto das restantes, o que leva a abordagens diferentes nos processos de avaliação. Uma vez que este assunto está sujeito a vários fatores externos, é fundamental incorporar um estudo temporal nos números de sinistros, de forma a facilitar as tomadas de decisões nos vários setores.

Neste contexto, faz sentido elaborar um estudo sobre as séries temporais dos números de sinistros registados pela companhia de seguros Ageas. Existem várias metodologias de estudos temporais que podem ser aplicadas. Estas podem ser categorizadas como paramétricas ou não paramétricas. Os métodos paramétricos assumem que o processo estocástico estacionário associado aos dados em estudo tem uma estrutura que pode ser descrita através de parâmetros. Ao contrário destes modelos, onde o processo de modelação consiste em ajustar os seus parâmetros, os métodos não paramétricos avaliam, explicitamente, o espectro ou a autocovariância temporal do processo estocástico subjacente aos dados sem assumir que este admite qualquer tipo de estrutura. Além disso, para se efetuarem previsões de uma série temporal através de metodologias paramétricas, apenas são necessários certos parâmetros dos dados. Contrariamente, os métodos não paramétricos necessitam apenas dos próprios valores dos dados. Estes também se distinguem significativamente na praticalidade uma vez que os modelos não paramétricos apenas mantêm a

estrutura empírica dos dados, ao contrário das abordagens paramétricas que implicam a verificação de vários pressupostos sobre a estrutura de autocovariância temporal e a distribuição dos dados que podem não ser verificados na prática. Além disso, os métodos paramétricos têm a vantagem de ser mais rapidamente processados, porque assumem conhecer a distribuição do processo estocástico associado aos dados.

Os modelos estudados nesta dissertação baseiam-se na metodologia clássica de Box-Jenkins, em métodos de alisamento exponencial e na metodologia TBATS da autoria de Robin Hyndman [De Livera et al., 2011]. Os modelos SARIMA, abrangidos pela temática de Box-Jenkins, têm uma utilidade bastante versátil, pelo que são usados em inúmeros casos e são sempre um ótimo ponto de partida para qualquer estudo de séries temporais. Os modelos Holt-Winters são modelos de alisamento exponencial não paramétricos. Esta ferramenta é bastante útil em casos práticos porque os pressupostos de normalidade e autocorrelação temporal nem sempre são verificados na prática. Assim, uma vez que se trata de uma metodologia não paramétrica, estes pressupostos não são necessários, o que torna este modelo numa ferramenta muito forte, principalmente em estudos de dados com bastante estocasticidade (que é o caso dos números de sinistros registados pela companhia de seguros Ageas estudados neste trabalho). Para além da alta variabilidade presente nos dados deste estudo, também se verificam sazonalidades múltiplas nestas séries. O modelo TBATS é uma ótima escolha nestes casos, já que Robin Hyndman desenvolveu esta metodologia para facilitar a modelação de séries temporais com sazonalidade complexa.

Toda a análise apresentada doravante é efetuada com um nível de significância de 5% no software R, recorrendo às bibliotecas: readxl, zoo, forecast, astsa, MASS, urca e dplyr.

1.1 Apresentação da companhia de seguros Ageas

A companhia de seguros Ageas é uma empresa multinacional, marcando presença em vários países, tais como a Bélgica, o Reino Unido, a França, a Turquia e Portugal, bem como em nove países da Ásia, onde se incluem a China, a Malásia, a Tailândia, a Índia, as Filipinas e o Vietname.

A Ageas Seguros (nome introduzido em 2017, após a aquisição da AXA Portugal e Direct), marca do Grupo Ageas Portugal, é uma seguradora que oferece vários produtos na área de seguros vida e não-vida. Esta empresa tem como principal objetivo ajudar os seus Clientes a mitigar riscos relacionados a bens, sinistros, vida e pensões da forma mais tranquila e acessível possível. Esta é já uma das marcas de seguradoras mais conhecidas a nível nacional.

A operar em Portugal desde 2005, a Ageas continua o seu crescimento através de parcerias e contribuindo para o desenvolvimento do país através da Fundação Ageas (IPSS, Instituição Particular de Solideriedade Social, fundada em 1998, agregadora de pessoas e parceiros, visando a prossecução de fins de solideriedade social na comunidade) e apoiando os seus Clientes na ges-

tão, antecipação e proteção contra riscos e imprevistos, de forma a que possam viver de forma segura com menos preocupações. O Grupo Ageas Portugal tem vindo a crescer ao longo do tempo, com a aquisição de várias marcas comerciais, englobando: Ageas Seguros, Ageas Pensões, Médis, Ocidental e Seguro Directo, tornando-se, assim, num líder do ranking segurador português.

A Ageas Seguros oferece produtos a particulares na área automóvel, da saúde, de acidents pessoais, de viagem e lazer, de animais domésticos, de habitação, de acidentes de trabalho e de vida. As restantes áreas são abrangidas pelo ramo não-vida.

1.2 Objetivos e estrutura de dissertação

O objetivo principal desta dissertação é construir modelos que consigam extrapolar o número de sinistros que irão ser registados futuramente, i.e., estabelecer modelos de previsão no contexto da análise de séries temporais. Desta forma, o Capítulo 2 expõe uma revisão de literatura sobre os métodos de previsão de séries temporais e alguns critérios de avaliação de forma a poder escolher-se entre diversas metodologias de previsão. Também é abordado o problema da escolha entre os diversos métodos de previsão, identificando alguns possíveis critérios de avaliação.

Nos Capítulos 3 e 4 estão descritas as componentes teóricas relacionadas com os métodos de modelação de séries temporais, nomeadamente uma descrição dos conceitos fundamentais de séries temporais, três metodologias que serão aplicadas aos dados e as medidas de avaliação que serão utilizadas para este contexto.

Em prol do objetivo principal, serão abordados dados diários no Capítulo 5 e dados mensais no Capítulo 6. Em ambos os casos, será feita uma separação entre o estudo dos números totais de sinistros e o estudo das dos números das suas categorias marginais. Estes capítulos começam com uma análise exploratória dos dados, onde são destacados alguns outliers. Neste âmbito, estes serão "suavizados" com o objetivo de melhorar os ajustes dos modelos estudados nesta dissertação. Para além do estudo das séries de contagens, o Capítulo 6 incorpora ainda uma modelação das taxas de frequência de sinistros (mensais), após ser calculada a exposição média ao risco ao longo dos anos. Esta taxa aparece como o rácio entre o número de sinistros registados e o número médio de apólices expostas ao risco de cada mês. Para finalizar estes capítulos, é realizada uma comparação entre os vários modelos ajustados, com o apoio das medidas de avaliação apresentadas.

Por fim, as conclusões finais do trabalho e algumas sugestões para uma futura investigação são discutidas no Capítulo 7.

Capítulo 2

Revisão de literatura

O estudo de séries temporais surge, principalmente, no século XX, quando Yule [Udny Yule, 1927] estabelece a noção de uma série estocástica, apesar de já haver registos de tentativas determinísticas no século XIX. Este postolou que cada série é uma realização de um certo processo estocástico indexado no tempo [De Gooijer and Hyndman, 2006]. É, então, em 1970 que Box e Jenkins, baseando-se nos estudos de Yule [Yule, 1926] e Wold [Wold, 1938], desenvolvem um conjunto de ferramentas práticas em prol da construção de modelos autorregressivos e de médias móveis integrados (autoregressive integrated moving average, ARIMA) que são constituídos por três etapas: identificação, estimação e validação [Zhang, 2003, Tsay, 2000]. Nas décadas de 50 e 60, são, também, publicados os trabalhos de Brown [Brown, 1960], Holt [Holt, 1957] e Winters [Winters, 1960], onde são estabelecidos os modelos de alisamento exponencial, designados por modelos de Holt-Winters.

Contudo, estes modelos assumem linearidade que, não obstante, é uma suposição com bastante utilidade em muitos casos, mas, no final dos anos 70 e no início dos anos 80, começa a mostrar-se incapaz de ser verificada em vários casos reais [De Gooijer and Hyndman, 2006]. Assim, com o objetivo de ultrapassar esta restrição, começam a surgir modelos não lineares, tais como modelos bilineares e modelos autorregressivos com limiares (threshold autoregressive, TAR) [Tong, 1990], que permitem explicar alguns padrões não lineares. Nos últimos anos, a área de Machine Learning tem sido bastante utilizada na previsão de séries temporais através das suas redes neuronais artificiais, já que os modelos anteriores são desenvolvidos para explicar padrões não lineares específicos e podem ser insuficientes para modelar outras componentes não lineares [Zhang, 2003].

Além destes modelos, houve também outras abordagens que consideravam estudos em simultâneo de várias séries temporais relacionadas entre si, os modelos autorregressivos vetoriais (vector autoregressive VAR). Esta metodologia, por necessidade, deu à luz os modelos autorregressivos e de médias móveis fracionalmente integrados (autoregressive fractionally integrated moving average,

ARFIMA), que se adaptam a séries que assumem funções de autocorrelações empíricas com decaimentos mais lentos do que nos modelos ARIMA, i.e., séries com maior memória [Tsay, 2000]. Os trabalhos publicados por De Gooijer e Hyndman [De Gooijer and Hyndman, 2006] exploram, com mais detalhe, os métodos de previsão de séries temporais nos últimos 25 anos.

Mais recentemente, Chu e Zhang [Chu and Zhang, 2003], comparam os modelos clássicos lineares de Box-Jenkins e de regressão linear múltipla e os modelos não lineares de redes neuronais. Em cada caso, as componentes sazonais são modeladas de maneiras distintas, já que esta é explicada por variáveis indicatrizes (dummy) ou por funções trigonométricas. Para além destas abordagens, os modelos de redes neuronais também possibilitam a sua modelação sem decompor a série temporal, i.e., considerando a série como um todo, ou elaborando um ajuste sazonal (seasonally adjusted). Esta última metodologia tem sido altamente utilizada, recentemente, como métodos de previsão em diversas áreas. Existe uma vasta bibliografia a comparar estes modelos com a teoria clássica de Box-Jenkins, na qual se encontram os trabalhos de Zhang e Qi [Zhang and Qi, 2005] e Kuvulmaz [Kuvulmaz et al., 2005].

Numa tentativa audaz, Pan [Pan, 2013], Aburto e Weber [Aburto and Weber, 2007] decidem combinar os modelos SARIMA com modelos de redes neuronais, seguindo as palavras de Zhang [Zhang, 2003], "combinar diferentes modelos pode aumentar a probabilidade de capturar diferentes padrões dos dados e melhorar o desempenho das previsões.Vários estudos sugerem que, combinando modelos diferentes, a precisão da previsão pode ser melhorada em relação ao modelo individual. Além disso, o modelo combinado é mais robusto no que diz respeito à possível mudança de estrutura nos dados". Clemen [Clemen, 1989] possui publicações sobre estes modelos combinados.

Susana Lima [Lima, 2018] estuda um conjunto de séries temporais relativos ao segmento do retalho, com o objetivo de avaliar a precisão de vários métodos de previsão. Entre estes encontram-se os modelos ARIMA e os modelos de alisamento exponencial de Holt-Winters, uma vez que estes apresentam uma forte capacidade de explicar tendências e flutuações sazonais. Para tal, foram utilizadas várias métricas de avaliação de ajuste e previsão, tais como: o erro quadrático médio (EQM), a raiz do erro quadrático médio (REQM), o erro percentual absoluto médio (EPAM), o erro escalado absoluto médio (EEAM) e a estatística U de Theil.

Alpuim e El-Shaarawi [Alpuim and El-Shaarawi, 2008] afirmam que a estimação de tendências em séries temporais é um processo recorrente em diversas áreas da aplicação da Estatística, e.g., a Econometria, a Epidemiologia e a Estatística Ambiental. Este procedimento, geralmente, recorre a modelos de regressão que assumem o tempo como variável independente, por vezes em conjunto com outras covariáveis. No entanto, nem sempre é verificada a ausência de autocorrelação dos resíduos dos modelos, o que pode tornar as estimações da variância dos estimadores de mínimos quadrados ordinários (OLS, Ordinary Least Squares) incorreta.

Um dos passos mais importantes na modelação de séries temporais é a escolha do modelo mais apropriado para os dados em estudo. Tanto os modelos clássicos, ARMA, ARIMA e SARIMA estudados por Box e Jenkins [Box et al., 2016], Lee, Ko [Lee and Ko, 2011], Pappas [Pappas et al., 2008] e Chen [Chen et al., 1995], como os modelos de alisamento exponencial trabalhados por Taylor [Taylor, 2003], Kostenko e Hyndman [Kostenko and Hyndman, 2008] são aptos para explicar séries sem padrões sazonais complexos. Para séries que admitam sazonalidades complexas, De Livera e Hyndman [De Livera et al., 2011] introduziram, há alguns anos, os modelos Trignometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal components (TBATS).

Para atenuar as limitações dos modelos de espaço de estados na previsão de séries temporais com sazonalidades complexas (sazonalidades de alta frequência, sazonalidades múltiplas, sazonalidade de calendário duplo, etc.), De Livera e os seus coautores [De Livera et al., 2011] efetuaram algumas alterações, dando origem aos modelos BATS (Box-Cox Transformation, ARMA errors, Trend and Seasonal components) e TBATS. Estas alterações envolvem transformações de Box-Cox, séries de Fourier com coeficientes não constantes e correções dos erros ARMA. Uma das maiores vantagens desta evolução é a diminuição de processamento computacional na estimação da máxima verosimilhança, para além da alta versatilidade que oferece em termos de aplicação, como mostram em três estudos empíricos. Este melhoramento também mostra como a formulação trignométrica pode ser vista como um método de decompor séries temporais complexas, que, por sua vez, possibilitam a recolha de informação fulcral, tal como a identificação de componentes sazonais que não seriam detetadas à priori.

O trabalho de Brożyna [Brożyna et al., 2018] ilustra a capacidade preditiva do modelo TBATS, que não tem restrições quanto à sazonalidade. Este diz, ainda, que independentemente da variável que se pretenda prever, é importante escolher o modelo que explica melhor o processo da série no passado. Geralmente a variação de uma série temporal pode ser decomposta em três componentes: a tendência, a componente sazonal/cíclica e a componente irregular. As previsões de séries temporais normalmente recorrem a vários valores passados para incorporar uma única sazonalidade e um período de previsão próprio. Isto é, para prever valores mensais à distância de um ano são utilizados, por exemplo, dados dos últimos 10 anos. O modelo TBATS tem a capacidade de calcular previsões a médio prazo com a exatidão de uma previsão a curto prazo, utilizando séries temporais específicas. Brożyna [Brożyna et al., 2018] recolheu dados sobre a procura de energia alétrica na Polónia num período de 14 anos, conseguindo incorporar três sazonalidades distintas no seu modelo.

Noutro estudo, Naim [Naim et al., 2018] compara os modelos TBATS e BATS para previsões a curto prazo de séries com sazonalidade complexas. Ao ajustar modelos ETS (Exponential Trigonometric Smoothing) e SARIMA, nota-se que estes são eficazes para descrever comportamentos sazonais simples, mas são insuficientes para modelar componentes sazonais complexas. O estudo

relembra que os casos de séries com sazonalidades complexas são cada vez mais comuns, tais como: consumos de combustíveis fósseis, o atendimento a call centers num certo período, etc.. Estes casos não admitem uma sazonalidade simples, mas sim uma sazonalidade dinâmica, o que leva a recorrer à aplicação dos modelos BATS e TBATS.

Puindi mostra, na sua tese de doutoramento [Puindi, 2018], que modelos estruturais de espaço de estados são eficazes para modelar séries com comportamentos sazonais complexos. Porém, muitas vezes o cerne de um estudo recai sobre o poder preditivo do modelo. Assim, qualquer informação sobre os dados é essencial para melhorar o desempenho do modelo. Este autor refere que apenas a abordagem TBATS está apta para modelar séries com sazonalidades complexas. Puindi também contribuiu para a elaboração de modelos estruturais dinâmicos, admitindo efeitos das covariáveis. Construiram-se os modelos SCov (Structural Models with Covariance), que renova a ideia de métodos tradicionais de suavização exponencial sazonal simples e os modelos TSCov (Trignometric Structural Models with Covariates), que se resume a uma extensão do modelo TBATS. Estes modelos definem-se à custa de três componentes não constantes ao longo do tempo e não visíveis à priori: o nível, a tendência e a sazonalidade.

Segundo Shu [Shu et al., 2014], a teoria clássica deBox-Jenkins é o método mais utilizado no estudo de séries temporais. Logo esta pode ser melhorada em termos de precisão dos modelos SARIMA, minimizando os resíduos com o auxílio de séries de Fourier.

Assim, dada a enorme oferta de métodos de previsão de séries temporais, surge a questão "Que critério utilizar para a escolha da metodologia a aplicar?". Yokuma e Armstrong [Yokuma and Armstrong, 1995] realizam dois diferentes estudos, com o intuito de encontrar uma resposta a esta pergunta. O primeiro estudo avalia se os critérios diferem de acordo com o papel principal do inquirido (e.g., acionista, investigador, etc.) e o segundo tenciona verificar se os critérios variam conforme a natureza das previsões (e.g., tipo de dados, quantidade de dados disponível, etc.). Ambos os estudos, em conjunto com outros anteriores, concluem que o fator mais importante é, de facto, a precisão, independentemente da função do inquirido ou da natureza das previsões. Contudo, nem sempre é possível conduzir um estudo com base apenas neste critério, pois, por vezes, a facilidade de implementação, o custo, a flexibilidade, o tempo disponível, entre outros contra-tempos, devem ter sidos em conta [Yokuma and Armstrong, 1995]. De facto, este critério é apoiado por vários estudos [Alon et al., 2001, Chu and Zhang, 2003, Ramos et al., 2015, da Veiga et al., 2016].

Para além dos critérios de escolha de metodologias, também é importante conseguir quantificar o desempenho preditivo dos modelos. Existem vários para o caso, mas nem todos são universais [Hyndman et al., 2006, Hyndman and Koehler, 2006]. Por exemplo, medidas que dependem da escala dos dados (e.g., o erro quadrático médio, EQM, a raiz do erro quadrático médio, REQM e o erro absoluto médio, EAM) são práticos para comparar diferentes modelos aplicados ao mesmo

conjunto de dados, mas não devem ser usados quando os modelos são aplicados a dados com escalas distintas. Para tal, existem erros percentuais que não dependem da escala dos dados, e.g., o erro percentual absoluto médio, EPAM, porém podem fornecer resultados inconclusivos, pois podem admitir valores indefinidos ou infinitos se os dados possuírem valores nulos. Para além dos erros percentuais, existem também erros relativos e erros escalados.

Utilizando três séries representativas da realidade, Hyndman e Koehler [Hyndman and Koehler, 2006] mostram a incapacidade de várias medidas de avaliação de precisão. Estas séries assumem valores nulos e negativos, pelo que os autores recomendam a utilização de medidas de erros escalados para estes casos, tais como o erro escalado absoluto médio (EEAM), que é de fácil interpretação, sempre finito e com alta aplicabilidade. Porém, estes autores admitem a existência de casos onde as medidas tradicionais continuam a ser mais favoráveis, pela sua simplicidade. Uma vez que não existe unanimidade na escolha dos critérios de avaliação de previsão, é recomendado o uso de diversas métricas para ajudar na avaliação preditiva dos modelos [Chu and Zhang, 2003, Ramos et al., 2015].

Capítulo 3

Séries temporais

A área de estudo de séries temporais consiste em trabalhar dados que se caracterizam como um conjunto de observações que se distribuem sequencialmente ao longo de um período de tempo, por exemplo: preços diários de ações, exportações mensais, temperatura, quantidade de chuva caída numa região, velocidade do vento, vendas mensais/semanais/anuais, eletrocardiogramas, etc..

Em todos estes casos, a ordem das observações é essencial. Para além disso, as observações são, geralmente, dependentes, i.e., o valor de X_t depende do(s) valor(es) de X_s para certo(s) $s < t$. Uma vez que se pretende:

1) Descrever propriedades da série (e.g. a tendência, fenómenos sazonais/cíclicos, outliers, etc.) e perceber o mecanismo gerador da série, ou seja, encontrar motivos ou acontecimentos que justifiquem o comportamento da série;

2) Usar a variabilidade de uma série para explicar a variabilidade de outra;

3) Predizer valores futuros com o auxílio de modelos estatísticos, baseando-se em observações passadas (sempre visando aumentar a exatidão o máximo quanto possível).

Esta dependência temporal é fulcral, pois só é possível fazer predições à custa dos valores a que se tem acesso, os valores x_s. Ou seja, havendo dependência, consegue-se estimar o valor de X_t com o(s) valor(es) x_s. No entanto, há casos de séries temporais de observações independentes que implicam uma abordagem diferente da clássica.

Costumam tratar-se de séries temporais como séries contínuas, mesmo quando estas são observações de fenómenos discretos [Chatfield, 2000, Chatfield, 2003], como é o caso das vendas mensais/semanais/anuais (note-se que os instantes de tempo serão sempre discretos) onde a variável pode ser 0 ou 15 000 euros, por exemplo. No tratamento de séries temporais contínuas,

frequentemente, estas são transformadas em discretas, através de amostragens em intervalos de tempo iguais (o mais pequeno possível com o intuito de se perder o mínimo de informação possível).

Para cada instante de tempo t deve existir uma observação x_t (uma realização da v. a. correspondente X_t; que corresponde a uma série temporal univariada). Caso se registem mais do que uma observação para cada instante t, a série temporal classifica-se como multivariada. Pressupõe-se, ainda, que[1]

$$t_i - t_{i-1} = c,\ c \in \mathbb{R}^+,\ \forall\, i = 2, 3, \ldots, \underset{t_i \in T_0}{max}\{i\},$$

i.e., a frequência de amostragem (intervalo de tempo entre duas observações consecutivas) deve ser constante ou, pelo menos, aproximadamente constante. Por exemplo, as observações diárias registadas todos os dias úteis por várias semanas apresenta aproximações dos intervalos de observação, uma vez que apenas se tem a garantia que as observações são registadas num certo dia, mas independentemente da hora (o que resulta numa aproximação do intervalo de observação). Por outro lado, se o estudo se basear em observações registadas a cada hora por vários dias, tem-se a garantia que a frequência de amostragem é de exatamente uma hora.

Caso esta frequência de amostragem não seja respeitada, surgem valores omissos que implicam um tratamento específico. Muitas vezes recorre-se a interpolação de forma a poder construir-se uma série temporal.

Nota 1. Um modelo estatístico não descreve a variação total das observações, mas sim a sua tendência, i.e., os dados podem ter qualquer tipo de tendência (crescente, decrescente ou nula), mas sempre com variabilidade presente. No entanto, estes modelos ajudam a descrever e prever fenómenos futuros. Este é um dos principais objetivos da análise de séries temporais.

3.1 Componentes de uma série temporal

Uma série temporal pode ser descrita através de quatro principais componentes [Chatfield, 2000]: a tendência ou componente sistemática (T), a componente sazonal (S), a componente cíclica (C) e a componente residual ou ruído (R). Sendo X_t o valor da série temporal no instante t, estas componentes podem ser combinadas de duas formas:

$$X_t = T + S + C + R, \qquad \text{(decomposição aditiva)}$$
$$X_t = T \times S \times C \times R\,. \qquad \text{(decomposição multiplicativa)}$$

[1] Geralmente considera-se $T_0 = \mathbb{N}$.

O modelo baseado na decomposição multiplicativa aplica-se quando a amplitude das oscilações sazonais da série dependem da tendência (caso contrário, opta-se usualmente pelo modelo de decomposição aditiva), i.e., se existir algum tipo de proporcionalidade (direta ou inversa) entre a tendência e a amplitude [Wheelwright et al., 1998]. Contudo, o modelo de decomposição multiplicativa, pode ser transformado logaritmicamente num modelo de decomposição aditiva (note-se que a componente cíclica C não depende do instante de tempo t)

$$ln\left(X_t\right) = ln\left(T_t \times S_t \times R_t\right) \iff ln\left(X_t\right) = ln\left(T_t\right) + ln\left(S_t\right) + ln\left(R_t\right) \ . \tag{3.1}$$

Esta transformação permite uma análise aditiva da série $Y_t = ln\left(X_t\right)$, no entanto apenas pode ser aplicada em casos de séries temporais de valores estritamente positivos.

Além destas duas decomposições, por vezes é apropriado um modelo de decomposição mista, que resulta em modelos com relações aditivas e multiplicativas, e.g., um modelo multiplicativo com erros aditivos

$$X_t = T_t \times S_t + R_t \ .$$

A escolha do modelo de decomposição é apoiada fortemente pelos gráficos de decomposição (instaurados por [Clevele and Terpenning, 1982]) que permitem a visualização gráfica das várias componentes isoladamente. Estes gráficos são obtidos com o uso de filtros de médias móveis.

A **Tendência (T)** é a inclinação da série temporal ao longo do período observado, que pode ser linear, não-linear, positiva (caso esta seja crescente), negativa (caso esta seja decrescente) ou nula (caso esta seja constante). Neste último caso diz-se que a série é estacionária para a média.

A **Sazonalidade (S)** traduz-se nas oscliações periódicas da série temporal. Esta é uma característica previsível, pois para além de se estar à espera que aconteça, é também possível prever os instantes de tempo em que se vai manifestar. Esta sazonalidade pode ser diária, semanal, mensal, anual, ou de qualquer outra periodicidade, podendo também ser caracterizada como aditiva ou multiplicativa. Este último caso equivale ao aumento ou diminuição da amplitude das oscliações sazonais consoante a monotonia da tendência da série temporal. Tome-se como exemplo a série temporal que representa as quantidades mensais (desde 1949 até 1960) de passageiros que viajam numa certa companhia aérea (Figura 3.1).

Esta série mostra uma sazonalidade anual, pois nota-se uma maior afluência a viagens aéreas nas épocas de verão. Para além disso, também é visível uma proporcionalidade direta entre a sazonalidade e a tendência da série temporal. Ou seja, o aumento da amplitude sazonal com o aumento da tendência evidencia um modelo de decomposição multiplicativa para a componente sazonal. O ajustamento sazonal é geralmente aplicado quando se pretende remover a oscilação

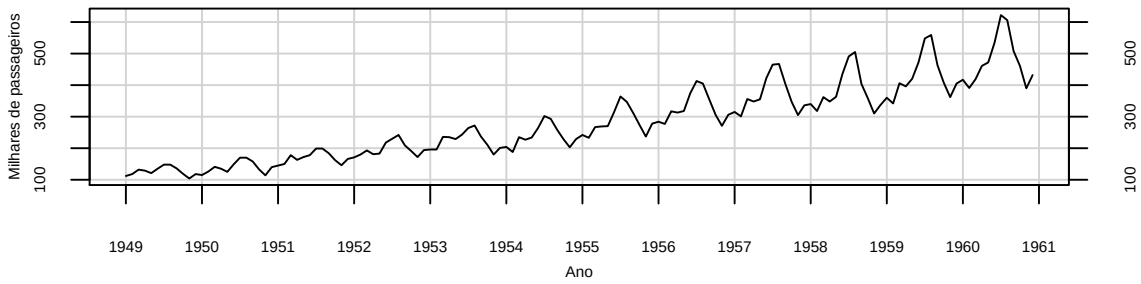


Figura 3.1: Série mensal do número de passageiros de uma certa companhia aérea (1949-1960).

da série temporal, que pode estar a omitir outras componentes descritivas da série, tais como a tendência.

Aplicando a transformação logarítmica referida anteriormente 3.1 obtém-se uma nova série temporal que admite uma decomposição aditiva (Figura 3.2).

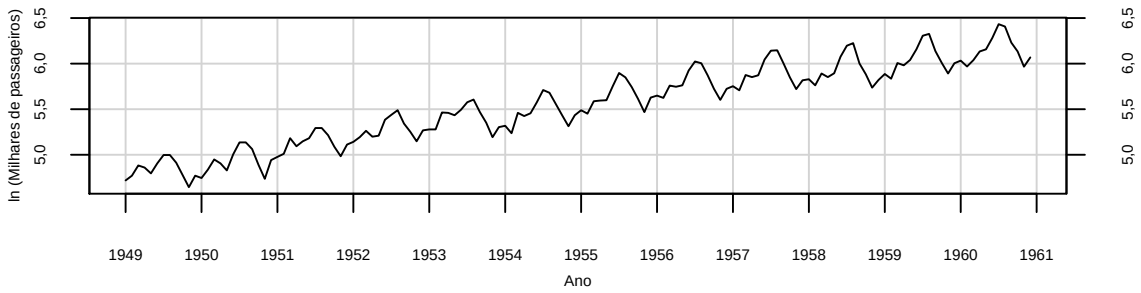


Figura 3.2: Série mensal do logaritmo do número de passageiros de uma certa companhia aérea (1949-1960).

A componente **Cíclica (C)** representa oscilações da série temporal em torno da sua tendência que, ao contrário da componente sazonal, não apresentam uma frequência de observação fixa (i.e., a sua duração não é constante). Esta componente é mais complicada de se prever, contudo é muitas vezes ignorada em séries temporais a curto prazo, uma vez que esta pode prolongar-se por vários anos. Tome-se como exemplo as medidas mensais da pressão do ar associadas às temperaturas na superfície do oceano Pacífico entre 1950 e 1987 (Figura 3.3).

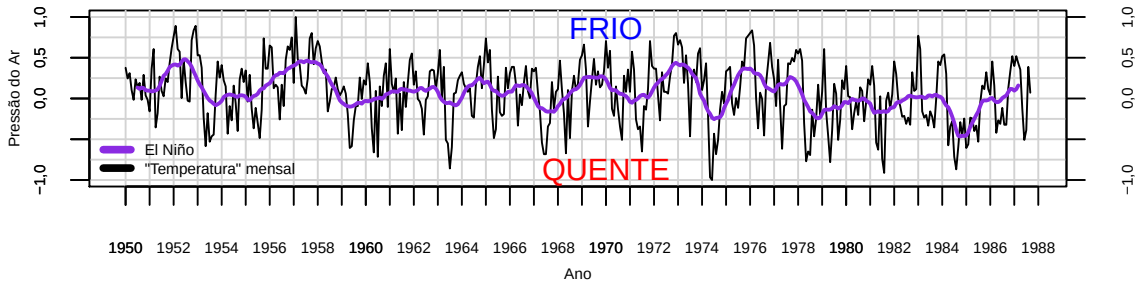


Figura 3.3: Série mensal da pressão do ar associadas às temperaturas na superfície do oceano Pacífico (1950-1987).

Este gráfico evidencia a influência do fenómeno conhecido por "El Niño"[2] na pressão do ar.

[2]Massa de água quente que é criada no centro/este-centro do Pacífico equatorial.

Note-se que, para além da sazonalidade anual da temperatura (devido às estações do ano), existe uma componente cíclica justificada pelo "El Niño" cuja duração varia entre cerca de 3 a 7 anos, o que resulta num obstáculo em termos preditivos.

A componente **Aleatória/Residual (R)** corresponde à variação da série temporal não explicada pelo modelo. Esta é puramente aleatória e também é conhecida como **ruído** ou **resíduo**.

3.2 Processos Estocásticos

Toda a série temporal está associada a um processo estocástico.

Definição 1. Um processo estocástico é uma família de variáveis aleatórias $\{X_t\}_{t\in T_0}$ ($T_0 \subset \mathbb{R}$ discreto) com contradomínio S (espaço de estados) que quando realizadas parcialmente (finitamente) representam uma série temporal. Para este ser devidamente caracterizado é necessário especificar todas as distribuições de todas as combinações de variáveis $(X_{t_0}, X_{t_1}, \dots, X_{t_n})$, para todos os valores possíveis de n e t_i, com $i = 0, 1, \dots, n$.

Uma vez que a especificação de todas as distribuições de todas as combinações de variáveis $(X_{t_0}, X_{t_1}, \dots, X_{t_n})$ é um assunto intricado, na prática resume-se a caracterizar os momentos do processo, em particular o primeiro e segundo momentos:

$$E[X_t] = \mu(t), \qquad \text{(primeiro momento)}$$

$$E\left[\left(X_{t_1} - \mu_{t_1}\right)\left(X_{t_2} - \mu_{t_2}\right)\right] = \gamma(t_1, t_2)\ . \qquad \text{(segundo momento)}$$

O segundo momento é conhecido como **função de autocovariância** e equivale a uma generalização da variância σ_t^2, pois esta última é o caso específico de quando $t_1 = t_2$, i.e., $\gamma(t,t) = E[(X_t - \mu_t)(X_t - \mu_t)] = \sigma_t^2$.

Tal como foi referido anteriormente, uma série temporal é apenas uma realização de um certo processo estocástico. Ou seja, um processo estocástico pode ter várias realizações resultando em várias séries temporais distintas. Assim, analogamente à inferência estatística onde se retiram conclusões baseadas numa amostra de uma certa população, nas séries temporais retiram-se conclusões sobre um processo estocástico baseando-se numa série temporal por este gerada [Cordeiro, 2011].

Os processos estocásticos podem (ou não) ser estacionários. Na Secção 3.2.1 serão introduzidas as noções de estacionariedade de um processo estocástico que é um requisito necessário para várias ferramentas no estudo de séries temporais.

3.2.1 Estacionariedade

Muitas vezes surgem séries não estacionárias com várias irregularidades, sobre as quais não se conseguem elaborar grandes estudos. Uma vez que estacionaridade é um requisito para muitas técnicas estatísticas utilizadas em análise de séries temporais, o tratamento usual começa por transformar os dados de forma a impor-se a estacionariedade. Esta característica é muito importante no estudo de séries temporais, pois é baseada nela que se fazem estimativas, uma vez que estabelece regularidade nos dados.

Definição 2. Um processo estocástico $\{X_t\}_{t\in T_0}$ diz-se fortemente estacionário se e só se a distribuição conjunta de qualquer conjunto $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ for a mesma que o conjunto $(X_{t_1+k}, X_{t_2+k}, \dots, X_{t_n+k})$ para $\underset{i=1,2,\dots,n}{max}\{t_i + k\} \in T_0$, ou seja, as funções de distribuição conjunta de dimensão finita são invariantes por translações no tempo, i.e.,

$$F_{(X_{t_1},X_{t_2},\dots,X_{t_n})}(x_1, x_2, \dots, x_n) = F_{(X_{t_1+k},X_{t_2+k},\dots,X_{t_n+k})}(x_1, x_2, \dots, x_n),$$

$$\forall\ (x_1, x_2, \dots, x_n) \in \mathbb{R}^n.$$

A estacionariedade forte é uma característica bastante útil, contudo demasiado difícil de ser verificada em casos práticos. Por este motivo, estabelece-se a definição de estacionariedade para a covariância (ou de segunda ordem).

Definição 3. Um processo estocástico $\{X_t\}_{t\in T_0}$ diz-se fracamente estacionário (ou estacionário para a covariância) se e só se os primeiro e segundo momentos de $\left(X_{t_1}, X_{t_2}, \dots, X_{t_n}\right)$ existirem e forem iguais aos momentos correspondentes de $\left(X_{t_1+k}, X_{t_2+k}, \dots, X_{t_n+k}\right)$. Isto é, se:

1. $E(X_t) = \mu, \forall\ t \in T_0$, o valor médio é finito e não depende do instante de tempo t;
2. $Var(X_t) = \sigma^2, \forall\ t \in T_0$, a variância é finita e não depende do instante de tempo t;
3. $Cov(X_t, X_s) = \gamma(|t-s|), \forall\ t, s \in T_0$, a covariância entre duas variáveis aleatórias depende apenas do seu desfazamento temporal.

Pelas Definições 2 e 3, se $\{X_t\}_{t\in T_0}$ é um processo fortemente estacionário com primeiro e segundo momentos finitos, então $\{X_t\}_{t\in T_0}$ também é fracamente estacionário, apesar da proposição recíproca não ser necessariamente verdadeira. Porém, se $\{X_t\}_{t\in T_0}$ admite uma distribuição Normal, então as definições são equivalentes (Tsay, 2010).[3]

A transformação de Box-Cox é o método mais comum de se obter estacionariedade para a variância de uma série. Esta metodologia transforma os dados de forma a que se consiga assumir

[3] Por questões de simplicidade, designar-se-ão processos fracamente estacionários por apenas processos estacionários.

que estes provêm de uma distribuição Normal (ou pelo menos que a sua distribuição se assemelha suficientemente de uma distribuição Gaussiana). Esta transformação é dada por

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & , \text{ se } \lambda \neq 0 \\ ln\, x_i & , \text{ se } \lambda = 0 \end{cases} , \tag{3.2}$$

onde x_i tem de ser estritamente positivo. O parâmetro λ é estimado maximizando a função de máxima verosimilhança das observações $x_i^{(\lambda)}$, partindo do princípio que $x_i^{(\lambda)} \dot{\sim} N(\mu, \sigma^2)$.

Desta forma obtém-se uma série $x_i^{(\lambda)}$ transformada, ou seja, a interpretação desta variável não é a mesma que a da x_i. Por este motivo, após ter sido ajustado um modelo à série $x_i^{(\lambda)}$, é necessário aplicar uma transformação inversa aos valores previstos pelo modelo, de modo a conseguir-se interpretar o valor de x_i, logo

$$x_i = \begin{cases} \sqrt[\lambda]{\lambda x_i^{(\lambda)} + 1} & , \text{ se } \lambda \neq 0 \\ e^{x_i^{(\lambda)}} & , \text{ se } \lambda = 0 \end{cases} . \tag{3.3}$$

3.2.2 Autocorrelação e autocovariância

No contexto de processos estacionários faz sentido falar em correlações e covariâncias entre variáveis aleatórias. Contudo, uma vez que um processo estocástico define-se à custa de uma variável aleatória avaliada ao longo do tempo, estes conceitos alargam-se para correlações e covariâncias entre uma variável num instante de tempo t e a própria desfazada no tempo (noutro instante de tempo $t+k$). Nestes casos falam-se de **autocorrelações** e **autocovariâncias**. Caso o valor da autocovariância seja pequeno, os valores da série temporal aproximam-se da média (reversão da média) mais rapidamente e mais lentamente caso contrário.

Definição 4. Dado um processo estocástico estacionário $\{X_t\}_{t \in T_0}$, define-se a função de autocovariância como

$$\gamma_k = Cov\left(X_t, X_{t+k}\right) = E\left[\left(X_t - \mu\right)\left(X_{t+k} - \mu\right)\right],$$

que quantifica a intensidade com que covaria um par de valores do processo estocástico com lag (desfazamento) de k unidades de tempo.

O estimador mais comum desta função perante um processo estocástico estacionário é dado por

$$\hat{\gamma}_k = \frac{1}{n} \sum_{i=1}^{n-k} \left(x_i - \overline{x}\right)\left(x_{i+k} - \overline{x}\right).$$

Quanto maior for o lag k, maior será o desvio entre o estimador $\widehat{\gamma}_k$ e a função γ_k. Por este motivo, não é habitual estimar esta função para além dos primeiros $\frac{n}{4}$ valores de k [Chatfield, 2000].

A função γ_k deve respeitar as seguintes propriedades:

1. $\gamma_0 = Cov\left(X_t, X_t\right) = Var\left(X_t\right) = \sigma^2$;
2. $\gamma_k = \gamma_{-k}$, i.e., a função depende apenas do valor absoluto do lag;
3. $|\gamma_k| \leqslant \gamma_0$ como consequência da desigualdade de Cauchy-Schwarz[4];
4. É semidefinida positiva, ou seja,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\gamma(|t_i - t_j|) \geqslant 0,$$

onde $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$ e $t_1, t_2, \ldots, t_n$ são os instantes de tempo.

A função de autocorrelação mede a correlação entre dois valores de uma série temporal com lag k. Assim, ρ_k representa uma medida de similitude entre estes valores [Murteira et al., 1993].

Definição 5. Dado um processo estocástico estacionário $\{X_t\}_{t\in T_0}$, define-se a função de autocorrelação (FAC) como

$$\rho_k = Corr\left(X_t, X_{t+k}\right) = \frac{Cov\left(X_t, X_{t+k}\right)}{\sqrt{Var\left(X_t\right)Var\left(X_{t+k}\right)}} = \frac{Cov\left(X_t, X_{t+k}\right)}{Var\left(X_t\right)} = \frac{\gamma_k}{\gamma_0}.$$

O estimador mais utilizado para esta função é descrito pela expressão

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{i=1}^{n-k}\left(x_i - \overline{x}\right)\left(x_{i+k} - \overline{x}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}.$$

A representação gráfica desta função designa-se por correlograma e ilustra características essenciais sobre a série que servem de apoio na escolha do modelo. Como seria de esperar numa série temporal, observações mais próximas tendem a ter mais influência, ou seja, tendem a mostrar maiores valores de ρ_k e de γ_k, o que se reflete numa proporcionalidade inversa entre k e estas funções. Desta forma, espera-se que $(k \to +\infty) \implies (\rho_k \to 0 \ \wedge \ \gamma_k \to 0)$. Além disso, a função de autocorrelação deve respeitar as seguintes propriedades:

1. $\rho_0 = Corr\left(X_t, X_t\right) = 1$;
2. $\rho_k = \rho_{-k}$, i.e., a função depende apenas do valor absoluto do lag;

[4] Desigualdade de Cauchy-Schwarz: $E\left(XY\right) \leqslant \sqrt{E\left(X^2\right)E\left(Y^2\right)}$.

3. $|\rho_k| \leqslant 1$, como consequência da desigualdade de Cauchy-Schwarz;

4. É semidefinida positiva, ou seja,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\rho(|t_i - t_j|) \geqslant 0,$$

onde $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$ e $t_1, t_2, \dots, t_n$ são os instantes de tempo.

As autocorrelações para instantes sucessivos são transitivamente dependentes, ou seja, estando o primeiro elemento de uma série correlacionado com o segundo e o segundo com o terceiro, então existe correlação entre o primeiro e o terceiro elementos. Porém, em muitos casos interessa estudar a correlação parcial entre duas variáveis X_t e X_{t+k}, sem o efeito das $k-1$ variáveis intermédias $X_{t+1}, X_{t+2}, \dots, X_{t+k-1}$ [Caiado, 2011]. Por este motivo, define-se a função de autocorrelação parcial.

Definição 6. Dado um processo estocástico estacionário $\{X_t\}_{t\in T_0}$, define-se a função de autocorrelação parcial (FACP) como o conjunto $\{\phi_{kk} : k \in \mathbb{N}\}$ de autocorrelações parciais de lag k, onde

$$\phi_{kk} = Corr\left[X_t, X_{t+k} \mid X_{t+1}, X_{t+2}, \dots, X_{t+k-1}\right] = \frac{|P_k^*|}{|P_k|}$$

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

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

Considere-se um modelo de regressão linear múltipla

$$X_{t+k} = \phi_{k1}X_{t+k-1} + \phi_{k2}X_{t+k-2} + \cdots + \phi_{kk}X_t + \epsilon_{t+k}, \tag{3.4}$$

onde $\phi_{ki},\ i = 1, 2, \dots, k$ são os coeficientes do modelo que assumem erros Gaussianos. Os erros $\epsilon_t \sim N(0, \sigma^2), t \in \mathbb{Z}$ e ϵ_{t+k} são independente de $\{X_{t+k-j}, j \geqslant 1\}$. Por definição, ϕ_{kk} é o resultado do valor esperado da variável resposta X_{t+k} quando $X_i = 0, \forall\ i \neq t$ e $X_t = 1$, ou seja, representa a variação média de X_{t+k} por cada incremento unitário da variável X_t quando as restantes variáveis $X_{t+1}, X_{t+2}, \dots, X_{t+k-1}$ permanecem constantes. Esta variação pode ser interpretada como a correlação parcial entre X_t e X_{t+k}.

Assumindo, sem perda de generalidade, que $E(X_t) = 0, \forall\ t \in T_0$, multiplicando ambos os membros de (3.4) por X_{t+k-j}, calculando os valores esperados e dividindo por γ_0 obtêm-se as equações de Yule-Walker:

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-k},\ j = 1, 2, \ldots, k. \tag{3.5}$$

Aplicando a regra de Cramer à equação (3.5) ou partindo diretamente da Definição 6, obtêm-se os seguintes resultados:

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

3.2.3 Ruído branco

O caso mais conhecido de um processo estocástico estacionário é designado por ruído branco. Este processo define-se como uma sucessão de variáveis aleatórias não correlacionadas e identicamente distribuídas $\{W_t\}_{t\in T_0}$ com média e variância constantes.

Definição 7. Dado um processo estocástico $\{W_t\}_{t\in T_0}$, este define-se como um processo de ruído branco se e só se verificar as seguintes condições:

1. $E[W_t] = \mu_W$ (geralmente, $\mu_W = 0$);
2. $Var[W_t] = \sigma_W^2$;
3. $Corr(W_t, W_{t+k}) = 0,\ k = \pm 1, \pm 2, \ldots, \pm \underset{t+k\in T_0}{max}\{k\}$.

Sendo as variáveis não correlacionadas entre si, deduz-se que estas apresentam covariância nula, pois

$$Cov(W_i, W_j) = Corr(W_i, W_j)\sqrt{Var(W_i)Var(W_j)} = 0 \times \sqrt{Var(W_i)Var(W_j)} = 0.$$

Uma vez que $Cov(W_i, W_j) = 0$ e as suas média e variância são constantes ao longo do tempo. Pela Definição 3, conclui-se que $\{W_t\}_{t\in T_0}$ é um processo fracamente estacionário. Ora, se estas variáveis aleatórias seguirem uma distribuição Normal, ou seja, se $W_t \sim N(\mu_W, \sigma_W^2)$, então este trata-se de um processo de ruído branco Gaussiano e é fortemente estacionário. Sendo as variáveis aleatórias não correlacionadas e Gaussianas com os mesmos parâmetros, então são independentes e identicamente distribuídas, logo a distribuição de qualquer conjunto $(W_{t_0}, W_{t_1}, \ldots, W_{t_n})$ é

a mesma que o conjunto $(W_{t_0+k}, W_{t_1+k}, \ldots, W_{t_n+k})$, respeitando todas as condições da Definição 2.

Um ruído branco é, portanto, um processo que admite funções de autocorrelação (FAC) e autocorrelação parcial (FACP) nulas para todos os lags $k \neq 0$. A Figura 3.4 ilustra o comportamento aleatório deste tipo de processos ao longo do tempo e, simultaneamente, as FAC e FACP correspondentes que apresentam baixos valores de autocorrelação e de autocorrelação parcial.

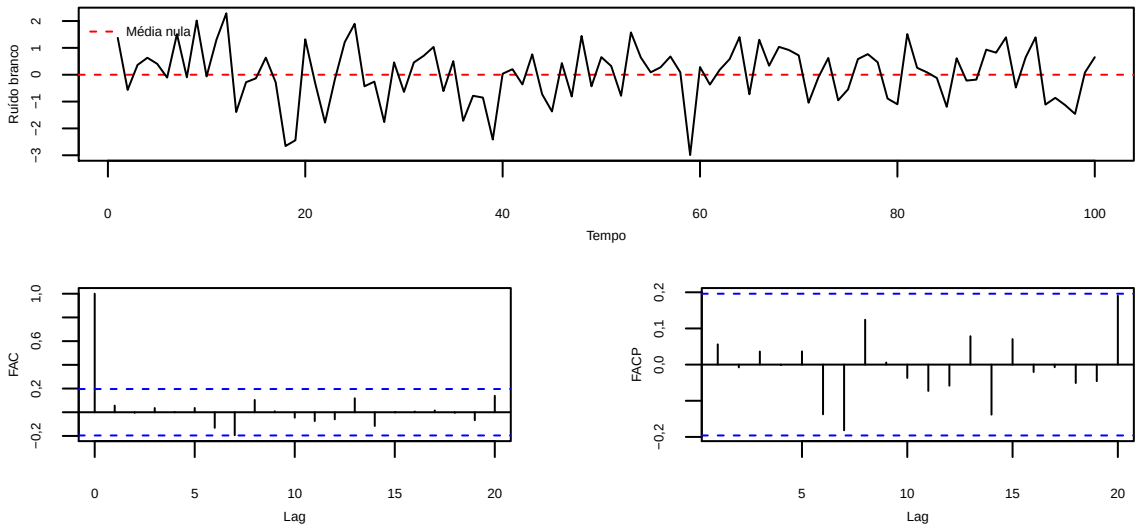


Figura 3.4: Simulação de um ruído branco de média nula e variância unitária e respetivas FAC e FACP empíricas.

Um bom modelo é aquele que explica toda a componente sistemática dos dados, consequentemente criando erros aleatórios com um comportamento análogo ao de um ruído branco [Caiado, 2011]. Por este motivo, este processo tem um papel muito importante na construção de modelos probabilísticos ou estocásticos. Contudo, o ruído branco não é frequentemente visualizado na modelação de séries reais, pois é extremamente difícil descobrir o processo gerador exato na exploração de casos reais.

3.2.4 Processos estocásticos não estacionários

A estacionariedade é uma característica essencial na modelação de séries temporais. Alguns modelos assumem que a série é estacionária na sua natureza ou que pode tornar-se estacionária após uma transformação [Jebb et al., 2015].

Muitas séries temporais não são estacionárias, porque a média ou a variância dependem do tempo em vez de serem constantes. Note-se que uma série estacionária em média não é necessariamente estacionária em variância. Tal como a transformação de Box-Cox, existem outras transformações que podem ser aplicadas a séries temporais de modo a introduzir estacionariedade.

Quando a estacionariedade está ausente tanto na média como na variância, deve-se estabilizar a variância antes de se estabilizar a média [Murteira et al., 1993, Caiado, 2011]. A Figura 3.5 ilustra 3 casos de ausência de estacionariedade: no topo, uma série não estacionária em variância mas estacionária em média (variância crescente e média constante), no centro, uma série não estacionária em média mas estacionária em variância (média crescente e variância constante) e, por último, uma série não estacionária em média nem em variância (média e variância crescentes).

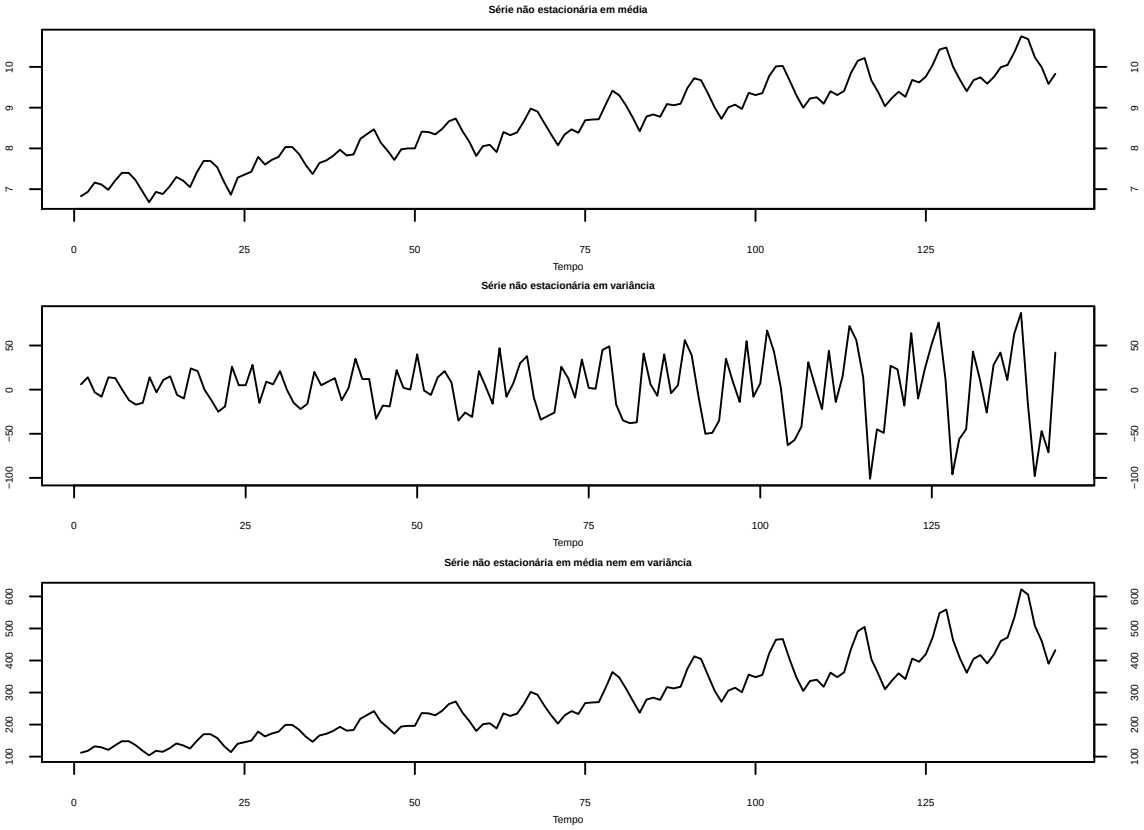


Figura 3.5: Exemplo de séries temporais não estacionárias.

Em casos de não estacionariedade, o procedimento usual parte por "remover" a tendência e a sazonalidade, seguindo os modelos de decomposição descritos na Secção 3.1. Uma vez tornada a série estacionária, os modelos estimam as componentes tendência (T) e sazonalidade (S) através de funções determinísticas (ou outro tipo de metodologias), de forma a que a série consiga ser modelada por um processo estacionário. Contudo, também é possível efetuar outras tranformações para implementar estacionariedade em séries não estacionárias.

Transformações para estabelecer estacionariedade

Um método bastante recorrente para estabilizar a média de uma série temporal consiste em efetuar diferenças (regulares), através do uso do operador atraso ∇, que consiste em criar uma

nova série correspondente aos incrementos da série original não estacionária, i.e., $\nabla X_t = X_t - X_{t-1},\ t = 2, 3, \dots, \underset{t \in T_0}{max}\{t\}$.

Se a série permanecer não estacionária, mesmo após efetuada uma diferenciação regular, pode-se, novamente, diferenciar a série, obtendo-se as diferenças de 2.ª ordem, i.e.,

$$\begin{aligned} \nabla^2 X_t = \nabla(\nabla X_t) &= \nabla(X_t - X_{t-1}) \\ &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= X_t - 2X_{t-1} + X_{t-2} \qquad , \quad t = 3, 4, \dots, \underset{t \in T_0}{max}\{t\}. \end{aligned}$$

O operador atraso de ordem d, para $d \in \mathbb{N}$ consiste portanto na aplicação deste operador d vezes, obtendo-se

$$\nabla^d X_t = \nabla(\nabla^{d-1} X_t),\ t = d+1, d+2, \dots, \underset{t \in T_0}{max}\{t\}.$$

Nota 2. Cada vez que o operador atraso é aplicado, a série resultante representa valores diferentes da série original. Por este motivo, não é recomendada a aplicação exaustiva deste operador, uma vez que a interpretação da série fica cada vez mais complexa à medida que é diferenciada regularmente, pois a variância aumenta após cada diferenciação.

Obtendo uma série estacionária, deve evitar-se esta diferenciação. Na verdade, o objetivo passa por obter estacionariedade com o menor número de diferenciações possível. Num panorama geral, se a série $\nabla^{d_0} X_t$ é estacionária, então todas as séries $\nabla^{d_1} X_t,\ d_1 > d_0$ também o são. Na prática, este operador costuma impor estacionariedade logo após a primeira ou a segunda diferenciação.

Se uma série não estacionária se tornar estacionária após d diferenciações regulares, então diz-se que esta é uma série integrável de ordem d e representa-se por I(d). Por convenção, um processo é estacionário se e só se for um processo integrável de ordem 0, ou seja, se for um processo I(0).

Passeio aleatório

Considere-se o modelo

$$X_t = X_{t-1} + \epsilon_t \tag{3.6}$$

com tendência estocástica, onde ϵ_t é um ruído branco. Este modelo é conhecido como modelo de passeio aleatório (random walk), pois o valor da série no instante t é descrito apenas pelo valor do instante anterior, $t-1$ e por um choque aleatório. O comportamento usual de um passeio aleatório descreve-se à custa de movimentos de tendência (crescente ou decrescente) em períodos

longos, seguidos de mudanças repentinas a favor ou contra o sentido da tendência [Caiado, 2011]. O valor médio de um processo aleatório é constante, uma vez que a tendência positiva é apenas a existência de uma maioria de valores positivos em relação a uma minoria de valores negativos e o recíproco para a tendência decrescente. Caso seja conhecido o estado inicial de um passeio aleatório, Y_0, o modelo (3.6) pode ser reescrito como

$$\begin{aligned} X_t &= \overbrace{(X_{t-2} + \epsilon_{t-1})}^{X_{t-1}} + \epsilon_t \\ X_t &= \underbrace{(X_{t-3} + \epsilon_{t-2})}_{X_{t-2}} + \epsilon_{t-1} + \epsilon_t \\ &\vdots \\ X_t &= X_0 + \sum_{i=1}^{t} \epsilon_i, \end{aligned} \tag{3.7}$$

onde se verifica que $E(X_t) = X_0$, ou seja, o valor esperado de X_t é constante. No entanto, este modelo é um processo não estacionário, pois a sua variância depende do instante de tempo t [Enders, 2015]. Contudo, aplicando o operador atraso uma vez, obtém-se

$$\nabla X_t = X_0 + \sum_{i=1}^{t} \epsilon_i - X_0 - \sum_{i=1}^{t-1} \epsilon_i = \epsilon_t,$$

que corresponde a um processo de ruído branco, que, por sua vez, é estacionário.

Este tipo de modelos é aplicado muitas vezes na modelação de séries financeiras não estacionárias, tais como, as séries dos preços das ações, cuja melhor previsão para um valor num dia t corresponde ao respetivo valor no dia $t-1$ [Caiado, 2011].

A Figura 3.6 mostra uma simulação de um passeio aleatório e a série correspondente à sua diferenciação de primeira ordem. O correlograma de um passeio aleatório apresenta valores positivos com um decaimento lento para zero.

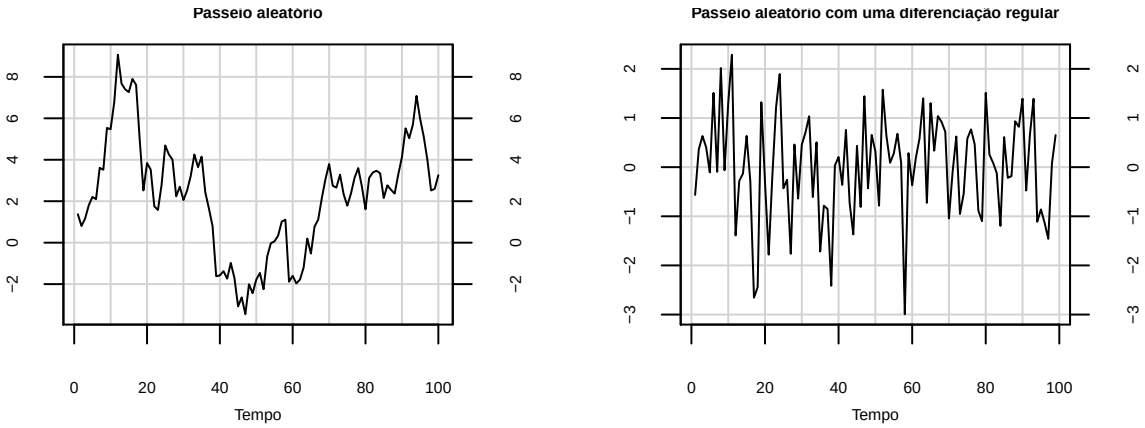


Figura 3.6: Simulação de um passeio aleatório e a série correspondente à sua diferenciação de 1.ª ordem.

Passeio aleatório com drift

Ao adicionar um termo constante, diga-se a_0, ao modelo anterior, obtém-se a expressão de um modelo de passeio aleatório com drift,

$$X_t = a_0 + X_{t-1} + \epsilon_t, \tag{3.8}$$

onde ϵ_t continua a ser um ruído branco. Ao contrário do passeio aleatório sem drift, o valor médio deste processo não é constante, pois depende do instante de tempo t. Isto porque, neste caso, a tendência é considerada parcialmente determinística e parcialmente estocástica. Conhecendo o valor inicial X_0, pode reescrever-se o modelo (3.8) como

$$\begin{aligned} X_t &= a_0 + \overbrace{(a_0 + X_{t-2} + \epsilon_{t-1})}^{X_{t-1}} + \epsilon_t \\ X_t &= 2a_0 + \underbrace{(a_0 + X_{t-3} + \epsilon_{t-2})}_{X_{t-2}} + \epsilon_{t-1} + \epsilon_t \\ &\vdots \\ X_t &= X_0 + a_0 t + \sum_{i=1}^{t} \epsilon_i, \end{aligned} \tag{3.9}$$

de onde vem que o valor esperado, $E(X_t) = X_0 + a_0 t$, não é constante. Por este motivo, este processo não é estacionário, mas é integrável de 1.ª ordem.

A Figura 3.7 mostra os gráficos de um passeio aleatório com drift ao longo do tempo e da respetiva diferenciação regular de 1.ª ordem, admitindo $a_0 = 1$.

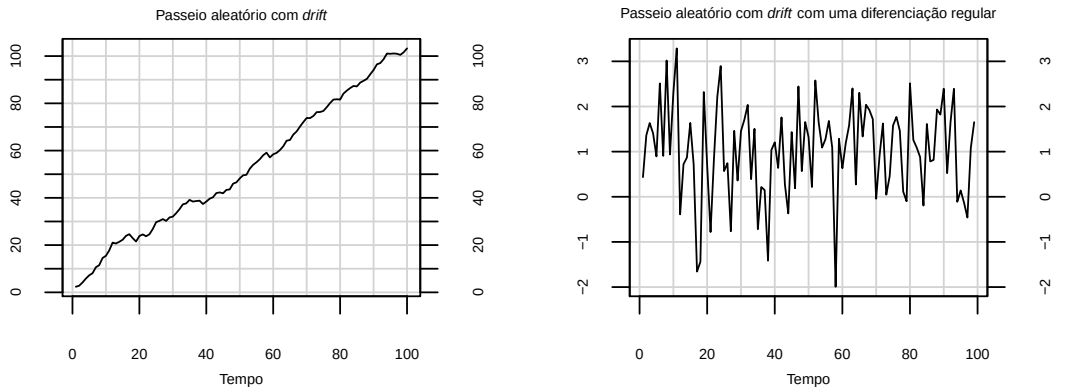


Figura 3.7: Simulação de um passeio aleatório com drift e a série correspondente à sua diferenciação de 1.ª ordem.

Como é visível no primeiro gráfico da Figura 3.7, a tendência determinística é dominante neste modelo. Porém, esta pode não ser tão visível com o aumento de $\sigma^2_{\epsilon_t}$ ou com a diminuição do valor absoluto de a_0 [Enders, 2015].

Análise de estacionariedade

Uma maneira empírica de avaliar a estacionariedade de uma série é representando o seu gráfico em função do tempo. No entanto, este é um critério bastante subjetivo, logo necessita de apoio estatístico. Para tal, existem vários testes de hipóteses de estacionariedade, sendo que a maioria consiste em encontrar uma raiz unitária. Segundo alguns autores, devem utilizar-se vários testes de estacionariedade para se poder inferir algo a seu respeito. Alguns dos testes mais conhecidos são: o teste de Dickey-Fuller (DF), o teste de Dickey-Fuller Aumentado (Augmented Dickey-Fuller, ADF), o teste de Phillips-Perron (PP) e o teste de Kwiatkowski-Phillips-Schmidt-Shin (KPSS). Estes testes dividem-se em duas abordagens: a procura de uma raiz unitária (i.e., a não estacionariedade) e na procura de estacionariedade (KPSS). No primeiro caso, se for encontrada uma raiz unitária, é fornecida a ordem de integração da série, i.e., o número de diferenciações regulares necessárias para impor estacionariedade (DF, ADF e PP). Encontra-se mais informação exposta sobre este tema em [Dickey and Fuller, 1979, Said and Dickey, 1984, Phillips and Perron, 1988, Kwiatkowski et al., 1992].

Teste de Dickey-Fuller

Considere-se o seguinte processo

$$X_t = \phi X_{t-1} + \epsilon_t, \tag{3.10}$$

onde $|\phi| \leqslant 1$ e ϵ_t é um ruído branco. Este processo é estacionário se $|\phi| < 1$ (tal como será explicado na Secção 4.1.1). Caso $|\phi| = 1$, trata-se de um processo de passeio aleatório, que, tal como foi visto, não é estacionário.

O processo (3.10) pode ser escrito à custa do operador atraso, obtendo-se

$$\begin{aligned} X_t = \phi X_{t-1} + \epsilon_t &\iff X_t - X_{t-1} = \phi X_{t-1} + \epsilon_t - X_{t-1} && (3.11)\\ &\iff \nabla X_t = X_{t-1}(\phi - 1) + \epsilon_t && (3.12)\\ &\iff \nabla X_t = \delta X_{t-1} + \epsilon_t, && (3.13) \end{aligned}$$

onde $\delta = \phi - 1$ e ϵ_t é um ruído branco, logo um processo estacionário. Então, o teste DF postula-se, em função de δ, como

$$H_0 : \delta = 0 \quad vs \quad H_1 : -2 < \delta < 0.$$

Note-se que, se $\delta = 0$ então X_t trata-se de um passeio aleatório, logo $X_t \sim I(1)$ é não estacionário. Caso H_1 se verifique, então X_t é estacionário. Dickey e Fuller [Dickey and Fuller, 1979] propõem a alteração

$$H_0 : \delta = 0 \quad vs \quad H_1 : \delta < 0,$$

mantendo-se a rejeição de estacionariedade caso H_0 não seja rejeitada. Caso se verifique H_1, então $\phi - 1 < 0 \iff \phi < 1$, o que indica que X_t é estacionário, sabendo as condições de estacionariedade deste processo.

Estes autores, consideram ainda duas outras equações, além da (3.13):

$$\nabla X_t = a_0 + \delta X_{t-1} + \epsilon_t, \tag{3.14}$$

$$\nabla X_t = a_0 + a_1 t + \delta X_{t-1} + \epsilon_t\ . \tag{3.15}$$

As três equações (3.13), (3.14) e (3.15) diferem na constante a_0 e/ou no termo determinístico $a_1 t$. Desta forma, ao considerar a equação (3.13), conclui-se que X_t é estacionário de média não nula se se rejeitar H_0. Considerando a equação (3.14), a rejeição da hipótese nula indica que X_t é estacionário com uma tendência determinística.

É essencial que se verifique a ausência de autocorrelações na série de ruído branco ϵ_t, pois, caso isto não aconteça, esta série pode exibir comportamentos determinísticos. Nestes casos,

deve ser aplicado o teste ADF [Cordeiro, 2011].

Teste de Dickey-Fuller Aumentado

Raramente se encontram séries com um grau de complexidade como o da série (3.10). O teste ADF é mais apropriado para estas situações, onde a série apresenta uma estrutra mais complexa. Este teste é uma generalização do teste DF criada por Said e Dickey [Said and Dickey, 1984] e tem tido bastante utilidade no estudo de estacionariedade de séries temporais.

Considere-se, portanto, o processo definido por

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t \ .$$

Aplicando o operador atraso, vem

$$\nabla X_t = \delta X_{t-1} + \sum_{i=1}^{p-1} \gamma_i \nabla X_{t-i} + \epsilon_t, \tag{3.16}$$

onde $\delta = \sum_{j=1}^{p} \phi_j - 1$, $\gamma_i = -\sum_{j=i+1}^{p} \phi_j$, $\nabla X_{t-i} = X_{t-i} - X_{t-i-1}$ e ϵ_t é um ruído branco. Esta diferenciação separa o modelo (3.16) em duas componentes: X_{t-1} e $\sum_{i=1}^{p-1} \gamma_i \nabla X_{t-i}$. Se X_t se tratar de um passeio aleatório, esta diferenciação separa o modelo numa componente não estacionária e em $p-1$ componentes estacionárias. Na notação usual, diz-se que foram aumentadas $p-1$ componentes ao modelo original, originando o termo ADF($p-1$).

O teste ADF estabelece, portanto, as hipóteses

$$H_0 : \delta = 0 \quad vs \quad H_1 : \delta < 0$$

onde H_0 significa que $\sum_{j=1}^{p} \phi_j = 1$, i.e., o modelo assume uma raiz unitária e H_1 implica a ausência de estacionariedade de X_t.

O valor crítico deste teste corresponde ao quantil (1-α)100% da respetiva distribuição, uma vez que é um teste unilateral à esquerda, ou seja, rejeita-se H_0 se o valor da estatística de teste for inferior ou igual a este quantil.

Tal como no teste DF, são propostos outros dois modelos:

$$\nabla X_t = a_0 + \delta X_{t-1} + \sum_{i=1}^{p-1} \gamma_i \nabla X_{t-i} + \epsilon_t \tag{3.17}$$

e

$$\nabla X_t = a_0 + a_1 t + \delta X_{t-1} + \sum_{i=1}^{p-1} \gamma_i \nabla X_{t-i} + \epsilon_t \ . \tag{3.18}$$

Estes diferem, mais uma vez, na constante a_0 e/ou na componente determinística $a_1 t$.

Contudo, surge a questão sobre quantos termos devem ser incluidos nestas equações, i.e., o valor de p. Para tal, existem várias abordagens:

1. Testar vários valores de p, até que os resíduos do modelo não apresentem qualquer estrutura de correlação;

2. Utilizar um critério de informação, e.g., o critério de informação de Akaike (Akaike Information Criterion, AIC);

3. Recorrer ao método proposto por Ng e Perron [Ng and Perron, 1995]. Em primeiro lugar, estabelecer um valor máximo para p, p_{max}. De seguida efetuar o teste ADF com $p = p_{max}$ e verificar o valor da estatística t. Se este valor for, em valor absoluto, superior a 1,6, então define-se $p = p_{max}$. Caso contrário, define-se p como o maior valor possível menor que p_{max}, tal que o valor da estatística t seja, superior, em módulo, a 1,6. Para estabelecer o valor de p_{max}, pode usar-se o método de Schwert [Schwert, 2002],

$$p_{max} = \left[12 \left(\frac{T}{100} \right)^{\frac{1}{4}} \right]$$

onde $[x]$ representa a parte inteira de x, i.e., o chão de x e T o número de observações.

Mesmo tendo um vasto leque de opções, é sempre necessário verificar se os resíduos do modelo se comportam, de facto, como um ruído branco.

Teste de Phillips-Perron

À semelhança do teste ADF, o teste de Phillips-Perron (PP) testa a estacionariedade de um processo utilizando as equações (3.16), (3.17) e (3.18) e as hipóteses $H_0 : \delta = 0$ e $H_1 : \delta < 0$. Este difere apenas na distribuição que os erros ϵ_t podem assumir, pois é permitido que estes sejam correlacionados e/ou heterocedásticos. Logo, rejeitando a hipótese nula, pode inferir-se que o processo é estacionária.

Teste de Kwiatkowski-Phillips-Schmidt-Shin

O teste de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) serve, também, para testar a estacionariedade de um processo. Ao contrário dos testes anteriores, este assume estacionariedade sob validade da hipótese nula, i.e.,

$$H_0 : \text{O processo é estacionário} \quad vs \quad \text{O processo não é estacionário .}$$

A equação deste teste divide um dado processo X_t em três componentes: T_t, uma tendência determinística, μ_t, um passeio aleatório e u_t, um erro estacionário. Ou seja,

$$X_t = T_t + \mu_t + u_t$$

onde $\mu_t = \mu_{t-1} + \epsilon_t$ e ϵ_t é um ruído branco.

Este teste é unilateral à direita, logo a hipótese nula de estacionariedade é rejeitada se o valor da estatística de teste for superior ou igual ao quantil (1-α)100% da respetiva distribuição.

Capítulo 4

Metodologias de previsão

Os modelos matemáticos servem para descrever processos geradores da natureza. Mais especificamente, os modelos estatísticos podem ser utilizados para calcular a distribuição de probabilidade de uma certa variável aleatória para um instante futuro. Estes modelos regem-se por leis de probabilidade, uma vez que são definidos por processos estocásticos. Para além de estimar valores observados e efetuar previsões pontuais, estes modelos estão aptos para calcular intervalos de previsão com um certo nível de significância.

Um método de previsão é uma ferramenta que, com base em valores observados no passado, prevê valores futuros. Assim, este pode ser um simples algoritmo de fácil aplicabilidade e sem bases estatísticas, como pode também ser um modelo construído com base em leis de probabilidade. Segundo Chatfield [Chatfield, 2003], existem três tipos de métodos de previsão: métodos subjetivos, que envolvem subjetividade com base em experiência empírica; métodos univariados, cujas previsões dependem apenas dos valores de apenas uma série temporal e métodos multivariados, cujas previsões dependem de várias variáveis.

Existe uma grande panóplia de métodos de previsão, onde cada um tem as suas vantagens e limitações. Isto torna a escolha destes métodos mais complicada, pois esta escolha nunca é efetiva, ou seja, nunca se deve descartar a ideia de abordar uma modelação de uma série temporal por diversos meios. Alguns métodos de previsão nasceram à custa da combinação de outros, com fim de se poder retirar ou diminuir as limitações. A escolha do método envolve muitas variáveis, tais como:

- o objetivo do estudo;
- a quantidade de dados adquirida;
- o conhecimento do analista na área;
- o horizonte de previsão;

- o tipo de série temporal (presença/ausência de valores negativos e/ou nulos, sazonalidades complexas, etc.);

- a disponibilidade financeira/geográfica/computacional.

Neste Capítulo, são discutidas três metodologias de previsão distintas: a teoria clássica de Box-Jenkins (modelos SARIMA), os modelos de alisamento exponencial de Holt-Winters (metodologia não paramétrica) e os modelos TBATS.

4.1 Metodologia Box-Jenkins

Com base no trabalho de Yule [Yule, 1926] e Wold [Wold, 1938], Box e Jenkins [Box et al., 2016] desenvolveram um conjunto de ferramentas práticas para abordar séries temporais, os conhecidos modelos SARIMA. Esta metodologia é constituída por três etapas principais: a identificação do modelo, a estimação dos seus parâmetros e a análise de diagnóstico (validação do modelo, geralmente feita através de uma análise aos resíduos). Este é um processo iterativo, pelo que é tipicamente aplicado várias vezes até se obter um modelo com um ajuste favorável.

A base desta abordagem consiste em assumir que os valores de uma série temporal apresentam autocorrelações entre si, sendo então possível prever valores futuros, à custa de valores anteriores. Assim, comparando comportamentos teóricos de autocorrelações, é possível identificar alguns modelos com potencial para atingir o objetivo da modelação. Consequentemente, foi proposto o uso da função de autocorrelação (FAC) e da função de autocorrelação parcial (FACP) para construir os modelos SARIMA [Zhang, 2003].

4.1.1 Processos autorregressivos (AR)

Quando se tem uma correlação significativa entre dois valores distintos X_{t_0} e X_{t_1}, de um certo processo $\{X_t\}_{t\in T_0}$ estacionário, significa que se trata de um processo autorregressivo (AR).

Definição 8. Um processo autorregressivo de ordem p, AR(p), é um processo que pode ser escrito como

$$X_t = \delta + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t \tag{4.1}$$

onde $\epsilon_t \sim RB(0, \sigma_\epsilon^2)$, (ruído branco de média nula e variância σ_ϵ^2), $\delta \in \mathbb{R}$, $\phi_i \neq 0$, $\forall\, i = 1, 2, \dots, p$.

Sem perda de generalidade, considerando $\delta = 0$, a equação 4.1 pode ser escrita à custa do

operador atraso, definido como sendo $B^k X_t = X_{t-k}$, obtendo-se

$$\epsilon_t = X_t - \sum_{i=1}^{p} \phi_i X_{t-i} \iff \epsilon_t = X_t - \sum_{i=1}^{p} \phi_i B^i X_t \iff \epsilon_t = X_t \underbrace{\left(1 - \sum_{i=1}^{p} \phi_i B^i\right)}_{\text{polinómio autorregressivo}}.$$

Nota 3. Dado um polinómio $P(z) = \alpha_n z^n + \alpha_{n-1} z^{n-1} + \cdots + \alpha_1 z + \alpha_0,\ z \in \mathbb{C}$ de grau n, com m raízes distintas ($m \leqslant n$), $z_1,\ z_2,\ \ldots,\ z_m$, pode-se fatorizá-lo como:

$$\begin{aligned} P(z) = \alpha_n \prod_{i=1}^{n} (z - z_i)^{m_i} &= \alpha_n \prod_{i=1}^{n} (-z_i)^{m_i} \left(-\frac{z}{z_i} + 1\right)^{m_i} \\ &= \underbrace{\alpha_n \prod_{i=1}^{n} (-z_i)^{m_i}}_{\beta} \prod_{i=1}^{n} \left(1 - \frac{z}{z_i}\right)^{m_i} \\ &= \beta \prod_{i=1}^{n} (1 - \lambda_i z)^{m_i}, \end{aligned}$$

onde m_i é a multiplicidade da raiz z_i e $\lambda_i = \frac{1}{z_i}$.

Sem perda de generalidade, reescrevendo este processo à custa das suas raízes, vem $\Phi_p(B) = \prod_{i=1}^{p}(1 - G_i B)$. Este é estacionário se e só se as raízes, $G_1^{-1},\ G_2^{-1},\ \ldots,\ G_p^{-1},$, do seu polinómio autorregressivo $\Phi_p(B)$ forem todas superiores, em módulo, a 1, ou seja, se e só se $|G_i^{-1}| > 1 \implies |G_i| < 1,\ i = 1, 2, \ldots, p$ [Cowpertwait and Metcalfe, 2009]. Qualquer processo autorregressivo estacionário é também invertível, o que na prática significa que quanto maior a distância temporal entre duas observações, menor é a dependência entre si.

As autocorrelações num processo AR(p) vão diminuindo com o aumento do lag. Assim, a sua função de autocorrelação parcial, ϕ_{kk}, é nula para todo o $k > p$. Isto significa que o seu gráfico apresenta uma queda repentina para zero a partir do lag $p + 1$ e o gráfico da sua FAC mostra um decaimento exponencial para zero.

A Figura 4.1 mostra os gráficos das FAC e FACP empíricas associadas ao processo estacionário AR(1), $X_t = -0,9 X_{t-1} + \epsilon_t$. A FAC apresenta um decaimento exponencial para zero de acordo com a expressão $\rho(h) = -0,9^h$ e o gráfico da FACP mostra uma queda abrupta para zero a partir do $lag = 2$.

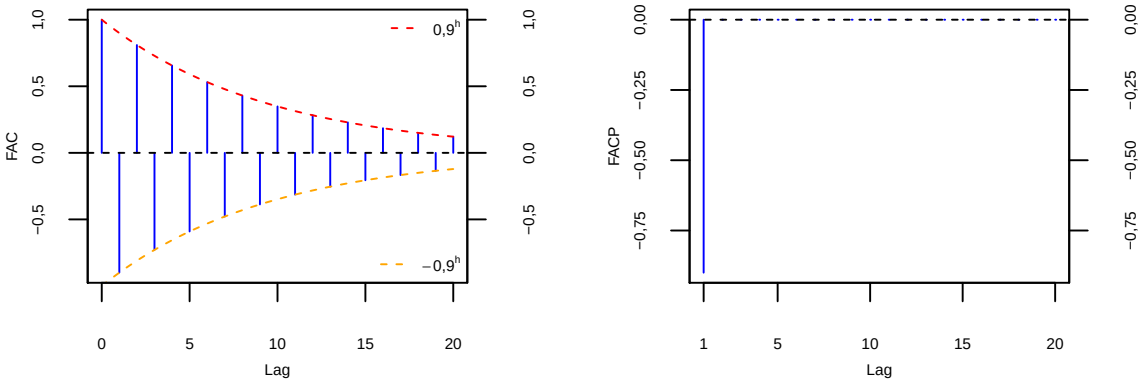


Figura 4.1: FAC e FACP de um processo autorregressivo de equação $X_t = -0,9X_{t-1} + \epsilon_t$.

4.1.2 Processos médias móveis (MA)

Definição 9. Seja ϵ_t um processo de ruído branco com média e variância constante. O processo $\{X_t\}_{t\in T_0}$ define-se como um processo de médias móveis de ordem q, MA(q), se

$$X_t = \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}. \tag{4.2}$$

A equação (4.2) pode ser escrita como $X_t = \Theta_q(B)\epsilon_t$, onde $\Theta_q(B) = 1 + \sum_{i=1}^{q} \theta_i B^i$ é o polinómio de médias móveis de ordem q a si associado. Este processo serve como que uma suavização para séries, uma vez que, para cada instante t, este processo calcula uma média ponderada de $q+1$ observações de um processo de ruído branco. Assim, todos os processos MA(q) são estacionários, pois qualquer ruído branco é estacionário e, consequentemente, qualquer combinação linear de ruídos brancos também o é.

Um processo MA(q) diz-se invertível se for possível escrevê-lo na forma de um processo autorregressivo estacionário de ordem infinita, i.e., AR(∞).

Nota 4. Um processo é estacionário se conseguir escrever-se na forma de um processo MA(∞). Considere-se, a título de exemplo, um processo estacionário AR(1), $X_t = \phi_1 X_{t-1} + \epsilon_t$. Utilizando o resultado de séries geométricas,

$$\frac{1}{1-x} = \sum_{i=0}^{+\infty} x^i, \ |x| < 1,$$

é possível escrever o processo $\{X_t\}_{t\in T_0}$ à custa do operador B como

$$X_t = \phi_1 B X_t + \epsilon_t \iff X_t - \phi_1 B X_t = \epsilon_t \iff X_t\,(1 - \phi_1 B) = \epsilon_t,$$

$$\therefore X_t = \frac{\epsilon_t}{1-\phi_1 B}\ .$$

Então,

$$X_t = \epsilon_t \times \frac{1}{1-\phi_1 B} = \epsilon_t \sum_{i=0}^{+\infty} (\phi_1 B)^i = \sum_{i=0}^{+\infty} \phi_1^i B^i \epsilon_t = \underbrace{\sum_{i=0}^{+\infty} \phi_1^i \epsilon_{t-i}}_{\text{processo MA}(\infty)}$$

onde ϵ_t é um ruído branco e $|\phi_1| < 1$.

Tal como nos processos AR(p), uma condição necessária e suficiente para que um processo MA(q) seja invertível é que as raízes do seu polinómio de médias móveis $\Theta_q(B)$ estejam todas fora do círculo unitário, i.e., que sejam todas, em módulo, maiores que 1 [Cowpertwait and Metcalfe, 2009].

As autocorrelações teóricas, ρ_k, num processo MA(q) são nulas para $k > q$, logo o gráfico da sua FAC mostra correlações significativamente superiores a zero até ao $lag = q$, registando-se uma queda abrupta para zero no $lag = q + 1$, diminuindo com o aumento do lag. Assim, a sua função de autocorrelação parcial, ϕ_{kk}, é nula para todo o $k > p$. Isto significa que o seu gráfico apresenta uma queda repentina para zero a partir do lag $p + 1$ e o gráfico da sua FACP mostra um decaimento exponencial para zero.

A Figura 4.2 mostra os gráficos das FAC e FACP empíricas associadas ao processo estacionário e invertível MA(1), $X_t = \epsilon_t - 0,9\epsilon_{t-1}$. A FACP apresenta um decaimento exponencial para zero e o gráfico da FAC mostra uma queda abrupta para zero a partir do $lag = 2$.

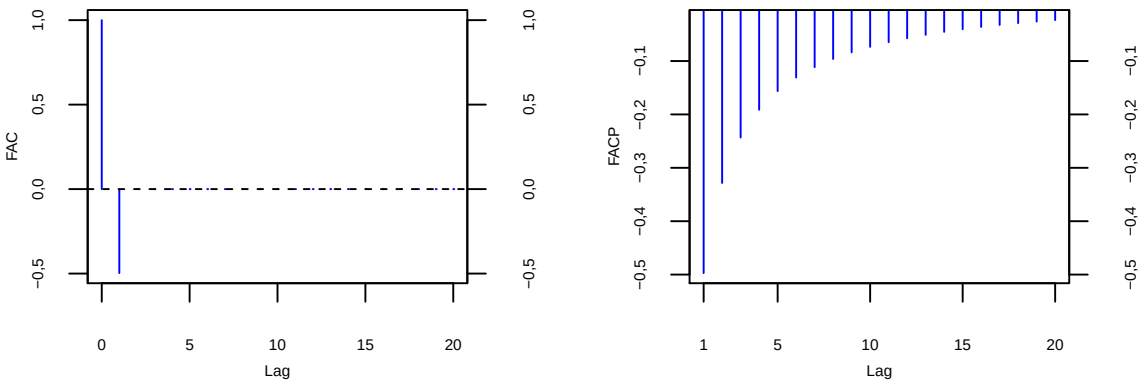


Figura 4.2: FAC e FACP de um processo de médias móveis de equação $X_t = \epsilon_t - 0,9\epsilon_{t-1}$.

4.1.3 Processos autorregressivos e de médias móveis (ARMA)

Como já foi referido, todos os processos estacionários admitem representações tanto autorregressivas como na forma de médias móveis. Porém estas representações podem resultar em processos com demasiados coeficientes, o que torna a sua estimação menos eficiente. Assim, pode-se escrever um processo mais parcimonioso contendo tanto coeficientes autorregressivos

como de médias móveis. Este trata-se de um processo autorregressivo e de média móveis de ordens p e q, ARMA(p,q).

Definição 10. Dado um processo $\{X_t\}_{t\in T_0}$, este diz-se autorregressivo e de médias móveis de ordem p e q, ARMA(p,q) se satisfizer a equação

$$X_t = \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_i \epsilon_{t-j}, \tag{4.3}$$

ou, utilizando o operador atraso B, a equação

$$\Phi_p(B) X_t = \Theta_q(B) \epsilon_t, \tag{4.4}$$

onde ϵ_t é um ruído branco com média nula, independente de X_{t-k} para $k \geqslant 1$, $\Phi_p(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ é o polinómio autorregressivo de ordem p e $\Theta_q(B) = 1 + \sum_{i=1}^{q} \theta_i B^i$ o polinómio de médias móveis de ordem q associados a este processo.

As funções de autocorrelação e autocorrelação parcial de um processo ARMA(p,q) são uma combinação das FAC e FACP de processos AR(p) e MA(q). A Figura 4.3 mostra as representações gráficas das FAC e FACP de um processo ARMA(2,2) estacionário $(1 - 0,5B + 0,3B^2)X_t = (1 - 0,7B + 0,1B^2)\epsilon_t$.

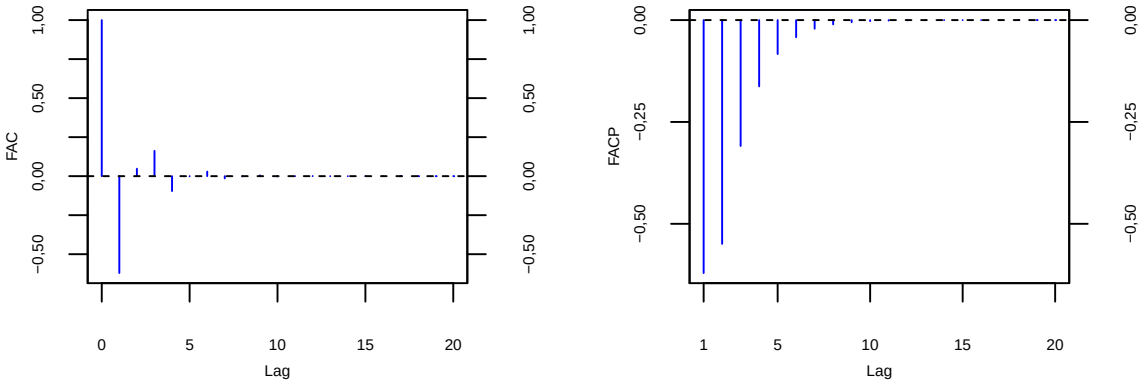


Figura 4.3: FAC e FACP de um processo autorregressivo e de médias móveis de equação $(1-0,5B+0,3B^2)X_t = (1 - 0,7B + 0,1B^2)\epsilon_t$.

As condições de estacionariedade de um processo ARMA(p,q) são equivalentes às anteriores, pois basta que as raízes dos polinómios $\Phi_p(B) = 0$ e $\Theta_q(B) = 0$ sejam, em módulo, superiores a 1.

4.1.4 Processos autorregressivos integrados e de médias móveis (ARIMA)

Na prática, a estacionariedade não está presente em muitas séries temporais. Como foi visto na Secção 3.2.4, processos não estacionários podem ser diferenciados para que assumam estacionariedade, adquirindo o termo de processos integrados. Ora, se um processo ARMA(p,q) necessitar de ser diferenciado d vezes até atingir estacionariedade, este denomina-se um processo autorregressivo de ordem p, integrado de ordem d e de médias móveis de ordem q, ou simplesmente ARIMA(p,d,q).

Definição 11. Dado um processo $\{X_t\}_{t\in T_0}$, este diz-se um processo autorregressivo de ordem p, integrado de ordem d e de médias móveis de ordem q, ARIMA(p,d,q), se admitir uma representação da forma

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)(1-B)^d X_t = \left(1+\sum_{j=1}^{q}\theta_j B^j\right)\epsilon_t, \tag{4.5}$$

ou simplesmente

$$\Phi_p(B)\nabla^d X_t = \Theta_q(B)\epsilon_t, \tag{4.6}$$

onde $\nabla^d X_t = (1-B)^d X_t$, com $d \geqslant 1$, é a série estacionária (depois de ter sido diferenciada regularmente d vezes), $\Phi_p(B)$ é o polinómio autorregressivo constituído pelos coeficientes autorregressivos $\phi_1,\ \phi_2,\ \ldots,\ \phi_p$ e $\Theta_q(B)$ é o polinómio de médias móveis constituído pelos coeficientes de médias móveis $\theta_1,\ \theta_2,\ \ldots,\ \theta_q$.

Sendo um processo não estacionário, um processo ARIMA apresenta correlações mais elevadas do que nos processos estudados anteriormente, fazendo com que o gráfico da sua FAC mostre um decrescimento mais lento para zero. A Figura 4.4 mostra as FAC e FACP de uma simulação de 100 observações de um processo ARIMA(2,1,1) de equação $(1+0,9B+0,7B^2)(1-B)X_t = (1+0,8B)\epsilon_t$.

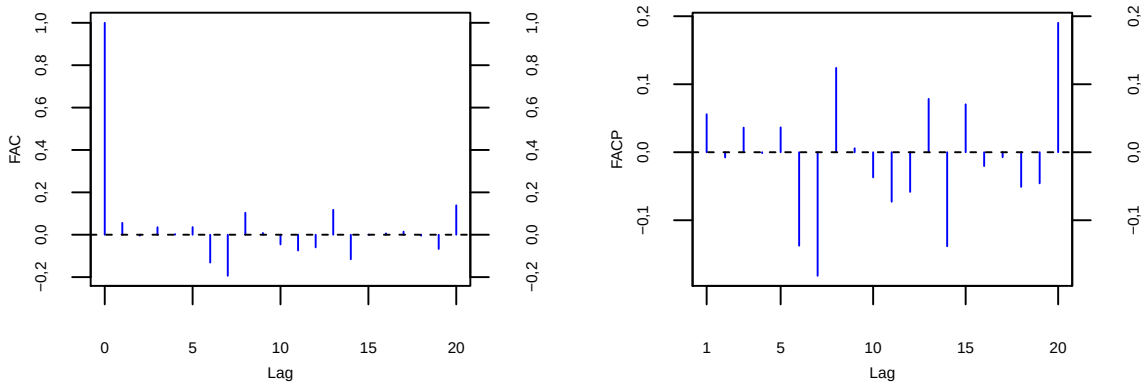


Figura 4.4: FAC e FACP empíricas de um processo autorregressivo integrado e de médias móveis de equação $(1+0,9B+0,7B^2)(1-B)X_t = (1+0,8B)\epsilon_t$ simulado.

4.1.5 Processos sazonais autorregressivos integrados e de médias móveis (SARIMA)

Para além de comportamentos autorregressivos e de médias móveis, muitas séries exibem padrões sazonais. Os modelos anteriores estão incompletos no sentido em que não incorporam esta componente. Geralmente, são adotados dois modelos ARIMA: um para a componente sistemática e outro para a componente sazonal de séries temporais. Isto significa que as componentes sazonais podem não ser estacionárias, o que, usualmente, implica que estas necessitem de uma diferenciação sazonal para se impor estacionariedade. Uma diferenciação sazonal é uma simples diferença entre uma observação do instante t e outra do instante $t-s$. Ou seja, quando uma série apresenta características sazonais, pode-se efetuar uma diferenciação sazonal, i.e.,

$$\nabla_s X_t = X_t - X_{t-s} = (1 - B^s)X_t.$$

A série resultante representa a mudança entre observações distanciadas s unidades de tempo. Tal como as diferenciações regulares, a diferenciação sazonal pode ser aplicada d vezes até se obter estacionariedade sazonal, com $d \geqslant 1$. Assim, surge o operador de diferenciação sazonal de ordem D, que se representa como

$$\nabla_s^D X_t = (1 - B^s)^D X_t.$$

Definição 12. Um processo $\{X_t\}_{t\in T_0}$ diz-se um processo autorregressivo integrado e de médias móveis sazonal, ou simplesmente SARIMA$(p,d,q)(P,D,Q)_s$ se satisfizer a equação

$$\Phi_p(B)N_p(B^s)\nabla^d\nabla_s^D X_t = \Theta_q(B)H_Q(B^s)\epsilon_t, \tag{4.7}$$

onde $\Phi_p(B)$, $N_p(B^s)$, $\Theta_q(B)$ e $H_Q(B^s)$ são os polinómios já referidos, d e D são as ordens das diferenciações regular e sazonal, respetivamente.

Uma diferenciação sazonal mostra-se útil quando a FAC de um processo mostra oscilações periódicas, i.e., quando mostra decrescimentos de s em s períodos de lag, seguidos de um crescimento [Shumway et al., 2000]. A Figura 4.5 mostra as FAC e FACP de uma simulação de 100 observações de um processo SARIMA$(2,1,1)(1,1,1)_7$ de equação $(1-1,2B+0,8B^2)(1-0,3B^7)(1-B)(1-B^7)X_t = (1+0,4B)(1+0,7B^7)\epsilon_t$. Nesta figura, nota-se uma clara sazonalidade de periodicidade 7, tanto na FAC como na FACP.

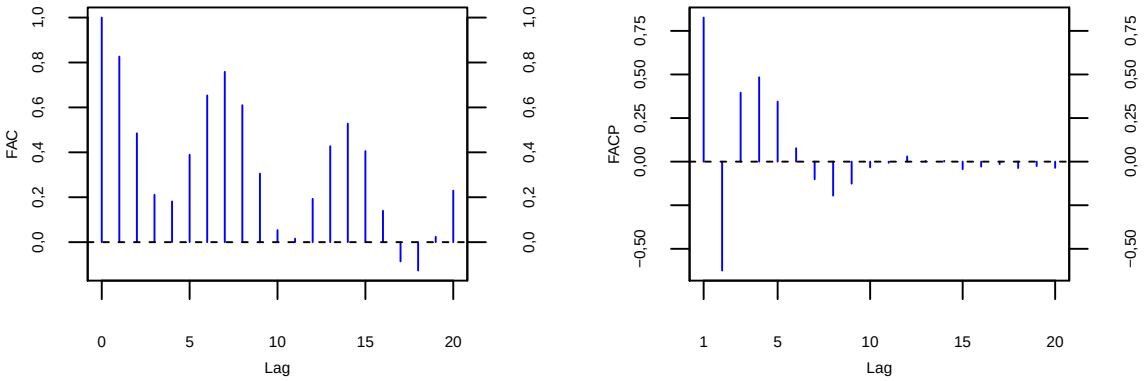


Figura 4.5: FAC e FACP empíricas de um processo autorregressivo integrado sazonal e de médias móveis de equação $(1-1,2B+0,8B^2)(1-0,3B^7)(1-B)(1-B^7)X_t = (1+0,4B)(1+0,7B^7)\epsilon_t$ simulado.

4.1.6 Estimação de modelos SARIMA

A construção do modelo completo da teoria de Box-Jenkins, SARIMA$(p,d,q)(P,D,Q)_s$ passa por três etapas principais já referidas anteriormente: a identificação, a estimação e a análise de diagnóstico.

Identificação do modelo SARIMA$(p,d,q)(P,D,Q)_s$

O primeiro passo consiste em representar graficamente a série e tentar retirar o máximo de informação empírica possível, por exemplo, visualizar quaisquer sazonalidades, tendências ou variância inconstante. Caso a série apresenta alguma destas características, deve-se, em primeiro lugar, retirar estes efeitos de modo a poder analisar-se o comportamento estacionário da série. O segundo objetivo na construção de um modelo SARIMA é analisar as FAC e FACP empíricas da série, tentando procurar reforço para as conclusões tiradas da análise gráfica do gráfico da série. Por exemplo, se uma série exibir uma forte sazonalidade de período s, então a sua FAC deve apresentar decaimentos periódicos de $lag = s$. A Tabela A.1 resume os comportamentos teóricos de cada tipo de modelo da teoria de Box-Jenkins. Depois de se identificarem os parâmetros p, d, q, P, D, Q e s do modelo, recorre-se às suas estimações.

Estimação dos parâmetros

Após fixada a escolha do modelo e dos parâmetros a estimar, o processo de modelação envolve a estimação dos mesmos. Os principais métodos de estimação de parâmetros são o dos mínimos quadrados e o da máxima verosimilhança. Estes são métodos númericos que, por vezes, apresentam um grau de complexidade elevado e, por este motivo, é aconselhado o uso de processamento computacional.

O método da máxima verosimilhança consiste em descobrir os valores dos parâmetros, tais que tornem mais verosímil a obtenção dos valores deveras observados. Este é um processo iterativo que maximiza a função de verosimilhança dos estimadores [Box et al., 2016].

O método dos mínimos quadrados passa por estimar os parâmetros, tais que minimizem o quadrado dos resíduos entre o modelo e os dados observados. Este é o método mais comum neste âmbito, porém, no caso dos modelos ARMA (com excessão de modelos AR), este não fornece estimadores centrados, i.e., estimadores cujo valor esperado corresponde ao parâmetro a estimar.

Análise de diagnóstico

Depois de identificado e estimado o modelo SARIMA, é necessário avaliar as estimativas dos parâmetros e a qualidade do ajuste do modelo aos dados observados. Para avaliar as estimativas dos parâmetros, recorre-se, usualmente, ao teste t de Student, onde se testam as hipóteses H_0 : $\beta_i = 0$ e $H_1 : \beta_i \neq 0$ para todo o $i = 1, 2, \dots, m$, onde m é o número de parâmetros estimados, β_i são os parâmetros estimados e $\widehat{\beta_i} \sim N(\beta_i, \sigma^2_{\beta_i})$. A hipótese nula deste teste é rejeitada se o valor da estatística de teste, T, for, em módulo, superior ou igual ao quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição t-Student($n - m$) onde n representa o número de observações, ou seja, se

$$|T| = \left| \frac{\widehat{\beta}_i}{\widehat{\sigma}_{\widehat{\beta}_i}} \right| \geqslant t_{1-\frac{\alpha}{2};n-m}.$$

Este teste tem bastante utilidade na prática, pois põe em consideração a parcimoniosidade do modelo, já que o objetivo é obter o modelo que melhor explique o comportamento dos dados, com a menor quantidade de parâmetros possível [Caiado, 2011].

Para além das estimações, é preciso verificar qual o comportamento dos resíduos do modelo. Estes devem ser uma realização de um processo de ruído branco de média nula e sem qualquer autocorrelação temporal. Caso algum destes dois critérios não se verifique, o modelo em causa deve ser rejeitado.

4.1.7 Análise de resíduos

Os resíduos de um modelo, e_t, são dados pela diferença entre os valores observados e os estimados, i.e., $e_t = X_t - \widehat{X}_t$. Tal como foi dito, estes devem ser representativos de um processo de ruído branco de média nula e sem autocorrelação temporal. Mais ainda, na construção de intervalos de previsão é conveniente que estes sigam uma distribuição, no mínimo, aproximadamente Normal com variância constante, pois estes são construídos com base em intervalos de confiança

de uma distribuição Gaussiana [Hyndman and Athanasopoulos, 2013].

Para verificar a condição de normalidade nos resíduos, deve ser efetuada uma análise gráfica e testes de hipóteses. A análise gráfica deve indicar que a distribuição dos resíduos deve ser, no mínimo, aproximadamente simétrica e a sua função de densidade de probabilidade empírica deve ser também, no mínimo, aproximadamente equivalente a uma curva de densidade Normal $N(0, \sigma^2_{e_t})$. Isto pode ser feito quer analisando o histograma dos resíduos, quer visualizando o QQ-plot, onde todos os pontos devem se posicionar ao longo de uma reta. Os testes de Shapiro-Wilk[1] e Kolmogorov-Smirnov (com correção de Lilliefors) são eficientes nesta análise, pois ambos testam as hipóteses

$$H_0 : \text{Os resíduos do modelo seguem uma distribuição Gaussiana}$$

$$vs$$

$$H_1 : \text{Os resíduos do modelo não seguem uma distribuição Gaussiana .}$$

Para verificar o pressuposto de autocorrelações não significativas, é geralmente utilizado um teste de Ljung-Box, para além da análise gráfica da FAC, onde não se devem observar correlações significativamente diferentes de zero. Para auxiliar estas inferências gráficas têm-se as bandas de Bartlett. No caso de um ruído branco Gaussiano, tem-se que $\hat{\rho}_i$ é independente de $\hat{\rho}_j$ para $i \neq j$ e, para n grande o suficiente,

$$\begin{aligned} \hat{\rho}_h \dot{\sim} N\left(0, \frac{1}{\sqrt{n}}\right) &\iff \sqrt{n}\hat{\rho}_h \dot{\sim} N(0,1) \\ &\iff P\left(-z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}} < \hat{\rho}_h \leqslant z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}}\right) \approx 1-\alpha, \end{aligned}$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição $N(0,1)$. Assim, um teste assintótico para testar se os resíduos de um modelo provêm de um processo de ruído branco (com $(1-\alpha)100\%$ de confiança), é verificar se $(1-\alpha)100\%$ das autocorrelações empíricas se encontram dentro do intervalo

$$\left(-z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}},\ z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}}\right).$$

Outro teste, este mais comummente efetuado, é o de Ljung-Box. Este baseia-se no facto de, perante um ruído branco Gaussiano, tem-se

$$\sqrt{n}\hat{\rho}_h \dot{\sim} N(0,1) \implies \left(\sqrt{n}\hat{\rho}_h\right)^2 \dot{\sim} \chi^2_1 \implies \widetilde{Q}_m = n\sum_{i=1}^{m} \hat{\rho}_h^2 \dot{\sim} \chi^2_m \,.$$

[1] O teste de Shapiro-Wilk não é recomendado para amostras muito grandes, por exemplo, superiores a 50 observações.

Logo é possível formular as hipóteses assintóticas $H_0 : e_t \sim RB(0, \sigma^2_{e_t})$ e $H_1 : e_t \nsim RB(0, \sigma^2_{e_t})$ que equivale a testar as hipóteses $H_0 : h \geqslant m \implies \rho_h = 0$ e $H_1 : \exists\, h \mid h \leqslant m \implies \rho_h \neq 0$. Para tal, de modo a obter-se um resultado assintótico mais aproximado, geralmente recorre-se à estatística corrigida

$$Q_m = n(n+2) \sum_{i=1}^{m} \frac{\widehat{\rho}_i^2}{n-i} \dot{\sim} \chi^2_m$$

que mede as autocorrelações acumuladas até ao $lag = m$. Portanto a hipótese nula é rejeitada com um nível de significância α se $Q_m > \chi^2_{m;1-\alpha}$, onde $\chi^2_{m;1-\alpha}$ é o quantil de ordem $1-\alpha$ da distribuição χ^2_m. Não existe nenhum critério para a escolha do valor de m, pelo que o melhor procedimento é efetuar este teste para vários valores distintos de m.

Para testar a média nula dos erros do modelo, recorre-se ao teste básico t de Student. As hipóteses postuladas neste caso são

$$H_0 : \mu_{e_t} = 0 \quad vs \quad H_1 : \mu_{e_t} \neq 0,$$

onde $|T| = \left|\frac{\overline{e}\sqrt{n}}{s_e}\right|$, sendo $\overline{e}$ a média dos resíduos obtidos, n a dimensão da amostra e s_e o desvio padrão dos resíduos. Note-se que este teste deve ser aplicado apenas se forem verificadas as condições explicadas anteriormente de não correlação e normalidade dos resíduos.

A homocedasticidade pode ser verificada empiricamente através da análise gráfica dos resíduos, onde se deve verificar que estes estão contidos entre dois limites (inferior e superior).

4.1.8 Previsão com modelos SARIMA

Depois de ajustado um modelo SARIMA a uma série temporal, pode partir-se para as previsões tanto pontuais como intervalares. As previsões pontuais são calculadas diretamente da fórmula do modelo escolhido. Na verdade, para extrapolar um valor a h passos, i.e., prever um valor no instante $t+h$, basta calcular o valor esperado condicionado $E\left(X_{t+h} \mid X_1, X_2, \ldots, X_t\right)$. Hyndman e Athanasopoulos [Hyndman and Athanasopoulos, 2013] resumem este processo em três fases: primeiro, reescrever a equação do modelo de forma a isolar o termo de X_t, segundo, substituir o índice t por $t+h$, e finalmente, substituir as observações futuras pelas suas previsões, os erros futuros por zero e os erros passados pelos respetivos resíduos. Desta forma, nota-se que o processo de previsão é iterativo, pois para extrapolar valores a h passos, é necessário ter todas as extrapolações até esse ponto. Assim, começa-se por prever o valor de X_{t+1}, seguido do valor de X_{t+2}, até se chegar ao valor de X_{t+h}.

Os intervalos de previsão são construídos com base em intervalos de confiança de uma distribuição Normal, já que os resíduos devem ser independentes entre si e apresentar um comportamento

Gaussiano. Assim, o intervalo de previsão para o instante $t+h$ é dado pela expressão

$$\left(\widehat{x}_{t+h|t} - z_{1-\frac{\alpha}{2}}\widehat{\sigma}_h, \widehat{x}_{t+h|t} + z_{1-\frac{\alpha}{2}}\widehat{\sigma}_h\right),$$

onde $z_{1-\frac{\alpha}{2}}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição N(0,1) e $\widehat{\sigma}_h$ é a estimativa do desvio padrão da previsão para o passo h.

Geralmente a amplitude destes intervalos aumenta proporcionalmente com o horizonte de previsão h. Porém, em modelos estacionários, as sucessões dos limites (inferior e superior) são convergentes, originando amplitudes idênticas independentemente do horizonte de previsão [Hyndman and Athanasopoulos, 2013].

4.2 Seleção de modelos

Em análises estatística, em particular em análises de séries temporais, existem vários modelos que se mostram adequados a descrever o comportamento dos dados e respeitar todas as regulamentações da análise de diagnóstico em simultâneo. Daqui surge a ambiguidade de escolha, que é resolvida com o apoio estatístico dos critérios de informação AIC (Akaike Information Criterion) e BIC (Bayesian Information Criterion).

O erro quadrático médio pode ser um critério de escolha já que escolhe o modelo com menores desvios entre valores estimados e observados. Contudo este critério não é eficaz em casos onde este apresenta um baixo valor e a função de verosimilhança conclui que o modelo é verosimil, simplesmente pelo aumento da quantidade de parâmetros do modelo.

Para evitar esta lacuna, a função de verosimilhança deve sofrer penalizações sempre que é acrescentado um novo parâmetro ao modelo, ou seja, deve-se aceitar um novo parâmetro apenas se o aumento que este causa na função de verosimilhança for superior à penalização causada pelo mesmo [Wheelwright et al., 1998].

Critério de informação de Akaike

Considere-se que um modelo com um total de m parâmetros foi ajustado a uma série de dimensão n. Akaike [Akaike, 1974] introduziu um critério para avaliar a qualidade do ajustamento de um modelo com base na quantidade de informação, expresso pela expressão

$$AIC = -2\ ln\ L + 2m,$$

onde L é a função de verosimilhança.

Nem sempre é possível determinar o valor exato deste critério porque a função de verosimi-

lhança pode ser difícil de descobrir. No entanto, é possível recorrer à aproximação $-2\ ln\ L \approx n(1 + ln\ 2\pi) + n\ ln\ \sigma^2$, de onde se deduz um valor aproximado para o AIC, da forma

$$AIC \approx n(1 + ln\ 2\pi) + n\ ln\ \hat{\sigma}^2 + 2m.$$

O AIC não fornece muita informação sobre o modelo, mas é capaz de quantificar as suas verosimilhança e parcimoniosidade. Por isso, este serve apenas para ser comparado com outros valores do AIC de outros modelos. Deve escolher-se o modelo que apresente o menor valor do AIC, tendo em conta que diferenças de até 3 ou 4 unidades[2] não são significativas. Nestas ocasiões, deve-se escolher o modelo em função da sua parcimoniosidade.

Critério de informação Bayesiano

O critério de informação Bayesiano (Bayesian Information Criterion, BIC) define-se pela expressão

$$BIC = -2\ ln\ L + m\ ln\ n,$$

onde, novamente, L é a função de verosimilhança, m a quantidade de parâmetros assumida pelo modelo e n a dimensão da série [Schwarz, 1978].

O BIC difere do AIC pelo valor de $ln\ n$, que, quanto maior for a dimensão da série, maior é a penalização do critério. Resolvendo a inequação $ln\ n > 2$ deduz-se que a penalização do BIC é superior à do AIC quando a série possui mais de 7 observações. Consequentemente, o BIC tende a selecionar modelos com menos parâmetros do que os selecionados pelo AIC, evitando, assim, a sobrestimação da quantidade de componentes.

4.3 Metodologias de alisamento exponencial

Os métodos de alisamento exponencial apareceram no final dos anos 50, nos trabalhos de Brown [Brown, 1960], Holt [Holt, 1957] e Winters [Winters, 1960] e proporcionam ferramentas capazes de se adaptarem às alterações dos dados, i.e., são capazes de captar mudanças no nível, no declive e no padrão sazonal. Tal como os modelos SARIMA, os modelos de alisamento exponencial efetuam previsões com base numa combinação ponderada de observações passadas, sendo que observações mais recentes possuem maior peso do que as mais antigas. Curiosamente, é daqui que surge o termo exponencial, pois as ponderações vão diminuindo o seu peso exponencialmente, quanto mais antiga for a observação.

É importante saber fazer a distinção entre modelos de alisamento exponencial e modelos de

[2] Estes valores são subjetivos.

espaço de estados. Um modelo de alisamento exponencial resume-se a uma ferramenta que calcula apenas previsões pontuais e não intervalares. O respetivo modelo de espaço de estados, para além das previsões pontuais, também consegue fornecer intervalos de confiança.

Um modelo de espaço de estados assume que um sistema pode ser determinado por um processo de vetores não observados $\theta_1, \theta_2, \ldots, \theta_n$ (estados) possivelmente associados a uma série $X_1, X_2, \ldots, X_n$. O modelo linear Gaussiano de espaço de estados é constituído pela equação de observação, $X_t = W_t\theta_t + e_t$, e pela equação de estado, $\theta_t = \Phi\theta_{t-1} + \theta_t$, onde W_t é uma matriz conhecida de dimensão $p \times m$, e_t é o vetor dos erros i. i. d. de dimensão $p \times 1$ com distribuição Normal de média nula e matriz de covariância H, i.e. $e_t \sim N(0, H)$ e $E(e_t e_s') = 0$ para $t = 1, 2, \ldots, n$ e $t \neq s$.

Estes métodos possuem algumas vantagens e desvantagens em relação a outros. A sua simplicidade de utilização, a facilidade de implementação e a sua robustez [Gardner Jr and McKenzie, 1985] são os seus principais pontos fortes, que se tornam bastante úteis em casos reais. Contudo, é necessária cautela nos processos de inicialização das suas componentes e constantes de alisamento. Por serem métodos não paramétricos, dificultam a capacidade de inferir conclusões estatísticas e, consequentemente, não permitem a criação de intervalos de previsão tão direta quanto nos casos de métodos paramétricos. Uma alternativa para estes intervalos passa por recorrer ao método de Bootstrap [Silva, 2013].

4.3.1 Alisamento exponencial simples

Dado um processo $\{X_t\}_{t\in T_0}$, pode-se ver a previsão $\widehat{X}_{t+1}$ como um ajuste em relação ao período anterior X_t por um erro de previsão $e_t = X_t - \widehat{X}_t$, ou seja,

$$\widehat{X}_{t+1} = \widehat{X}_t + \alpha \underbrace{(X_t - \widehat{X}_t)}_{e_t}, \tag{4.8}$$

onde α é uma constante de alisamento que varia entre 0 e 1. Esta constante tem uma interpretação simples, pois quantifica o ajuste de X_{t+1} em relação a X_t. Se α for um valor próximo de zero, conclui-se que o ajuste efetuado é pequeno, caso contrário, caso α seja próximo de 1, conclui-se que o ajuste é maior [Hyndman et al., 2008].

Reescrevendo a equação (4.8) como

$$\widehat{X}_{t+1} = \alpha X_t + (1-\alpha)\widehat{X}_t, \tag{4.9}$$

pode interpretar-se a previsão $\widehat{X}_{t+1}$ como uma média ponderada entre o valor observado no instante anterior e a respetiva estimativa. Substituindo $\widehat{X}_t$ por $\alpha X_{t-1} + (1-\alpha)\widehat{X}_{t-1}$ na equação

(4.8), tem-se

$$\begin{aligned}
& \widehat{X}_{t+1} = \alpha X_{t-1} + (1-\alpha)\widehat{X}_{t-1} + \alpha(X_t - \left[\alpha X_{t-1} + (1-\alpha)\widehat{X}_{t-1}\right]) \iff \\
\iff\ & \widehat{X}_{t+1} = \alpha X_{t-1} + (1-\alpha)\widehat{X}_{t-1} + \alpha(X_t - \alpha X_{t-1} - (1-\alpha)\widehat{X}_{t-1}) \iff \\
\iff\ & \widehat{X}_{t+1} = \alpha X_{t-1} + (1-\alpha)\widehat{X}_{t-1} + \alpha X_t - \alpha^2 X_{t-1} - \alpha(1-\alpha)\widehat{X}_{t-1} \iff \\
\iff\ & \widehat{X}_{t+1} = \alpha X_t + \alpha X_{t-1} - \alpha^2 X_{t-1} + (1-\alpha)\widehat{X}_{t-1}(1-\alpha) \iff \\
\iff\ & \widehat{X}_{t+1} = \alpha X_t + \alpha X_{t-1}(1-\alpha) + \widehat{X}_{t-1}(1-\alpha)^2,
\end{aligned}$$

de onde se deduz a equação

$$\widehat{X}_{t+1} = (1-\alpha)^t \widehat{X}_1 + \alpha \sum_{i=0}^{t-1} (1-\alpha)^i X_{t-i}. \tag{4.10}$$

Na expressão (4.10) nota-se que o peso dado às observações passadas, X_{t-i}, cresce ou decresce exponencialmente. Se α for um valor próximo de 1, uma observação mais recente tem mais peso do que outra mais antiga; caso α seja próximo de zero, o recíproco acontece. Assim, este parâmetro pode ser visto como um regulador de sensibilidade, pois sendo um valor alto (próximo de 1), maior é o peso atribuído à observação mais recente X_t, tornando a previsão mais sensível ao nível [Wang, 2006].

Sendo este um método recursivo, um valor no instante t necessita sempre do valor no instante anterior. Por este motivo, Wheelwright [Wheelwright et al., 1998] propõe que se considere $\widehat{X}_1 = X_1$. No que toca a extrapolação de valores, estes são sempre constantes por este motivo, i.e., $\widehat{X}_{t+h} = \widehat{X}_{t+1}$, para $h \geqslant 2$.

4.3.2 Alisamento linear de Holt

O método de alisamento exponencial simples é insuficiente para lidar com séries que tenham tendência. Por este motivo, Holt [Holt, 1957] estende este método, ao incorporar ferramentas que ajudem a tratar séries que assumam tendência, criando o método de alisamento linear de Holt. Este modelo resume-se, em vez de apenas modelar em termos de X_t, também modelar o nível, l_t e o declive, b_t, originando as equações

$$l_t = \alpha X_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{4.11}$$

$$b_t = \beta(l_t + l_{t-1}) + (1-\beta)b_{t-1} \tag{4.12}$$

$$\widehat{X}_{t+h} = l_t + hb_t, \tag{4.13}$$

para $0 < \alpha < 1$, $0 < \beta < 1$ e $h \geqslant 1$.

Mais uma vez, é necessário os valores iniciais l_1 e b_1 para se aplicar o método recursivo. Para tal, Wheelwright [Wheelwright et al., 1998] recomenda que se considere $\widehat{l}_1 = X_1$ e $\widehat{b}_1 = X_2 - X_1$ ou $\widehat{b}_1 = \frac{X_4 - X_1}{3}$.

4.3.3 Alisamento de Holt-Winters

O método lienar de Holt é incompleto no sentido em que não tem em consideração a sazonalidade de uma série. Daqui surge o método de alisamento exponencial de Holt-Winters que adiciona, para além do nível e do declive, um parâmetro γ ($0 < \gamma < 1$) para a sazonalidade. Antes de expor as equações iterativas deste modelo, é importante perceber as noções de sazonalidade aditiva e multiplicativa, uma vez que este método diferencia os dois casos. Quando uma série apresenta uma sazonalidade independente do nível, diz-se que esta é aditiva. Por outro lado, se a amplitude da sazonalidade de uma série aumentar proporcionalmente com o nível, diz-se que esta possui uma sazonalidade multiplicativa. Estes dois casos resumem-se nas decomposições aditiva, $X_t = T_t + S_t + \epsilon_t$, e multiplicativa, $X_t = T_t \times S_t + \epsilon_t$. A Tabela 4.1 expõe as equações dos modelos de Holt-Winters aditivo e multiplicativo, onde $h_s^+ = [(h-1) \bmod s] + 1$.

Tabela 4.1: Equações do modelo de alisamento exponencial de Holt-Winters.

Modelo aditivo	**Modelo multiplicativo**
$l_t = \alpha(X_t - s_{t-s}) + (1-\alpha)(l_{t-1} + b_{t-1})$	$l_t = \alpha \frac{X_t}{s_{t-s}} + (1-\alpha)(l_{t-1} + b_{t-1})$
$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$	$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$
$s_t = \gamma(X_t - l_t) + (1-\gamma)s_{t-s}$	$s_t = \gamma \frac{X_t}{l_t} + (1-\gamma)s_{t-s}$
$\widehat{X}_{t+h} = l_t + hb_t + s_{t-s+h_s^+},\ h \geqslant 1$	$\widehat{X}_{t+h} = (l_t + hb_t)s_{t-s+h_s^+},\ h \geqslant 1$

Tal como em todos os métodos de alisamento exponencial, são necessários valores iniciais para o nível, o declive e para os índices de sazonalidade. Estes últimos necessitam de s valores iniciais (onde s corresponde à periodicidade da sazonalidade) e a estimação incial do declive de $2s$. Assim sendo, este método exige que a série tenha, pelo menos, $2s$ observações. A Tabela 4.2 resume as expressões de atualização dos valores do nível, $\widehat{l}_s$, do declive, $\widehat{b}_t$ e dos fatores sazonais $\widehat{s}_i$ ($i = 1, 2, \ldots, s$).

Tabela 4.2: Expressões para a inicialização do método de alisamento exponencial de Holt-Winters.

Modelo aditivo	**Modelo multiplicativo**
$\widehat{l}_s = \frac{1}{s}\sum_{i=1}^{s} X_i$	$\widehat{l}_s = \frac{1}{s}\sum_{i=1}^{s} X_i$
$\widehat{b}_s = \frac{1}{s^2}\left(\sum_{i=s+1}^{2s} X_i - \sum_{i=1}^{s} X_i\right)$	$\widehat{b}_s = \frac{1}{s^2}\left(\sum_{i=s+1}^{2s} X_i - \sum_{i=1}^{s} X_i\right)$
$\widehat{s}_i = X_i - \widehat{l}_s,\ 1 \leqslant i \leqslant s$	$\widehat{s}_i = \frac{X_i}{\widehat{l}_s},\ 1 \leqslant i \leqslant s$

Para calcular os valores iniciais de l_s, b_s e s_i, $1 \leqslant i \leqslant s$, pode decompor-se a série nas componentes sistemática (tendência) e sazonal através de um processo MA.

4.3.4 Intervalos de previsão por amostragem de Bootstrap

O método de reamostragem de Bootstrap, proposto por Efron [Efron, 1992], surgiu para casos onde a amostra obtida não possui dimensão grande o suficiente para efetuar estudos estatísticos. No seu cerne, a metodologia de Bootstrap resume-se a reconhecer a amostra obtida como representativa da sua população, fazendo com que se possa elaborar processos de reamostragem com reposição.

Considere-se a amostra $(x_1, x_2, \dots, x_n)$ de uma realização de $(X_1, X_2, \dots, X_n)$. Esta amostra vai representar o papel de população de Bootstrap, da qual se retiram B amostras com reposição[3], cada uma denominada de amostra de Bootstrap.

Seja F_k a função de distribuição empírica do conjunto de previsões $\{\hat{x}^b_{n+k},\ 1 \leqslant b \leqslant B\}$, onde x^b_i, com $1 \leqslant b \leqslant B$, é uma das B previsões reamostradas por Bootstrap. O intervalo de previsão pode ser escrito, então, como

$$\left(F_k^{-1}\left(\frac{\alpha}{2}\right),\ F_k^{-1}\left(1-\frac{\alpha}{2}\right)\right). \tag{4.14}$$

O método de Holt-Winters combina com o de Bootstrap uma vez que, sendo este um método de previsão não paramétrico, é incapaz de calcular intervalos de previsão. O método de Bootstrap é útil, não só nesta metodologia, para construir intervalos de previsão através do método do percentil. Para tal, dada uma série temporal $(x_1, x_2, \dots, x_n)$, basta considerar a primeira metade da série, $(x_1, x_2, \dots, x_{n-k})$, para o processo de modelação e a segunda, $(x_{n-k+1}, x_{n-k+2}, \dots, x_n)$, para o processo de previsão. Assim, os intervalos de previsão são obtidos percorrendo ordenadamente os seguintes passos:

1. Aplicar o método de Holt-Winters (aditivo e multiplicativo) de forma a obeter os valores estimados $\hat{x}_i,\ i = 1, 2, \dots, n$, sendo n o número de observações;

2. Calcular os resíduos $e_j = x_j - \hat{x}_j$ para $j = 1, 2, \dots, n-k$ e os EQM de cada modelo, com o intuito de avaliar a qualidade dos ajustamentos;

3. Escolher o modelo (aditivo ou multiplicativo) que obtiver menor valor do EQM;

4. Verificar se os resíduos do modelo escolhido seguem uma distribuição Normal, com o apoio de testes de hipóteses (Shapiro-Wilk ou Kolmogorov-Smirnov com correção de Lilliefors);

[3] Geralmente o valor de B é alto, por exemplo $B = 2000$, para poder obter-se resultados assintóticos mais corretos.

5. Verificar o pressuposto da independência temporal da série dos resíduos através da análise gráfica das FAC e FACP, em conjunto com o teste de Ljung-Box;

6. Caso seja detetada alguma estrutura de dependência temporal nos resíduos, ajustar um modelo ARIMA aos dados e usá-lo para obter novos resíduos;

7. Recolher uma amostra de Bootstrap, de dimensão $n-k$, dos resíduos com reposição, e^*, da série dos resíduos originais, obtendo um conjunto $(e_1^*, e_2^*, \dots, e_{n-k}^*)$;

8. Contruir uma amostra de Bootstrap $(x_1^*, x_2^*, \dots, x_{n-k}^*)$, onde $x_j^* = \hat{x}_j + e_j^*$ para $j = 1, 2, \dots, n-k$;

9. Repetir os 7.º e 8.º passos B vezes;

10. Utilizando as B séries de Bootstrap obtidas no 8.º passo, considerar as estimativas/previsões pontuais, fixando k para cada uma, como as medianas dos conjuntos $\{\hat{x}_{n+k}^b,\ 1 \leqslant b \leqslant B\}$;

11. Construir os intervalos de previsão de Bootstrap pelo método do percentil, tal como o intervalo (4.14), para um nível de significância α desejado.

4.4 Modelos modificados

Na prática, raramente surgem séries temporais com padrões sazonais simples. Muitas vezes estas estruturas são complexas, por exemplo, períodos sazonais múltiplos (sazonalidades múltiplas), efeito de duplo calendário (derivado de anos bissextos), sazonalidade com periodicidade decimal (não inteira), etc.. Isto motivou a introdução de alterações aos modelos de alisamento exponencial. Desta forma, surgem dois métodos propostos por De Livera [De Livera et al., 2011]: BATS (Box-Cox Transformation, ARMA errors, Trend and Seasonal Components) e TBATS (Trignometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal Components). Considerando o modelo linear de Holt-Winters estudado na Secção 4.3.3, incorporam-se os casos de sazonalidades complexas, erros não Gaussianos e autocorrelacionados, são incluídos(as) erros ARMA, transformações de Box-Cox (de acordo com a expressão 3.2) e T padrões sazonais, obtendo-se:

$$x_t^{(\lambda)} = \begin{cases} \dfrac{x_t^{\lambda} - 1}{\lambda} & , \text{ se } \lambda \neq 0 \\ ln\ x_t & , \text{ se } \lambda = 0 \end{cases}$$

$$x_t^{(\lambda)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{s-m_i}^{(i)} + d_t$$

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

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

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

$$d_t = \sum_{i=1}^{p} \varphi d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t,$$

onde $x_t^{(\lambda)}$ é a observação transformada por uma transformação de Box-Cox de parâmetro λ; $m_1, m_2, \ldots, m_T$ são os períodos sazonais; l_t é o nível estocástico local; b_t é a tendência a curto prazo no instante t; $s_t^{(i)}$ é a i-ésima componente sazonal do instante t; d_t é um processo ARMA(p, q) (representativo do comportamento autocorrelacionado dos erros); ϵ_t é um ruído branco (com média nula e variância finita e constante); α, β e γ_i, $i = 1, 2, \ldots, T$ são as constantes de alisamento e T é o número de padrões sazonais presentes na série. Seguindo as recomendações de Gardner Jr., McKenzie e Arca [Gardner Jr and McKenzie, 1985, Arca et al., 2006], é adicionado um parâmetro de amortecimento ϕ completado com uma tendência de longo prazo b. Esta intervenção garante que as extrapolações de b_t convergem para b em vez de para zero.

4.4.1 Modelos BATS

O modelo BATS é um modelo de espaço de estados capaz de lidar com séries que admitem estruturas complexas de sazonalidade e erros correlacionados. Este é um acrónimo anglicista da expressão Box-Cox Transformation, ARMA errors, Trend and Seasonal Components e representa-se por BATS$(\omega, p, q, \phi, m_1, m_2, \ldots, m_T)$. O parâmetro ω corresponde ao parâmetro da transformação de Box-Cox, anteriormente designado por λ. Caso $\omega = 1$, diz-se que não foi aplicada qualquer transformação [Box and Cox, 1964]. O parâmetro ϕ representa a constante de amortecimento, que, tal como ω, quando $\phi = 1$ este não tem efeito [Gardner Jr and McKenzie, 1985]. Os erros, uma vez que este modelo considera erros correlacionados, são modelados por um processo ARMA(p, q) [Anderson, 1976, Chen et al., 1996] e as constantes $m_1, m_2, \ldots, m_T$ são os períodos das sazonalidades assumidas pelo modelo. Apesar deste modelo se adaptar a casos de sazonalidade múltiplas, estas não podem assumir pariodicidades não inteiras nem altas frequências e a componente de sazonalidade contém m_T valores iniciais, pelo que este modelo, facilmente, pode envolver um grande número de estados. No entanto, este modelo, para além de previsões pontuais, está apto a calcular intervalos de previsão, não necessita do fornecimento de valores iniciais (ao contrário do modelo de Holt-Winters, onde estes têm que ser estimados à priori) e, regra geral, apresenta melhor desempenho do que um modelo de espaço de estados tradicional.

4.4.2 Modelos TBATS

As limitações do modelo BATS, expostas anteriormente, motivaram a formulação do modelo TBATS (Trignometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal Components). Este novo modelo baseia-se nos mesmos conceitos que o modelo BATS, sendo que acrescenta a modelação de séries de Fourier para a componente sazonal [Harvey and Fernandes, 1989, West and Harrison, 1989], obtendo as equações:

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

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} cos\left(\frac{j}{m_i}2\pi\right) + s_{j,t-1}^{*(i)} sen\left(\frac{j}{m_i}2\pi\right) + \gamma_1^{(i)} d_t,\ (j = 1, 2, \dots, k_i)$$

$$s_{j,t}^{*(i)} = s_{j,t-1}^{*(i)} cos\left(\frac{j}{m_i}2\pi\right) - s_{j,t-1} sen\left(\frac{j}{m_i}2\pi\right) + \gamma_2^{(i)} d_t,\ (j = 1, 2, \dots, k_i),$$

onde $\gamma_1^{(i)}$, $\gamma_2^{(i)}$ são as constantes de alisamento, k_i o número de termos a considerar na série de Fourier da i-ésima componente sazonal, $i = 1, 2, \dots, T$, $s_{j,t}^{(i)}$ o nível de estocasticidade da i-ésima componente sazonal e $s_{j,t}^{*(i)}$ o aumento de estocasticidade da i-ésima componente sazonal necessário para explicar a variação na componente sazonal ao longo do tempo. Logo o modelo descreve-se como

$$TBATS(\omega, p, q, \phi, \{m_1, k_1\}, \{m_2, k_1\}, \dots, \{m_T, k_1\}, \{m_1, k_2\}, \dots, \{m_T, k_T\}),$$

onde $k_1, k_2, \dots, k_T$ são os números representativos das quantidades de termos de Fourier utilizados em cada período sazonal e os parâmetros $m_1, m_2, \dots, m_T$, ω, p, q e ϕ têm a mesma função que no modelo BATS.

Obviamente, se $p = q = 0$, tem-se que os erros não são modelados por um processo ARMA e podem, por vezes, não assumir uma distribuição Normal de média nula e variância finita e constante, o que acaba por se tornar numa limitação desta metodologia. Além disso, este método tem um peso computacional (especialmente se a série apresentar um número elevado deobservações) elevado quando comparado com as metodologias descritas anteriormente e não apresenta muita robustez para grandes horizontes de previsão. Os seus intervalos de previsão podem, também, ter amplitudes grandes.

Seleção do número de harmónicos nos modelos TBATS

As previsões de um modelo TBATS dependem do número de harmónicos, i.e., dependem das quantidades de termos de Fourier utilizados em cada período sazonal, k_i. Hyndman

[Kostenko and Hyndman, 2008] recomenda a utilização da regressão linear

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

para determinar os valores de k_i, $i = 1, 2, \ldots, T$. Para isso, deve-se reger pelo seguinte procedimento:

1. Dar início ao processo considerando apenas um harmónico. Vai-se adicionando um harmónico de cada vez e testa-se a sua significância através do teste F até que se atinja um limite de harmónicos significativos, k_i^*;
2. Ajustar o modelo TBATS assumindo $k_i = k_i^*$ e calcular os respetivos critérios de informação (AIC e BIC);
3. Fixar os valores de k_j para $j \neq i$ e calcular os critérios de informação, à medida que se aumenta k_i, até se obter o menor valor do critério de informação escolhido.

Seleção das ordens p e q do processo ARMA dos resíduos

Para além do número de harmónicos a considerar, também é necessário escolher as ordens de autorregressão e de médias móveis que são incorporadas no modelo ARMA(p, q) dos erros. Para tal, deve-se, em primeiro lugar, ajustar o modelo TBATS sem a componente ARMA e calcular os respetivos resíduos. De seguida, estima-se um modelo ARMA aos resíduos, assumindo que estes são estacionários. Assim fica definida a componente ARMA do modelo TBATS, responsável por modelar os erros. Logo, resta apenas ajustar, novamente, um modelo TBATS, mas agora considerando a componente ARMA(p, q) identificada anteriormente, sendo os seus coeficientes estimados todos em conjunto. Obviamente, em prol da parcimoniosidade, deve-se apenas considerar a componente ARMA se o modelo resultante admitir valores de critérios de informação (AIC ou BIC) inferiores aos do modelo não considerando componente ARMA.

4.5 Medidas de avaliação

De modo a avaliar, quer a qualidade de ajuste, quer a qualidade de previsão, vários autores recomendam as suas métricas de avaliação de preferência. Na verdade, dado que não existe nenhuma medida consensual para avaliação da qualidade de um modelo, é altamente recomendado o uso de várias. Nomeadamente, uma medida dependente da escala das osbervações, o erro quadrático médio, a respetiva raiz que apresenta unidades na escala das observações, a raiz do erro quadrático médio, uma medida percentual, o erro percentual absoluto médio, uma

medida escalada, o erro escalado absoluto médio e a estatística U de Theil, que são as medidas de avaliação utilizadas nesta dissertação.

Erro quadrático médio

O erro quadrático médio, EQM, define-se como a média dos quadrados da diferença entre o valor observado e o estimado, ou seja,

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2.$$

Esta medida depende da escala dos dados, uma vez que é calculada com base no erro de previsão a 1-passo, i.e., $e_t = X_t - \widehat{X}_t$. A sua raiz, $REQM = \sqrt{EQM}$, é frequentemente utilizada em seu lugar, pois admite valores na escala efetiva dos dados. Assim, o modelo que apresentar o valor mais baixo de EQM, ou de REQM, é considerado o mais preciso.

Erro percentual absoluto médio

O EQM e, consequentemente, o REQM são medidas bastante sensíveis a outliers, i.e., os seus valores são facilmente alterados por causa de apenas um (ou mais) valor(es), originando numa ideia errada de avaliação do modelo. O erro percentual absoluto médio, EPAM, que se traduz na percentagem média do erro $e_t = X_t - \widehat{X}_t$ em relação à escala dos dados, não apresenta esta sensibilidade e é definido por

$$EPAM = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i - \widehat{X}_i}{X_i}\right| \times 100\ \%.$$

Deste modo, considera-se o modelo mais preciso como aquele que apresentar o menor valor de EPAM.

Apesar do EPAM ter a vantagem da falta de sensibilidade em relação a outliers, este não pode ser calculado quando a série admite valores nulos e aumenta drasticamente se se tratar de uma série com valores próximos de zero. Para tal, introduziu-se o erro percentual absoluto médio simétrico (EPAMS),

$$EPAMS = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i - \widehat{X}_i}{X_i + \widehat{X}_i}\right| \times 200\ \%,$$

que contorna estas limitações [Caiado, 2011]. Porém, Hyndman e Koehler criticam esta medida, pois afirmam que retorna erros elevados quando a série assume valores nulos [Hyndman and Koehler, 2006].

Erro escalado absoluto médio

O erro escalado absoluto médio, EEAM, define-se como

$$EEAM = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i - \widehat{X}_i}{\frac{1}{n-s}\sum_{i=s+1}^{n}|X_i - X_{i-s}|}\right|,$$

onde s é a sazonalidade da série (no caso de ausência de qualquer sazonalidade, considera-se $s = 1$). Esta medida retorna a relação entre o erro de previsão $e_t = X_t - \widehat{X}_t$ e o erro absoluto médio da previsão naïve, $\frac{1}{n-s}\sum_{i=s+1}^{n}|X_i - X_{i-s}|$.

Definição 13. O método de previsão naïve é aquele que projeta o último valor ou (período sazonal) para o futuro, ou seja, $\widehat{X}_{n+1} = X_{n+1-s}$, para uma sazonalidade $s \geqslant 1$.[4]

Esta medida é útil para comparar modelos ajustados a séries de diferentes grandezas, pois é independente de escalas e possui ainda a vantagem de ser sempre finita exceto em casos de séries constantes, i.e., $X_i = X_{i-1},\ \forall\ i = 2, 3, \ldots n$. Sendo esta medida uma relação, indica que se o seu valor for superior à unidade, então as previsões do método escolhido são, em média, menos precisas do que as naïve, logo considera-se o modelo mais preciso aquele que apresentar menor valor do EEAM.

Estatística U de Theil

A estatística U de Theil combina duas vantagens das medidas anteriores. À semelhança do EQM, este considera o quadrado dos erros, o que dá mais peso aos erros mais elevados, quando comparados com erros inferiores; e oferece uma relação entre qualquer método de previsão e o método naïve. Esta é dada pela expressão

$$U = \sqrt{\frac{\sum_{i=1}^{n-1}\left(\frac{\widehat{X}_{i+1} - X_{i+1}}{X_i}\right)^2}{\sum_{i=1}^{n-1}\left(\frac{X_{i+1} - X_i}{X_i}\right)^2}} \tag{4.15}$$

onde o numerador corresponde ao somatório dos quadrados das diferenças entre $f_{i+1} = \frac{\widehat{X}_{i+1} - X_i}{X_i}$, que representa a variação relativa prevista, e $a_{i+1} = \frac{X_{i+1} - X_i}{X_i}$, que representa a variação relativa real.

Da expressão (4.15) retira-se que:

[4] Note-se que $s = 1$ entende-se como ausência de sazonalidade.

- $U = 0$ se $f_{i+1} = a_{i+1}$, $i = 1, 2, ..., n-1$, ou seja, se o modelo escolhido efetua previsões sem qualquer erro;
- $U = 1$ se $f_{i+1} = 0$, ou seja, se os erros de previsão do modelo escolhido são iguais aos do método naïve. Nesta situação diz-se que ambos os métodos são igualmente precisos;
- $U > 1$ se f_{i+1} tiver o sinal oposto de a_{i+1}, pois isto indica que o numerador é superior ao denominador. Nesta ocasião, diz-se que o método de previsão escolhido é menos preciso do que o método naïve, logo deve ser descartado;
- $U < 1$ se f_{i+1} tiver o mesmo sinal que a_{i+1}, pois isto implica que o numerador é inferior ao denominador. Neste caso, afirma-se que o método de previsão escolhido é mais preciso do que o método naïve.

À semelhança das medidas de avaliação já explicadas, quanto menor for o valor da estatística U de Theil de um modelo, mais preciso este é.

Capítulo 5

Aplicação das metodologias de previsão aos dados diários

O mercado das seguradoras é afetado por inúmeros fatores externos, pelo que pode ser extremamente variável. Um exemplo são as alterações climáticas que podem gerar mais acidentes rodoviários, danos em habitações, depressões, ou até suicídios. Por este motivo, será espectável algum erro na modelação temporal, principalmente em séries diárias que são as mais suscetíveis a esta variabilidade. Será, então, efetuada uma análise exploratória dos dados diários, seguida da modelação dos mesmos através das metodologias SARIMA, Holt-Winters e TBATS.

Entende-se por séries marginais, as séries das categorias **DNA**, **REL**, **TMP** e **restantes causas**. A categoria **DNA** corresponde a sinistros causados por água (e.g. infiltrações e inundações). Será de esperar alguma correlação entre esta e a categoria **TMP**, pois chuvas intensas (ocorrências que correspondem a sinistros da categoria **TMP**) podem causar problemas de infiltrações ou inundações. Para além desta, também a categoria **REL** pode estar correlacionada com a **TMP**, uma vez que trovoadas podem causar vários tipos de falhas elétricas, que correspondem a sinistros da categoria **REL**. As séries das categorias **DNA**, **REL**, **TMP** e **restantes causas** são as constituintes da série **total** que representa a soma dos valores de todas as séries marginais, para cada instante t.

5.1 Análise descritiva dos dados diários

A análise descritiva é um passo fulcral em qualquer estudo estatístico, pois estabelece as bases necessárias que permitem a exploração de ferramentas estatísticas. Para além de identificar características intrínsecas nos dados, esta análise, por vezes, fornece uma ligeira perceção de possíveis resultados ou problemas que poderão ser encontrados na exploração das metodologias que se tencionam estudar. O presente estudo tem como foco a análise dos números diários de sinistros de

habitação registados pela companhia de seguros Ageas. Inicialmente foram disponibilizadas estas contagens a partir do dia 1 de janeiro de 1995, que se dividiram em quatro categorias principais: **danos por água** (**DNA**), **riscos elétricos** (**REL**), **tempestades** (**TMP**) e **restantes causas**.

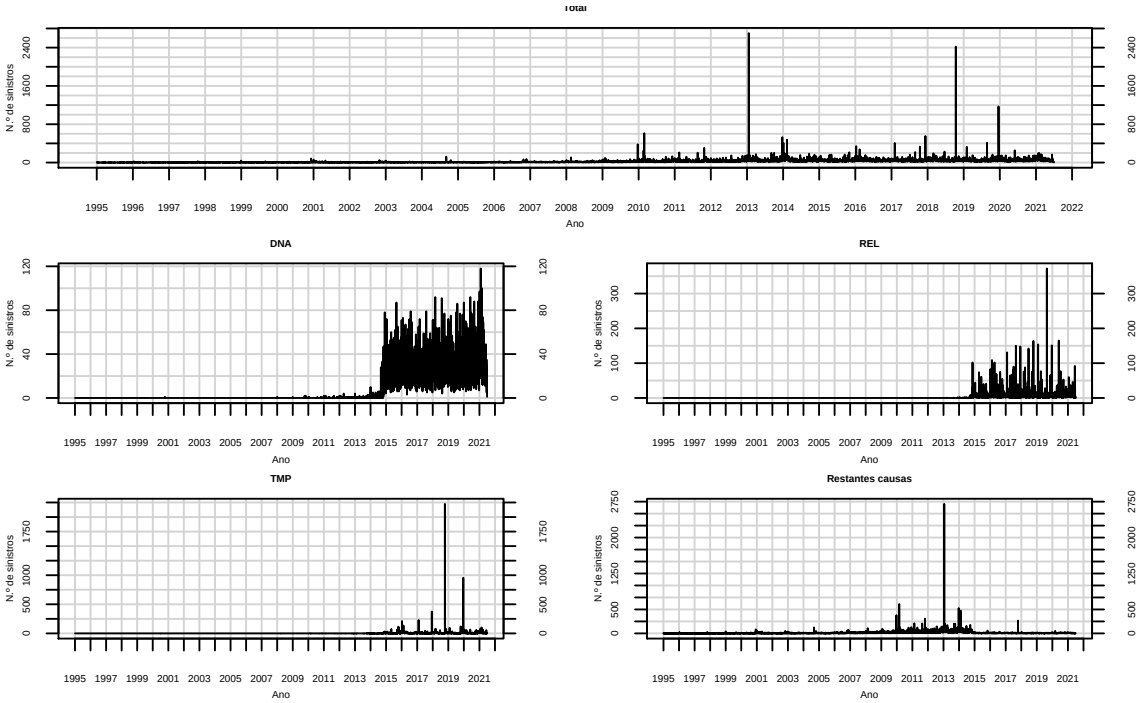


Figura 5.1: Representação gráfica do número de sinistros no período observado.

Os critérios de categorização de sinistros não são sempre os mesmos; as categorias atuais regem-se por critérios diferentes dos utilizados em 1995. Este fenómeno é visível na Figura 5.1 onde se nota uma maior variabilidade nos anos mais recentes (mais precisamente a partir do ano 2015) em todas as séries diárias, exceto na categoria **restantes causas**. Este fenómeno deve-se ao facto de, até ao princípio de 2015, muitos sinistros serem categorizados como "Geral" ou "Geral Migrado" (atualmente pertencentes à classe **restantes causas**), ao contrário dos dias de hoje onde se distinguem novas categorias (e.g. **danos por água**, **riscos elétricos** e **tempestades**). Esta evidência é apoiada pelas medidas descritivas correspondentes (presentes no Apêndice B). Por este motivo, considerar-se-ão apenas as contagens a partir do dia 1 de janeiro de 2015 (esta restrição traduz-se numa diminuição da amostra de 9 678 para 2 373 dias).

Como seria de esperar, ao trabalhar com dados reais, depara-se frequentemente com outliers. Estas séries são pouco informativas, pois não fornecem qualquer informação sobre sazonalidade nem tendência. Assim, antes de partir para qualquer exploração descritiva ou análise temporal, é necessário suavizá-las, tratando os outliers adequadamente. É possível verificar que muitos outliers da série diária **total** (série que corresponde à soma das séries diárias marginais) são também visíveis na série diária **TMP**, ou seja, grande parte do comportamento instável do número

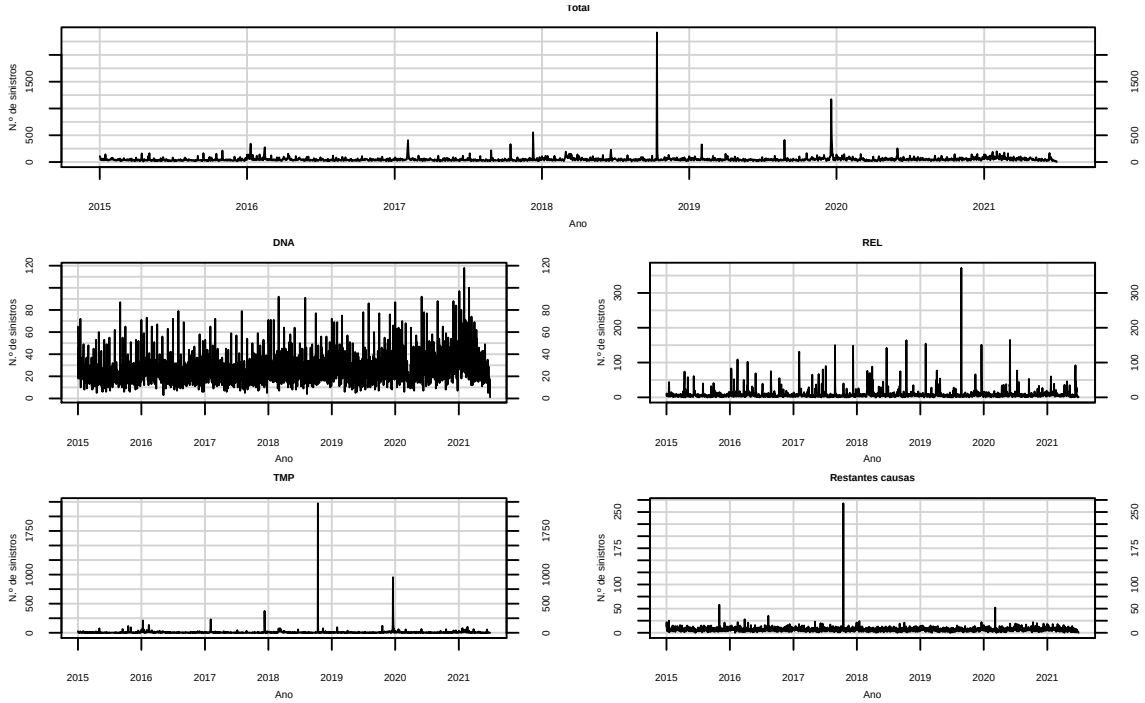


Figura 5.2: Representação gráfica do número de sinistros de janeiro de 2015 a junho de 2021.

total de sinistros é justificado por tempestades, nomeadamente:

- **Tempestade Ana (dia 10/12/2017):**

A tempestade Ana gerou 3 187 ocorrências distribuídas por quedas de árvores, inundações, limpezas de vias e movimentos de massa. Além de uma vítima mortal de 45 anos e cinco vítimas com ferimentos ligeiros, houve danos visíveis em vários pontos do país; problemas na circulação de comboios e uma quebra de energia que provocou falhas nos sistemas de comunicação um pouco por todo o país. Os distritos mais afetados, por ordem de mais ocorrências, foram Lisboa, Porto, Aveiro, Viseu, Braga, Coimbra, Leiria e Viana do Castelo.

- **Furacão Leslie (dia 13/10/2018):**

Formado a 22 de setembro de 2018 no oceano Atlântico, o furacão Leslie foi o mais forte a atingir Portugal desde 1842. Passou a algumas centenas de quilómetros a sudoeste do arquipélago dos Açores e atingiu o norte do continente português na noite de 13 de outubro de 2018. Provocou 27 feridos ligeiros, 61 desalojados, prejuízos de cerca de 120 milhões de euros (em território nacional) e originou a participação de 28 000 sinistros às companhias seguradoras.

- **Tempestade Elsa e Depressão Fabien (de 18/12/2019 a 21/12/2019):**

Originando a participação de mais de 10 mil sinistros (90% dos quais respeitam a seguros de casas e de atividades comerciais e industriais) a nível nacional, a tempestade Elsa (de 18/12/2019 a 20/12/2019) seguida da depressão Fabien (ocorrida no dia 21/12/2019) causou estragos no custo de cerca de 18,2 milhões de euros. Esta tempestade fez três mortos e deixou mais de 100 pessoas desalojadas, provocando danos em habitações, linhas de comboio, vias rodoviárias e na rede elétrica.

Os diagramas em caixa de bigodes da Figura B.1 do Apêndice B mostram as distribuições anuais das quantidades diárias de sinistros. É possível visualizar a quantidade excessiva de outliers presentes nestas séries. Por este motivo, não serão "suavizados" todos estes, pois esta transformação seria excessiva e resultaria num estudo temporal ilógico, uma vez que os dados estariam longe da realidade, ou seja, deixariam de ser representativos das suas populações.

Partindo deste princípio, conclui-se que tem de ser escolhido um método de suavização de outliers que respeite as características dos dados. Assim sendo, uma possível opção será estudar as sazonalidades das séries diárias e substituir os outliers por um valor que vá ao encontro dos mesmos. Os gráficos anteriores (Figura 5.2) são pouco informativos dada a alta densidade de observações, logo não é possível notar a presença de padrões sazonais nem o comportamento dos dados, por cada semana, mês ou ano, através de uma análise visual. No entanto, a função de autocorrelação parcial pode ser útil, pois mostra a autocorrelação temporal entre duas observações desfasadas no tempo, o que pode evidenciar possíveis sazonalidades.

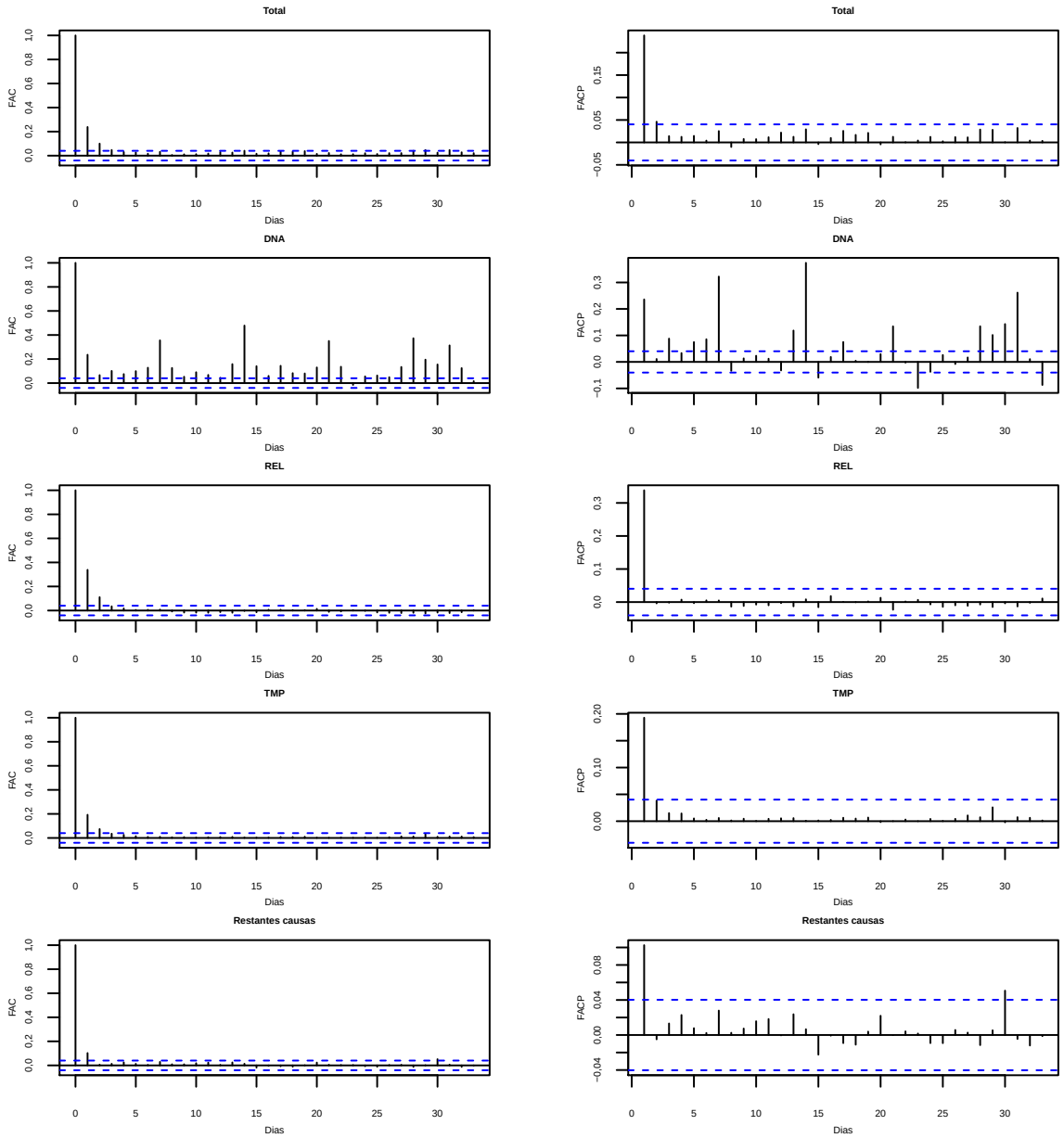


Figura 5.3: FAC e FACP das séries diárias.

Os gráficos correspondentes à FACP das séries diárias **total**, **DNA** e **restantes causas** da Figura 5.3 mostram autocorrelações relevantes para lags de 7 e 14 dias, ou seja, as contagens parecem estar autocorrelacionadas semanalmente. Esta ideia sugere uma transformação baseada numa média semanal. Será, então, cada observação outlier substituída pela média das restantes observações presentes na semana a que esta pertence

$$x_o = \sum_{\substack{i=a \\ i \neq o}}^{b} \frac{x_i}{6}$$

onde:

- o é o índice cuja observação corresponde a um outlier (x_o);
- a é o primeiro dia da semana a que pertence a observação x_o (segunda-feira);[1]
- b é o último dia da semana a que pertence a observação x_o (domingo).[1]

Nota 5. Esta transformação apenas faz sentido se não existir mais do que uma observação outlier por semana, pois caso contrário o denominador desta expressão deixaria de fazer sentido e estar-se-ia a acrescentar informação a mais aos dados ao considerar outliers no peso desta média (este não é o caso do presente estudo).

Uma vez decidida a transformação a efetuar, têm que se estabelecer os critérios que definem observações outliers, tal como foi referido anteriormente. Serão, portanto, definidos parâmetros empíricos na consideração de outliers, de modo a tentar ao máximo manter estas características. Para cada categoria, exceto **DNA** (que não apresenta um comportamento alterado por outliers), será estipulado um "nível máximo" para que uma contagem diária seja considerada razoável. Caso o valor observado seja igual ou superior ao "nível máximo" estabelecido, será efetuada a transformação descrita. Após terem sido atribuídos vários valores aos "níveis máximos" e analisadas as respetivas representações gráficas e contagens, ficaram definidos os máximos apresentados na Figura 5.4:

[1] Assume-se que uma semana começa à segunda-feira e termina ao domingo.

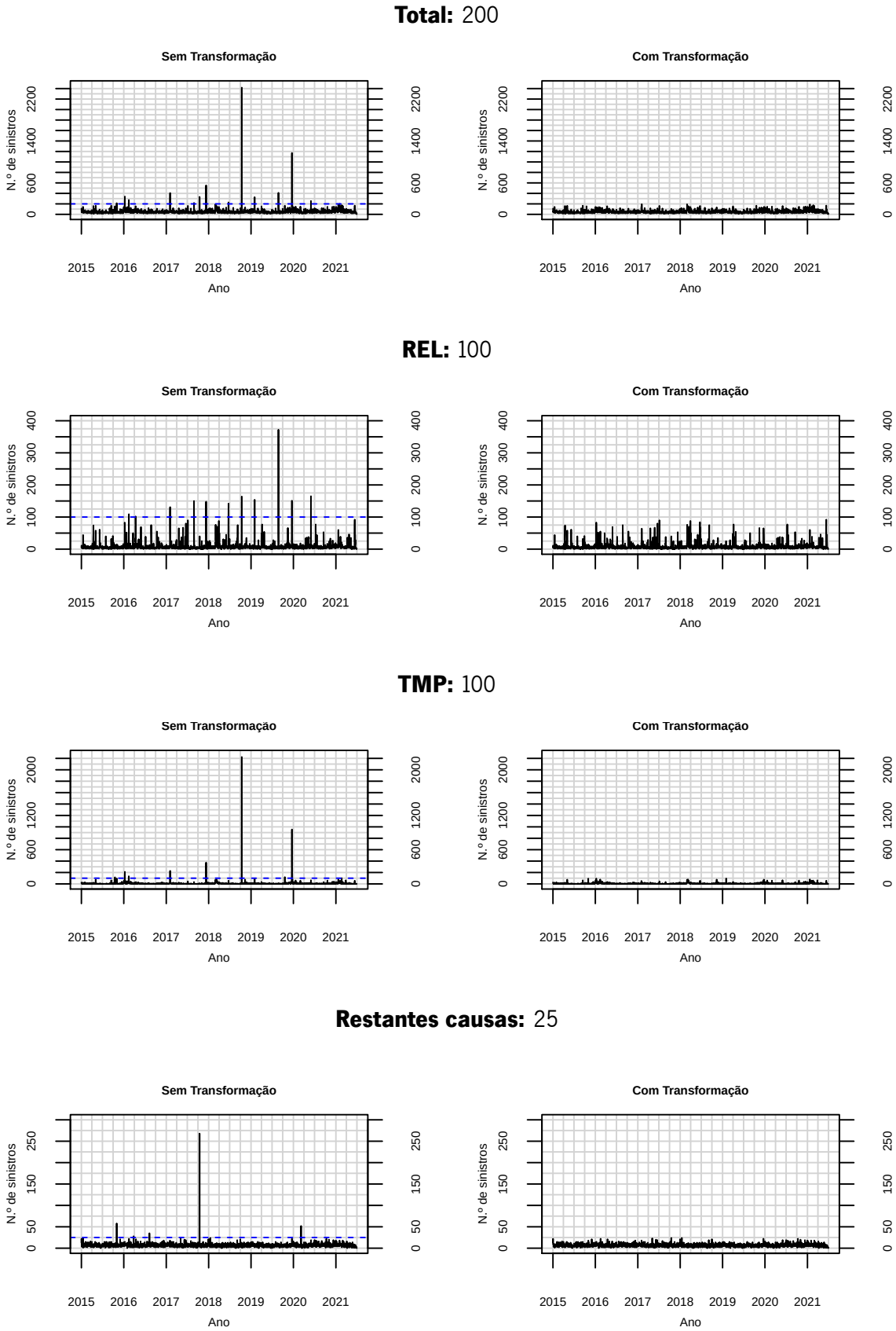


Figura 5.4: Representação gráfica do efeito da suavização de outliers considerando o respetivo nível máximo.

A transformação mostra-se empiricamente eficaz, uma vez que já não é possível identificar outliers severos através das análises gráficas das contagens diárias. Foram suavizadas 57 observações no total: 12 na série diária **REL**, 17 na **TMP**, 6 na **restantes causas** e 22 na **total**. Serão então estas novas séries a considerar para alcançar os objetivos definidos.[2]

Com base nas séries diárias suavizadas, faz sentido partir para a análise descritiva dos dados. Deste modo, a Tabela 5.1 mostra as medidas descritivas das séries diárias suavizadas a partir do ano 2015.[3]

Tabela 5.1: Medidas descritivas das séries diárias.

	Categoria				
	DNA	**REL**	**TMP**	**Restantes causas**	**Total**
Início	01/01/2015	01/01/2015	01/01/2015	01/01/2015	01/01/2015
Fim	30/06/2021	30/06/2021	30/06/2021	30/06/2021	30/06/2021
Dimensão	2373	2373	2373	2373	2373
N.º de zeros	0	35	192	9	0
Amplitude	1 - 118	0 - 92	0 - 96	0 - 24	2 - 194
$Q_{0,25}$	18	4	2	5	32
Mediana	24	6	4	6	42
$Q_{0,75}$	33	9	7	9	56
Média	26,93	8,11	6,52	6,87	48,19
Desvio padrão	13,40	9,88	9,73	3,44	25,32
Variância	179,56	97,57	94,69	11,81	641,04
Coeficiente de variação	0,50	1,22	1,49	0,50	0,53
N.º de outliers	90	189	232	42	145

Comparando as Tabelas 5.1 e B.1 (presente no Apêndice B)[4] conclui-se que a suavização dos outliers foi extremamente eficaz, pois reduziram-se os números de outliers e a amplitude drasticamente.

A alta variabilidade das séries diárias é visível nos correspondentes diagramas em caixa de bigodes (quando substituídos os outliers).

A Figura 5.5 mostra uma tendência e variância crescentes nas séries diárias das categorias **total** e **DNA**, enquanto que as restantes aparentam média constante. Estes pressupostos terão de ser verificados para as metodologias utilizadas nesta secção, logo terá de ser aplicada alguma transformação às séries diárias de modo a admitirem, pelo menos, estacionariedade fraca.

[2] Por fins de simplicidade, estas séries não serão referidas como suavizadas.
[3] A Tabela 5.1 considera outliers teóricos e não os escolhidos para serem suavizados.
[4] As Tabelas 5.1 e B.1 consideram outliers teóricos e não os escolhidos para serem suavizados.

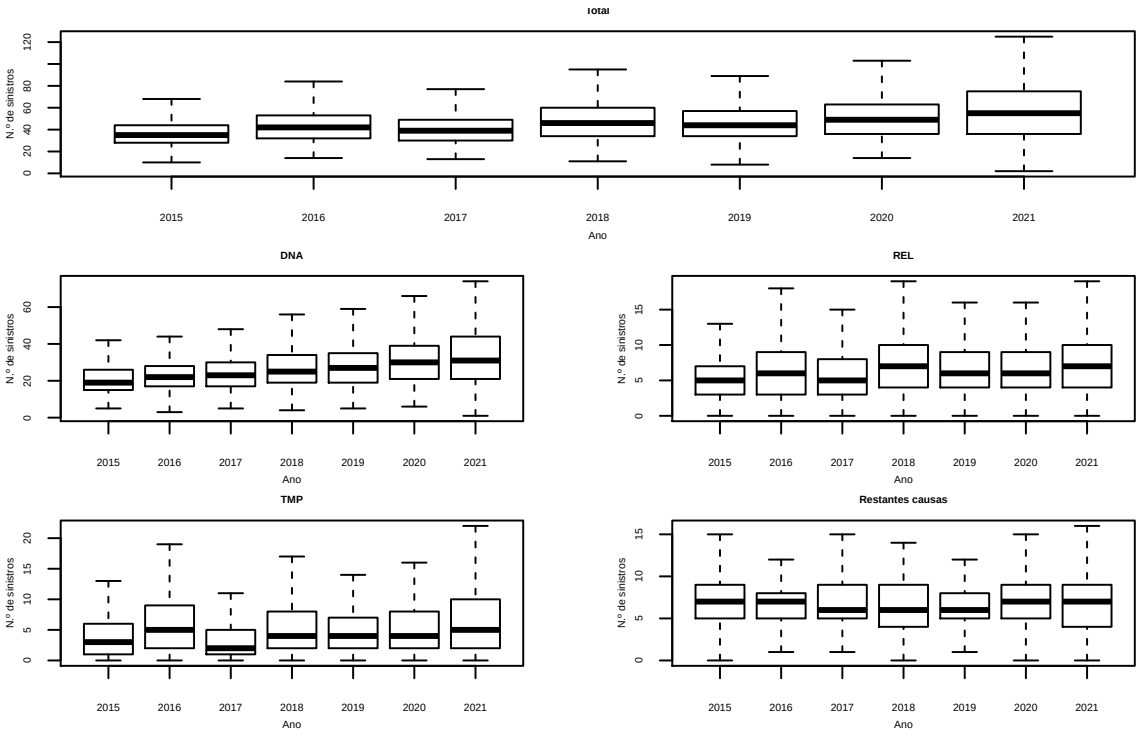


Figura 5.5: Diagramas em caixa de bigodes das séries diárias sem outliers.

Tal como foi referido anteriormente, as correlações entre as classes de sinistros devem ser alvo de estudo. A Tabela 5.2 enuncia os valores do coeficiente de correlação de Pearson correspondente a todos os pares de classes.

Tabela 5.2: Coeficientes de correlação de Pearson entre as séries diárias.

Categoria	DNA	REL	TMP	Restantes causas
DNA	1	0,129	0,399	0,385
REL		1	0,412	0,090
TMP			1	0,192
Restantes causas				1

Ao contrário do que era esperado, os valores dos coeficientes de correlação de Pearson não mostram grandes correlações lineares entre as categorias de sinistros. O facto de estes valores serem pequenos pode se dever à alta variabilidade das séries diárias. Provavelmente, estes irão mostrar correlações mais altas para séries mensais. Apesar disso, nota-se que os valores mais altos correspondem às correlações de **TMP** com **DNA** e **REL**, o que vai de acordo com o referido anteriormente.

5.2 Aplicação dos métodos de previsão aos dados diários

Após a exploração descritiva e tratamento de outliers dos dados diários, segue-se a aplicação das metodologias de previsão explicadas no Capítulo 4: modelos SARIMA, Holt-Winters e TBATS. Em todas estas abordagens, o primeiro passo consiste em transformar (à excessão do método de Holt-Winters) as séries de modo a ficarem estacionárias (tanto para a variância como para a média). A estacionariedade para a variância será atingida com a transformação de Box-Cox e a estacionariedade para a média será estudada através da análise gráfica das FAC e FACP empíricas de cada série, com o auxílio dos testes de estacionariedade ADF e KPSS. Contudo, de acordo com a expressão (3.2), as séries não podem admitir valores não positivos. A Tabela 5.1 mostra que as séries das categorias **REL**, **TMP** e **restantes causas** apresentam valores nulos.[5] De forma a poder aplicar-se esta transformação, irá ser somada uma unidade a todas as observações das mesmas, de forma a excluir valores nulos e preservar as proporcionalidades dentro de cada série.[6]

Nota 6. A transformação de Box-Cox assume uma expressão de dois parâmetros (λ_1 e λ_2) que apenas necessita de verificar a condição $x_i > -\lambda_2$, ou seja, x_i pode assumir valores negativos. No entanto, esta expressão é mais complexa do que a transformação de Box-Cox de um parâmetro, pelo que a interpretação dos valores transformados $x_i^{(\lambda_1,\lambda_2)}$ se torna mais complicada.

5.2.1 Caso I: Série diária do número total de sinistros

Nesta secção, será exposto todo o processo de modelação da série diária **total**, enquanto que as modelações das séries marginais (**DNA**, **REL**, **TMP** e **restantes causas**) serão descritas na Secção 5.2.2. Inicialmente, dividem-se os dados em conjuntos de treino e de teste, de forma a conseguir avaliar-se tanto o ajustamento como a capacidade preditiva dos modelos. Tendo observações desde o dia 1 de janeiro de 2015 até ao dia 30 de junho de 2021, decidiu-se definir o conjunto de treino como as primeiras 2192 observações (correspondentes ao período de 1 de janeiro de 2015 até 31 de dezembro de 2020) e as restantes 181 como o conjunto de teste. Desta forma, os conjuntos de treino e de teste representam, respetivamente, cerca de 92% e 8% da totalidade dos dados. Todos os modelos de agora em diante serão ajustados ao conjunto de treino e avaliados pelo seu poder preditivo no conjunto de teste.

Modelo SARIMA

A escolha de modelos SARIMA respeita as três etapas principais da metodologia de Box-Jenkins: identificação do modelo, estimações dos seus parâmetros e análise de diagnóstico.

[5] A Tabela 5.1 considera outliers teóricos e não os escolhidos para serem suavizados.

[6] Dada a natureza (número diário de sinistros) e a amplitude das séries, esta alteração não influencia gravemente o estudo.

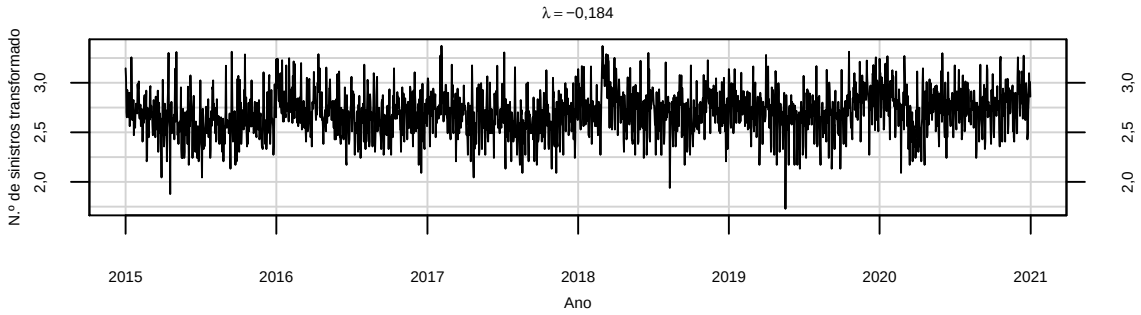


Figura 5.6: Representação gráfica da série diária da categoria **total** após a transformação de Box-Cox.

A Figura 5.6 ilustra o efeito da transformação de Box-Cox (λ = -0,184) na série diária. Esta aparenta variância constante, logo será a série escolhida para o ajuste do modelo SARIMA. Para fins de estabilização da média, irão ser analisados os gráficos das FAC e FACP empíricas da série após a transformação de Box-Cox (Figura 5.7), com o auxílio dos testes de estacionariedade ADF e KPSS. O número de lags (valor de p) é o mesmo tanto para o teste ADF como para o KPSS, que se baseia na regra de Ng & Perron [Ng and Perron, 1995] exposta no Capítulo 3. No teste ADF, rejeitar a hipótese nula significa que a série dos resíduos é estacionária, ao contrário do teste KPSS que a rejeição da hipótese nula implica que a série não é estacionária.

Segundo estes testes, a estacionariedade é aceite, pois, para um nível de significância de 5% e para 24 lags, as suas estatísticas de teste são $20,1891$ para o teste ADF e $0,0697$ para o teste KPSS. Os valores destas estatísticas quando comparadas com os respetivos valores críticos, $6,250$ e $0,146$, não permitem rejeitar esta hipótese.

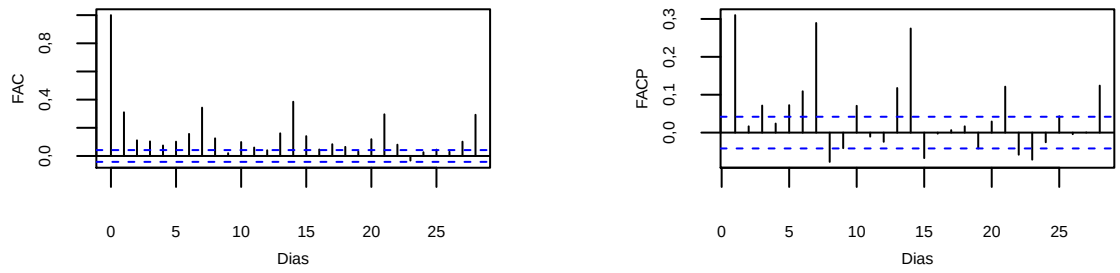


Figura 5.7: FAC e FACP da série diária da categoria total após a transformação de Box-Cox.

Partindo da série estacionária, ajusta-se então a parte sazonal do modelo. Na representação gráfica das FAC e FACP (Figura 5.7) nota-se uma forte sazonalidade semanal. Além disso, o gráfico da série (Figura 5.6) apresenta uma sazonalidade anual. Constata-se que esta série cresce nos finais e inícios de cada ano (períodos correspondentes aos invernos, onde existem maiores registos de tempestades). Portanto, serão consideradas as duas possíveis sazonalidades, $s = 7$ e $s = 365$.

Nota 7. No caso da sazonalidade anual considera-se que $s = 365$ apesar de um ano ter aproximadamente 365,24 dias, seguindo a sugestão de Benth & Benth (2010) de remover as observações registadas a 29 de fevereiro (passando, então, a exatamente 365 observações por ano). Porém, os modelos ajustados com estas sazonalidades não poderão efetuar previsões para o dia 29 de fevereiro. Além disso, os modelos SARIMA têm uma limitação computacional que não permite ajustar uma sazonalidade superior a 350 (excede a memória da máquina), ou seja, este caso não será considerado nesta metodologia.

Os restantes parâmetros do modelo (p, d, q, P, D e Q) são estimados percorrendo várias combinações para os seus valores, fazendo p, q, P e Q percorrer todos os valores entre entre 0 e 2 e a diferenciação sazonal (D) entre 0 e 1. Pelas conclusões dos testes ADF e KPSS, escolhem-se os modelos com $d = 0$, uma vez que a série é estacionária. Através da análise gráfica da Figura 5.7 vê-se, claramente, um decaimento exponencial de 7 em 7 lags, logo será de esperar um modelo com parâmetros p, q, P e Q não nulos. A Tabela 5.3 apresenta os cinco modelos com menor AIC e os cinco modelos com menor BIC dentro das 162 combinações de parâmetros descritas anteriormente.

Tabela 5.3: Ajustamento de modelos SARIMA para a série diária da categoria **total** com transformação de Box-Cox.

Modelo	**AIC**	**Modelo**	**BIC**
SARIMA$(2,0,2)(2,1,1)_7$	$-1248,61$	SARIMA$(2,0,1)(2,1,1)_7$	$-1208,46$
SARIMA$(2,0,1)(2,1,1)_7$	$-1248,28$	SARIMA$(2,0,1)(0,1,1)_7$	$-1204,35$
SARIMA$(2,0,2)(2,1,2)_7$	$-1246,63$	SARIMA$(2,0,2)(2,1,1)_7$	$-1203,10$
SARIMA$(2,0,1)(2,1,2)_7$	$-1246,32$	SARIMA$(2,0,1)(2,1,2)_7$	$-1200,80$
SARIMA$(1,0,2)(2,1,2)_7$	$-1240,08$	SARIMA$(1,0,2)(2,1,1)_7$	$-1200,26$

Analisando os valores do AIC, nota-se que a diferença de qualidade de estimação dos dois primeiros modelos não é significativa, contudo, o modelo SARIMA$(2,0,1)(2,1,1)_7$ apresenta o menor valor do BIC, logo este será o modelo escolhido para modelar os dados com a metodologia SARIMA. Claramente, este modelo não é inalterável, ou seja, conforme for necessário poderão considerar-se outros modelos. A Tabela 5.4 mostra os detalhes do modelo escolhido (todos os coeficientes pertencem aos intervalos de 95% de confiança correspondentes, logo são todos estatisticamente significativos para um nível de significância de 5%).

Tabela 5.4: Características do modelo SARIMA ajustado à séria diária do número total de sinistros.

SARIMA$(2,0,1)(2,1,1)_7$	$AIC \approx -1248,28$		$BIC \approx -1204,35$		$\widehat{\sigma} \approx 0,1805$	
Parâmetro	ϕ_1	ϕ_2	θ_1	ν_1	ν_2	η_1
Estimativa	$1,1944$	$-0,2124$	$-0,9241$	$0,0152$	$0,0965$	$-0,9885$
Erro padrão	$0,0299$	$0,0256$	$0,0190$	$0,0228$	$0,0220$	$0,0041$

O próximo passo nesta metodologia passa por realizar uma análise dos resíduos. Estes devem ser temporalmente não correlacionados e apresentar uma distribuição aproximadamente Gaussiana com média nula e variância constante.

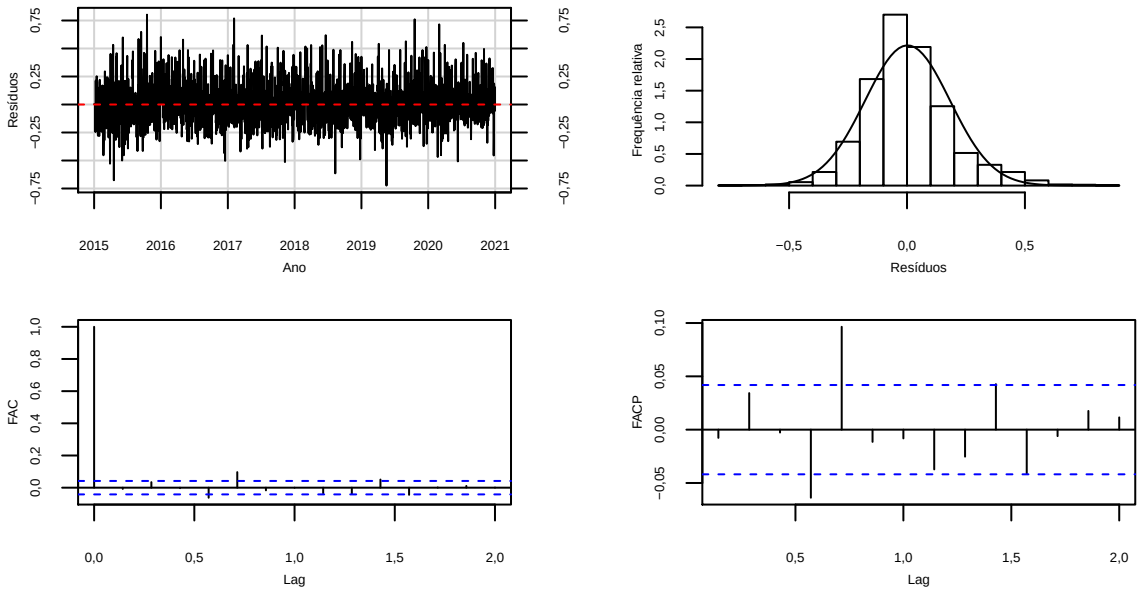


Figura 5.8: Série dos resíduos do modelo SARIMA ajustado à série diária da categoria **total** após uma transformação de Box-Cox e respetivo histograma, FAC e FACP estimadas.

Apesar de o gráfico apresentar variância constante em torno da média nula e o histograma parecer relativamente simétrico com uma distribuição Gaussiana, os erros apresentam correlações temporais e as hipóteses nulas dos testes de Kolmogorov-Smirnov (com correção de Lilliefors) e de Shapiro-Wilk são rejeitadas[7]. O teste de Ljung-Box[8] também não rejeita a hipótese alternativa de os resíduos apresentarem correlações temporais. Isto significa que o modelo pode não estar a explicar toda a correlação temporal das observações, pois os resíduos obtidos através deste método devem ser caracterizados como ruído branco Gaussiano (sem correlação temporal, média nula e variância constante).

[7] Todos os modelos descritos anteriormente estão nas mesmas condições.

[8] O teste foi efetuado com $m = 35$.

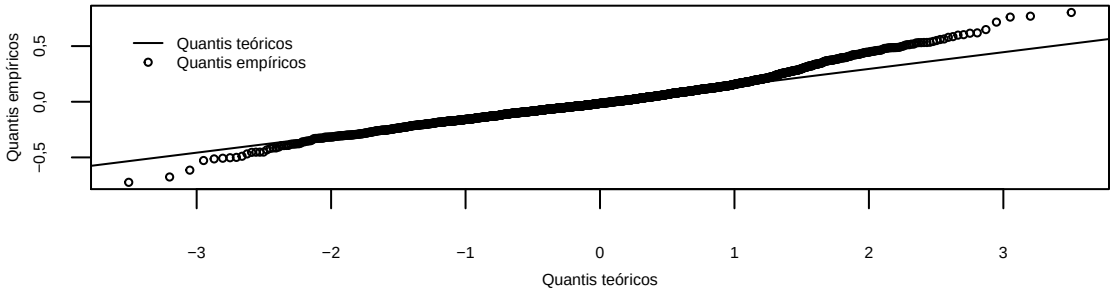


Figura 5.9: Representação gráfica dos quantis teóricos de uma distribuição $N(0, 0, 1805^2)$ em função dos quantis empíricos dos resíduos do modelo SARIMA ajustado à série diária do número total de sinistros.

A Figura 5.9 mostra que os quantis empíricos da série diária do número total de sinistros se desviam ligeiramente dos quantis teóricos de uma distribuição Normal.

Nota 8. Apesar de a teoria de Box-Jenkins e a modelação TBATS requererem os pressupostos de os erros dos modelos seguirem uma distribuição aproximadamente Normal e serem independentes, neste estudo estes pressupostos são relaxados devido à natureza dos dados reais. Serão sempre obtidas as previsões em cada processo de modelação, pois este é o principal objetivo deste trabalho.

Apesar de os testes de Kolmogorov-Smirnov (com correção de Lilliefors) e de Shapiro-Wilk rejeitarem a hipótese de normalidade na série, a Figura 5.8 mostra que os resíduos apresentam uma distribuição simétrica forte em torno do valor nulo e a curva da função de distribuição empírica tende a estar ao nível das barras do histograma, logo os pressupostos de normalidade $\varepsilon_t \sim N(0, \sigma^2_{\varepsilon_t})$ necessários para o método SARIMA podem ser relaxados.

A Figura 5.10 ilustra o ajuste (a vermelho), as previsões intervalares a 95% (tracejado azul) e pontuais (linha contínua azul) ao número diário total de sinistros. Tanto os valores ajustados como as previsões do modelo estão transformados(as) conforme a expressão (3.3), de modo a facilitar as suas interpretações gráficas.

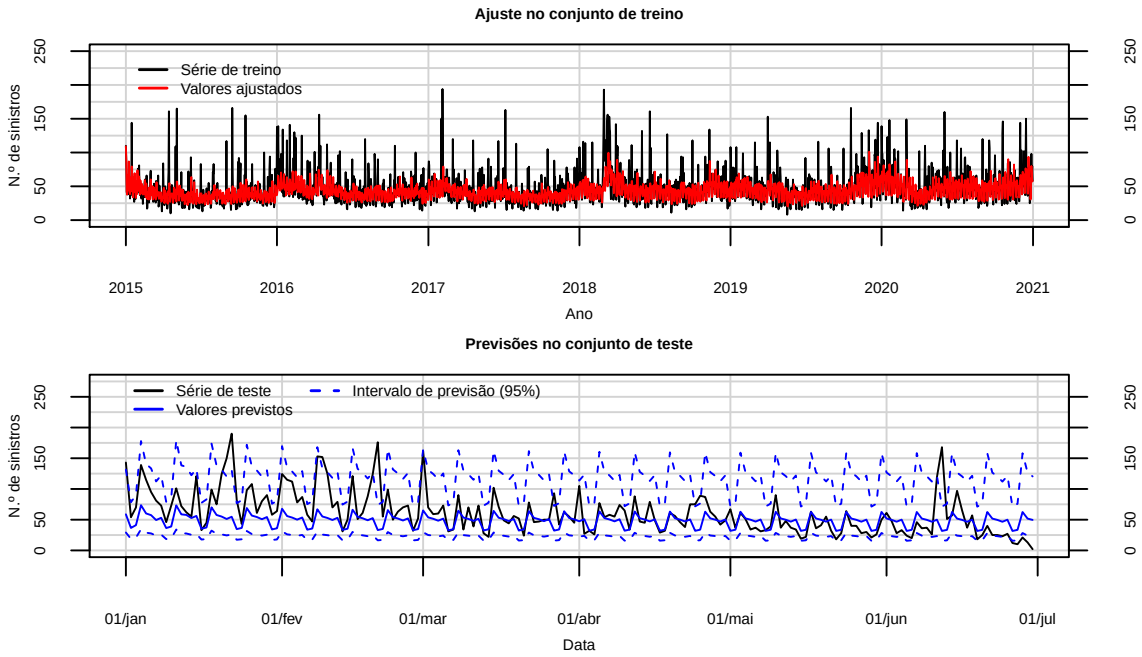


Figura 5.10: Ajuste do modelo SARIMA (no período de treino) e previsões intervalares a 95% e pontuais (no período de teste) sobrepostas à série diária do número total de sinistros.

Como seria de esperar após uma análise gráfica da Figura 5.10, as previsões pontuais não estão todas contidas nos intervalos de previsão de 95% de confiança e o ajuste do modelo é melhor na série treino do que na de teste, o que vai de acordo com os valores apresentados na Tabela 5.17. Os intervalos de previsão a 80% de confiança apresentam uma taxa de cobertura de 71,27% e os de 95% de confiança cobrem cerca de 91,71% dos valores observados. Na Tabela 5.5 são apresentadas as previsões pontuais e os respetivos intervalos de previsão de 80% e 95% de confiança para os primeiros 7 dias do conjunto de teste.

Tabela 5.5: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as primeiras 7 observações do conjunto de teste, relativos ao modelo SARIMA ajustado à série diária do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x}_i - x_i$
59	$(36,7763\ ;\ 98,1722)$	$(29,1410\ ;\ 131,5386)$	143	-84
37	$(23,4254\ ;\ 59,4876)$	$(18,7603\ ;\ 78,3934)$	54	-17
41	$(26,0253\ ;\ 67,8181)$	$(20,7278\ ;\ 90,1400)$	70	-29
74	$(44,2531\ ;\ 128,9266)$	$(34,4286\ ;\ 177,9318)$	139	-65
60	$(36,7957\ ;\ 103,2982)$	$(28,8489\ ;\ 140,7742)$	115	-55
57	$(35,2182\ ;\ 98,1688)$	$(27,6533\ ;\ 133,4620)$	95	-38
49	$(30,6757\ ;\ 83,2791)$	$(24,2206\ ;\ 112,2162)$	81	-32
Taxa de cobertura	71,3%	91,7%		$\sum_{i=1}^{7}(\widehat{x}_i - x_i) = -320$

Modelo Holt-Winters

A série diária do número total de sinistros é fortemente marcada por uma sazonalidade semanal, tal como visto no gráfico da sua FACP (Figura 5.3), tanto como uma sazonalidade anual, já referida anteriormente, visível na Figura 5.15. Este método permite estabelecer modelos com decomposição sazonal aditiva ou multiplicativa. Esta última serve para explicar sazonalidades cuja amplitude varia proporcionalmente com a tendência, enquanto que a decomposição aditiva é mais adequada para séries que apresentem uma amplitude sazonal aproximadamente constante ao longo do tempo. Através da análise do gráfico da série (Figura 5.15)[9], não é claro qual o modelo mais apropriado, pois a alta variabilidade da série não facilita a sua interpretação gráfica. Ora, serão então ajustados os modelos de Holt-Winters aditivo e multiplicativo.

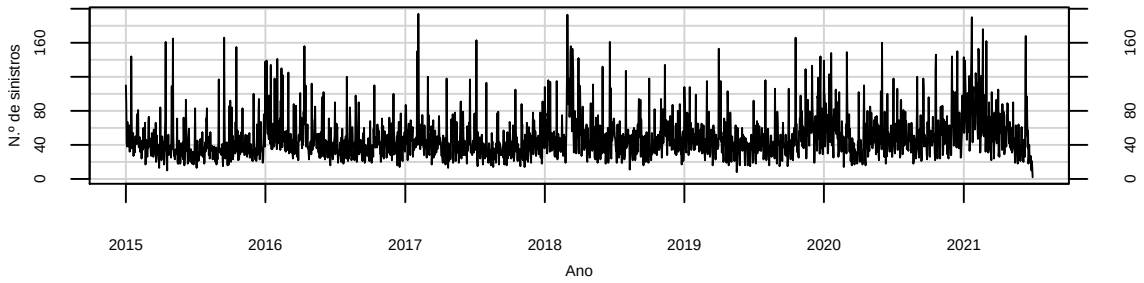


Figura 5.11: Representação gráfica da série do número total de sinistros.

Tabela 5.6: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, correspondentes ao modelo aditivo de Holt-Winters (considerando $s = 7$) ajustado ao conjunto de treino da série diária do número total de sinistros.

Modelo aditivo (considerando $s = 7$)				$REQM \approx 21{,}921$
$\widehat{\alpha} \approx 0{,}1220$	$\widehat{\beta} \approx 0{,}0107$	$\hat{\gamma} \approx 0{,}0374$	$\hat{l}_1 \approx 56{,}5890$	$\hat{b}_1 \approx 0{,}0727$
$\hat{s}_1 \approx 9{,}3101$	$\hat{s}_2 \approx -10{,}2287$	$\hat{s}_3 \approx -7{,}6926$	$\hat{s}_4 \approx 24{,}7635$	$\hat{s}_5 \approx 15{,}1836$
$\hat{s}_6 \approx 8{,}1457$	$\hat{s}_7 \approx 4{,}5189$			

Tabela 5.7: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, correspondentes ao modelo multiplicativo de Holt-Winters (considerando $s = 7$) ajustado ao conjunto de treino da série diária do número total de sinistros.

Modelo multiplicativo (considerando $s = 7$)				$REQM \approx 21{,}9496$
$\widehat{\alpha} \approx 0{,}1009$	$\widehat{\beta} \approx 0{,}0108$	$\hat{\gamma} \approx 0{,}0245$	$\hat{l}_1 \approx 58{,}7167$	$\hat{b}_1 \approx 0{,}0774$
$\hat{s}_1 \approx 1{,}1522$	$\hat{s}_2 \approx 0{,}7482$	$\hat{s}_3 \approx 0{,}7869$	$\hat{s}_4 \approx 1{,}4517$	$\hat{s}_5 \approx 1{,}2315$
$\hat{s}_6 \approx 1{,}139$	$\hat{s}_7 \approx 1{,}0727$			

[9] Dado que o método de Holt-Winters é não paramétrico, será considerada a série sem transformação de Box-Cox.

Tabela 5.8: Estimativas iniciais para o nível, o declive e os primeiros 10 fatores sazonais e estimativas das constantes de alisamento, correspondentes ao modelo aditivo de Holt-Winters (considerando $s = 365$) ajustado ao conjunto de treino da série diária do número total de sinistros.

Modelo aditivo (considerando $s = 365$)				$REQM \approx 22{,}5777$
$\widehat{\alpha} \approx 0{,}1199$	$\widehat{\beta} \approx 0$	$\hat{\gamma} \approx 0{,}4220$	$\hat{l}_1 \approx 71{,}2622$	$\hat{b}_1 \approx 0{,}0151$
$\hat{s}_1 \approx 65{,}6535$	$\hat{s}_2 \approx 33{,}0461$	$\hat{s}_3 \approx 25{,}9465$	$\hat{s}_4 \approx 7{,}4176$	$\hat{s}_5 \approx 4{,}3721$
$\hat{s}_6 \approx 13{,}0787$	$\hat{s}_7 \approx 9{,}1699$	$\hat{s}_8 \approx 13{,}0372$	$\hat{s}_9 \approx 15{,}9862$	$\hat{s}_{10} \approx 31{,}6958$

Tabela 5.9: Estimativas iniciais para o nível, o declive e os primeiros 10 fatores sazonais e estimativas das constantes de alisamento, correspondentes ao modelo multiplicativo de Holt-Winters (considerando $s = 365$) ajustado ao conjunto de treino da série diária do número total de sinistros.

Modelo multiplicativo (considerando $s = 365$)				$REQM \approx 25{,}2283$
$\widehat{\alpha} \approx 0$	$\widehat{\beta} \approx 0$	$\hat{\gamma} \approx 0{,}4654$	$\hat{l}_1 \approx 69{,}5379$	$\hat{b}_1 \approx 0{,}0151$
$\hat{s}_1 \approx 2{,}1080$	$\hat{s}_2 \approx 1{,}5501$	$\hat{s}_3 \approx 1{,}4143$	$\hat{s}_4 \approx 0{,}9865$	$\hat{s}_5 \approx 0{,}8848$
$\hat{s}_6 \approx 1{,}0429$	$\hat{s}_7 \approx 0{,}9794$	$\hat{s}_8 \approx 1{,}0037$	$\hat{s}_9 \approx 1{,}0589$	$\hat{s}_{10} \approx 1{,}3437$

Depois de efetuada a estimação dos quatro modelos (Tabelas 5.6, 5.7, 5.8 e 5.9) ao conjunto de treino da série diária do número total de sinistros, conclui-se que os modelos mais aptos a explicar a série, usando como critério o erro quadrático médio a 1-passo, são os que consideram sazonalidade semanal ($s = 7$). A escolha entre o modelo aditivo e o multiplicativo não é óbvia uma vez que os erros quadráticos médios a 1-passo diferenciam-se apenas por cerca de uma unidade. As Tabelas 5.6 e 5.7 mostram que todos os valores dos parâmetros são bastante próximos de zero em ambos os modelos. Isto significa que os valores previstos pelos modelos terão muito pouca variabilidade, pois oscilarão pouco em relação à amplitude original da série.

Ambos os modelos apresentam REQM relativamente altos, provavelmente justificado pelo facto de as previsões terem uma baixa amplitude de oscilação derivada dos baixos valores dos parâmetros. Na verdade, sendo esta uma série diária, está sujeita a uma alta variabilidade que não consegue ser explicada na totalidade por modelos estatísticos. Porém, o modelo multiplicativo é mais apropriado para casos onde a oscilação da série depende da sua tendência, característica que não está presente no gráfico desta série. Logo, de maneira a manter um critério de seleção para cada metodologia, dentro do método de Holt-Winters, escolhe-se o modelo aditivo para a modelação da série diária do número total de sinistros, pois é aquele que tem o menor REQM. Assim sendo, as equações correspondentes a este modelo são:

$$
\begin{aligned}
l_t &\approx 0,122(X_t - s_{t-7}) + 0,878(l_{t-1} + b_{t-1}) \\
b_t &\approx 0,0107(l_t - l_{t-1}) + 0,9893 b_{t-1} \\
s_t &\approx 0,0374(X_t - l_t) + 0,9626 s_{t-7} \\
\widehat{X}_{t+h} &\approx l_t + h b_t + s_{t-7+[(h-1) \bmod 7]+1}.
\end{aligned}
$$

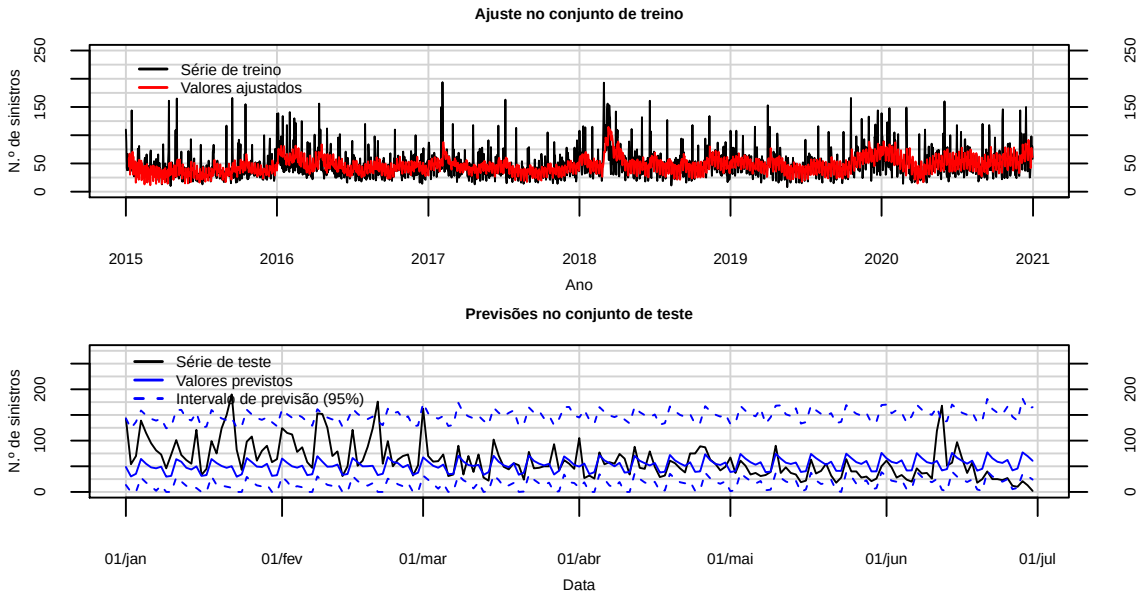


Figura 5.12: Ajuste do modelo aditivo de Holt-Winters (no período de treino) e previsões intervalares a 95% e pontuais (no período de teste) sobrepostas à série diária do número total de sinistros.

A Figura 5.12 ilustra o ajuste (a vermelho), as previsões intervalares a 95% de confiança (tracejado azul) e pontuais (linha contínua azul) ao número diário total de sinistros. Uma vez que o método de Holt-Winters é não paramétrico, os intervalos de previsão não podem ser construídos baseados na normalidade dos erros. Assim, os intervalos de previsão são construídos pelo método de Bootstrap explicado na Secção 4.3.4.

A Tabela 5.10 mostra que todos os intervalos de previsão com 95% de confiança contêm os respetivos valores observados, pois, estes são construídos de forma a serem conservadores, i.e., de forma a terem uma grande amplitude [Wang and Cai, 2009]. Ainda assim, a Figura 5.12 mostra que existem valores observados que não estão contidos nos respetivos intervalos de previsão.

Uma análise mensal mostra que os melhores meses previstos foram março e abril com $REQM \approx 25$ e $REQM \approx 24$, respetivamente, ou seja, um erro diário a rondar as 24 e 25 unidades nestes meses. Nota-se, claramente, que os restantes meses apresentam um $REQM$ superior, pois a Figura 5.12 ilustra bem as diferenças entre os valores previstos e os observados, chegando a atingir $REQM \approx 39$ em fevereiro e $REQM \approx 49$ em junho. Este último valor, provavelmente, é resultado da recentidade dos dados, uma vez que estes podem ainda não representar a totalidade dos sinistros registados neste mês, porque ainda não existe acesso a estes

dados.[10]

[10]Muitas vezes, o registo de sinistros ocorre longe da data das suas ocorrências.

Tabela 5.10: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as primeiras 7 observações do conjunto de teste, relativos ao modelo Holt-Winters ajustado à série diária do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x}_i - x_i$
49	(21,677100 ; 103,2493)	(13,5894 ; 143,2917)	143	-95
30	(0,919300 ; 87,1604)	(0 ; 123,8255)	54	-22
35	(4,9647 ; 89,5031)	(0 ; 136,4802)	70	-36
64	(37,213 ; 122,4720)	(27,7885 ; 158,6715)	139	-75
55	(27,988 ; 112,7908)	(18,2812 ; 149,4559)	115	-61
48	(21,0051 ; 105,8256)	(11,17870 ; 142,4907)	95	-48
46	(17,4510 ; 97,9205)	(2,2691 ; 138,9366)	81	-35
Taxa de cobertura	79,6%	93,9%		$\sum_{i=1}^{7} (\widehat{x}_i - x_i) = -372$

Modelo TBATS

Tal como nos modelos de Holt-Winters, no modelo TBATS serão abordados os casos de sazonalidade semanal ($m_1 = 7$) e anual ($m_1 = 365$). Apesar de o modelo TBATS conseguir incorporar sazonalidades complexas, i.e., pode modelar simultaneamente sazonalidade semanal e anual, estes casos terão que ser abordados em modelos diferentes, pois, no caso de sazonalidade anual ($m_1 = 365$), são retiradas observações necessárias para a sazonalidade semanal ($m_1 = 7$).[11] Este modelo é ajustado automaticamente, ou seja, não necessita que sejam atribuídos valores iniciais aos parâmetros. A função tbats fornece as estimativas de máxima verosimilhança dos valores de estados iniciais, estimativas dos parâmetros de alisamento, escolhe o número de harmónicos necessário para modelar a componente sazonal dos dados e os parâmetros p e q do processo ARMA. Os modelos que minimizam o AIC fornecidos pela função são: TBATS(1 , {2,4} , – , {7,3}) para a sazonalidade semanal e TBATS(1 , {4,2} , 0,8 , {365,1}) para a anual. O parâmetro $\omega = 1$ indica que não foi efetuada nenhuma transformação de Box-Cox à série, pois estes modelos foram ajustados ao conjunto de treino com a transformação de Box-Cox efetuada anteriormente (λ = -0,184). Esta transformação é necessária porque esta metodologia supõe que $\varepsilon_i \sim N(0, \sigma^2)$, tal como visto na Secção 4.4.2. O facto de o modelo com $m_1 = 7$ não incorporar β nem ϕ, implica que o modelo não possui parâmetros de amortecimento, ou seja, a série não necessita de ser suavizada. Os seus erros são modelados por um processo ARMA(2,4), o que significa que, apesar de ter sido efetuada uma transformação de Box-Cox, os erros continuam a ter correlações significativas.

Pela análise dos parâmetros dos modelos, nota-se claramente que o modelo considerando $m_1 = 7$ é mais parcimonioso, pois o modelo assumindo sazonalidade anual ($m_1 = 365$) necessita de parâmetros de amortecimento e apresenta fortes correlações temporais nos erros, dado que estes são modelados por um processo ARMA(4,2). Além disso, o primeiro modelo apresenta um valor do AIC = 9370,0951 que é inferior ao do segundo modelo (AIC = 9957,0189). Logo, dada a parcimoniosidade e o menor valor do AIC, escolhe-se o modelo com sazonalidade semanal ($m_1 = 7$) para modelar o número total diário de sinistros. As estimativas dos seus parâmetros encontram-se na Tabela 5.11.

Tabela 5.11: Parâmetros do modelo TBATS ajustado à série diária do número total de sinistros.

	TBATS(1 , {2,4} , – , {7 , 3})					$\widehat{\sigma} \approx$ **0,1792**		$AIC \approx$ **9370,0951**	
Parâmetro	α	γ_1	γ_2	ϕ_1	ϕ_2	θ_1	θ_2	θ_3	θ_4
Estimativa	0,0548	0,0001	-0,0001	-1,7641	-0,9803	2,0272	1,5192	0,4076	0,0840

[11] Foi também ajustado um modelo com $m_1 = 7$ e $m_2 = 365,25$ cujas características estão expostas na Tabela C.1 do Apêndice C

Desta forma, as equações do modelo TBATS para o número total diário de sinistros são escritas da seguinte forma:

$$
\begin{aligned}
x_t^{(\lambda)} &= l_{t-1} + s_{t-7} + d_t \\
l_t &= l_{t-1} + 0{,}0548 d_t \\
d_t &= 1{,}7641 d_{t-1} - 0{,}9803 d_{t-2} + \Theta \\
\Theta &= 2{,}0272 \varepsilon_{t-1} + 1{,}5192 \varepsilon_{t-2} + 0{,}4076 \varepsilon_{t-3} + 0{,}084 \varepsilon_{t-4} + \varepsilon_t \\
s_t &= \sum_{j=1}^{3} s_{j,t} \\
s_{j,t} &= s_{j,t-1} cos\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1} sen\left(\frac{2\pi j}{7}\right) + 0{,}0001 d_t \\
s^*_{j,t} &= -s_{j,t-1} sen\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1} cos\left(\frac{2\pi j}{7}\right) - 0{,}0001 d_t
\end{aligned}
$$

onde foram utilizados 7 valores iniciais para $s_{j,0}$ e $s^*_{j,0}$.

Na Figura 5.13 apresentam-se o ajuste (a vermelho), as previsões intervalares a 95% de confiança (tracejado azul) e pontuais (linha contínua azul) ao número diário total de sinistros.

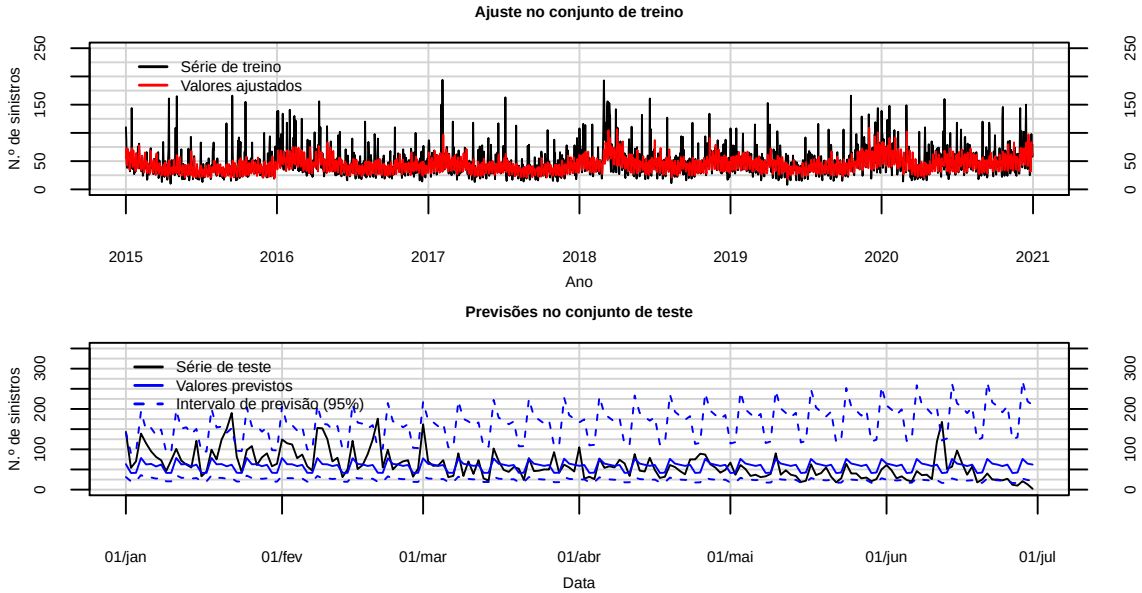


Figura 5.13: Ajuste do modelo TBATS (no período de treino) e previsões intervalares a 95% e pontuais (no período de teste) sobrepostas à série diária do número total de sinistros.

A Tabela 5.12 mostra os valores das previsões pontuais e os respetivos intervalos de previsão a 80% e 95% de confiança dos primeiros 7 dias do conjunto de teste.

Por fim, é necessário efetuar uma análise de resíduos de forma a validar o modelo escolhido. Tal como na metodologia SARIMA, os resíduos do modelo devem apresentar uma distribuição Gaussiana, com média nula, variância constante e sem correlações temporais. Tanto o gráfico em

Tabela 5.12: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as primeiras 7 observações do conjunto de teste, relativos ao modelo TBATS ajustado à série diária do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x}_i - x_i$
63	(39,4055 ; 105,8179)	(31,1851 ; 142,0653)	143	-80
41	(26,0696 ; 67,9638)	(20,7612 ; 90,3470)	54	-13
42	(26,1894 ; 69,0111)	(20,8092 ; 92,0669)	70	-28
79	(46,7023 ; 139,4366)	(36,1502 ; 194,0603)	139	-60
62	(37,8905 ; 108,2497)	(29,5988 ; 148,4013)	115	-53
64	(38,5967 ; 111,1870)	(30,0984 ; 152,8599)	95	-31
57	(34,9899 ; 98,8992)	(27,3940 ; 135,0906)	81	-24
Taxa de cobertura	75,7%	94,5%		$\sum_{i=1}^{7} (\widehat{x}_i - x_i) = -289$

função do tempo como o histograma da Figura 5.14 evidenciam uma distribuição Normal, com média nula e variância constante para os resíduos do modelo TBATS. Contudo, tanto o teste de normalidade de Kolmogorov-Smirnov (com correção de Lilliefors) como o teste de dependência temporal de Ljung-Box[12] rejeitam as suas hipósteses nulas. Isto significa que as previsões efetuadas pelo modelo não explicam toda a variabilidade dos dados.

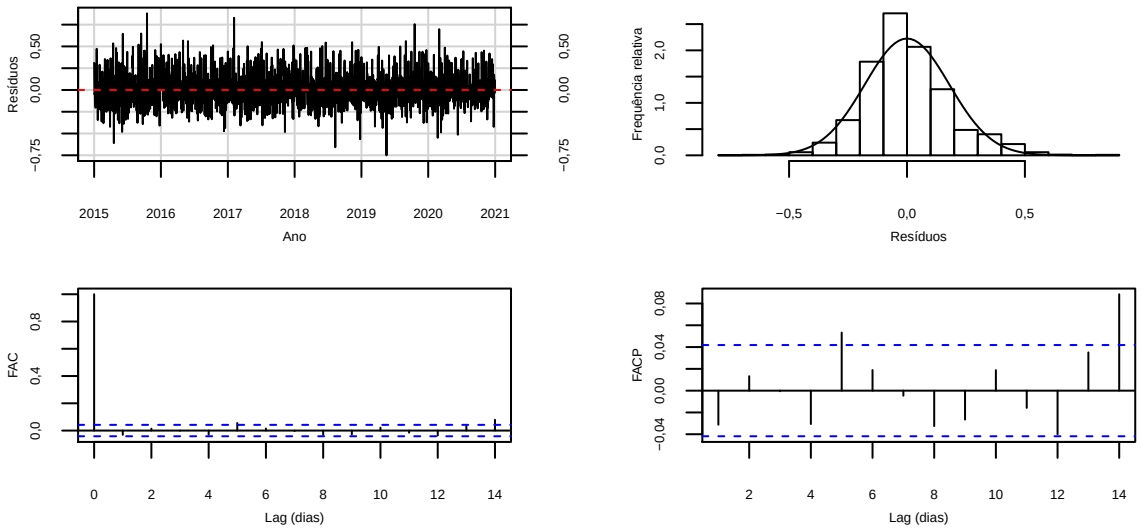


Figura 5.14: Série dos resíduos do modelo TBATS ajustado à série diária da categoria **total** após uma transformação de Box-Cox e respetivo histograma, FAC e FACP estimadas.

5.2.2 Caso II: Séries diárias marginais dos números de sinistros

Como referido anteriormente, esta secção aborda o processo de modelação das séries diárias marginais: **DNA**, **REL**, **TMP** e **restantes causas**. Relembrando as informações descritas no Caso I, os dados estão divididos em conjuntos de treino e de teste, com o intuito de se conseguir avaliar o ajustamento e a capacidade preditiva dos modelos. O conjunto de treino corresponde às primeiras 2192 observações e o conjunto de teste às restantes 181; novamente, representando cerca de 92% e 8% da totalidade dos dados, respetivamente. Todos os modelos doravante serão ajustados ao conjunto de treino e avaliados, principalmente, pelo seu poder preditivo no conjunto de teste.

Modelo SARIMA

Comparando a Figura C.2, do Apêndice C, e a Figura 5.15, verifica-se a ausência de estacionariedade nas séries marginais, tanto na média como na variância. Logo, para as etapas de modelação

[12] O teste foi efetuado com $m = 35$.

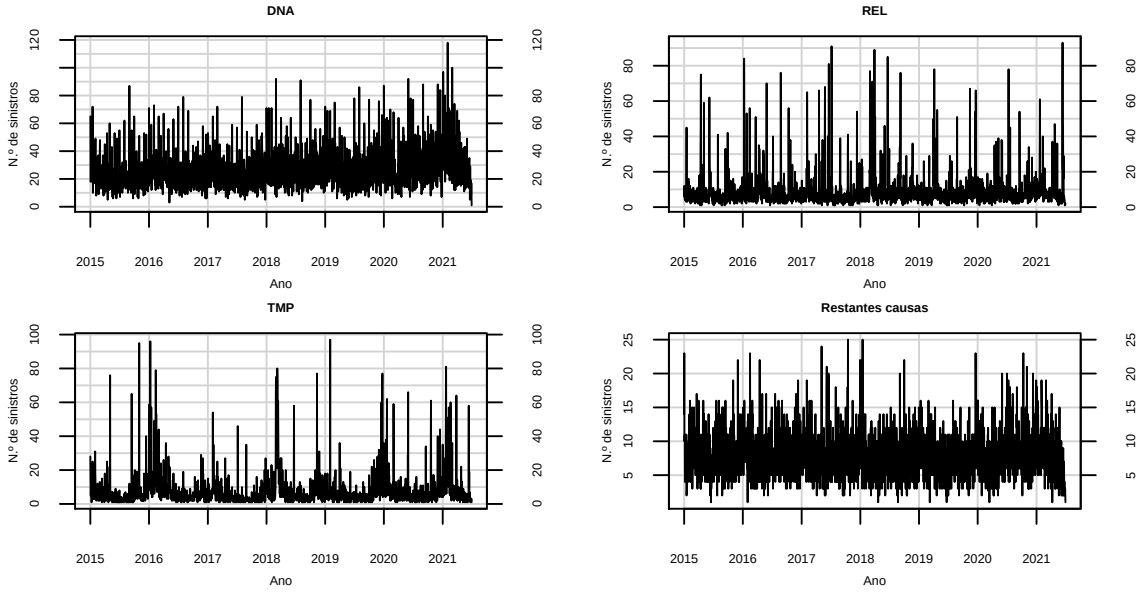


Figura 5.15: Representação gráfica das séries diárias marginais.

seguintes, será necessário considerar as séries com transformações de Box-Cox. Para este efeito, foram testados vários valores para λ e concluiu-se que os melhores valores deste parâmetro são $\lambda = 0{,}123$ para a categoria **DNA**, $\lambda =$ -0,160 para a categoria **REL**, $\lambda =$ -0,134 para a categoria **TMP** e $\lambda = 0{,}335$ para a categoria **restantes causas**.

Uma vez estabilizada a variância, recorrem-se aos testes de estacionariedade ADF e KPSS (mais uma vez, o número de lags é baseado na regra de Ng & Perron [Ng and Perron, 1995] apresentada no Capítulo 3). A Tabela 5.13 resume os testes de estacionariedade, onde se verifica que a estacionariedade não é rejeitada (com 95% de confiança) em todas as séries. Logo, o estudo procederá admitindo estacionariedade em todos os casos.

Tabela 5.13: Resumo dos testes ADF e KPSS para as séries diárias marginais.

Teste de hipóteses	Categoria	Estatística de teste	Número de lags usado
ADF	**DNA**	19,2482	23
	REL	20,7103	21
	TMP	7,8344	21
	Restantes causas	20,4994	20
KPSS	**DNA**	0,0561	23
	REL	0,2683	21
	TMP	0,1137	21
	Restantes causas	0,2508	20

De acordo com a Nota 7 (página 64), será apenas considerado o caso de sazonalidade semanal. Assim, serão ajustados vários modelos para todas as séries marginais, tal como explicado no Caso I.

A Tabela 5.14 mostra os modelos SARIMA escolhidos e as respetivas características para as modelações das séries diárias marginais, tendo em conta os menores valores do AIC e do BIC, tal como foi feito no caso da série **total**.

Tabela 5.14: Características dos modelos SARIMA escolhidos para modelar as séries diárias marginais com transformação de Box-Cox.

DNA	SARIMA$(2,0,1)(2,1,1)_7$	AIC $\approx 3541,68$		BIC $\approx 3581,50$		$\widehat{\sigma} \approx 0,54$	
	Parâmetro	ϕ_1	ϕ_2	θ_1	ν_1	ν_2	η_1
	Estimativa	1,1250	-0,1362	-0,9551	-0,0098	0,1279	-0,9850
	Erro padrão	0,0247	0,0228	0,0111	0,0226	0,0221	0,0043
REL	SARIMA$(2,0,1)(1,0,1)_7$	AIC $\approx 2859,29$		BIC $\approx 2859,34$		$\widehat{\sigma} \approx 0,46$	
	Parâmetro	ϕ_1	ϕ_2	θ_1	ν_1		η_1
	Estimativa	1,1630	-0,2250	-0,8539	0,9983		-0,9884
	Erro padrão	0,0585	0,0346	0,0520	0,0017		0,0058
TMP	SARIMA$(2,0,1)(2,1,1)_7$	AIC $\approx 3712,19$		BIC $\approx 3752,01$		$\widehat{\sigma} \approx 0,56$	
	Parâmetro	ϕ_1	ϕ_2	θ_1	ν_1	ν_2	η_1
	Estimativa	1,1708	-0,1883	-0,8790	0,0070	0,0729	-1
	Erro padrão	0,0326	0,0289	0,0222	0,0223	0,0218	0,0071
Restantes causas	SARIMA$(1,0,1)(0,1,1)_7$	AIC $\approx 5438,14$		BIC $\approx 5460,90$		$\widehat{\sigma} \approx 0,83$	
	Parâmetro	ϕ_1		θ_1			η_1
	Estimativa	0,9358		-0,8773			-0,9896
	Erro padrão	0,0269		0,0364			0,0046

Os testes de Kolmogorov-Smirnov (com correção de Lilliefors) e Shapiro-Wilk rejeitam as hipóteses de normalidade presente nas séries dos resíduos. Porém, o teste de Ljung-Box permite assumir que os resíduos dos modelos ajustados são temporalmente independentes (com confiança de 95%) nos casos das categorias **REL** e **restantes causas**. Este argumento pode ser visualizado na Figura C.6 onde estão ilustradas as bandas de Bartlett nos gráficos das FAC e FACP das várias categorias. As Figuras C.4 e C.5 (do Apêndice C) mostram o ajuste e as previsões (pontuais e intervalares a 95% de confiança) destes modelos às respetivas séries diárias.

Modelo Holt-Winters

Tal como no Caso I, as séries diárias marginais serão analisadas estabelecendo-se modelos Holt-Winters assumindo sazonalidade semanal (Figura 5.3). Após serem estudados os modelos aditivos e multiplicativos de cada categoria, conclui-se que, uma vez mais, os modelos aditivos são os que apresentam melhores valores das medidas de ajustamento. A Tabela 5.15 mostra as estimativas iniciais para os níveis, os declives e os fatores sazonais e as estimativas das constantes de alisamento destes modelos.

Tabela 5.15: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, correspondentes aos modelos de Holt-Winters aditivos ajustados aos conjuntos de treino da séries diárias marginais.

DNA	$REQM \approx$ 10,97				
	$\widehat{\alpha} \approx$ 0,0508	$\widehat{\beta} \approx$ 0,0227	$\hat{\gamma} \approx$ 0,0546	$\hat{l}_1 \approx$ 31,4474	$\hat{b}_1 \approx$ 0,0111
	$\hat{s}_1 \approx$ 3,6475	$\hat{s}_2 \approx$ -9,3714	$\hat{s}_3 \approx$ -6,5090	$\hat{s}_4 \approx$ 17,5467	$\hat{s}_5 \approx$ 10,8161
	$\hat{s}_6 \approx$ 7,6074	$\hat{s}_7 \approx$ 2,4846			
REL	$REQM \approx$ 9,6033				
	$\widehat{\alpha} \approx$ 0,1422	$\widehat{\beta} \approx$ 0,0014	$\hat{\gamma} \approx$ 0,0187	$\hat{l}_1 \approx$ 9,3983	$\hat{b}_1 \approx$ -0,0024
	$\hat{s}_1 \approx$ 2,8274	$\hat{s}_2 \approx$ 0,8309	$\hat{s}_3 \approx$ -1,3029	$\hat{s}_4 \approx$ 1,8567	$\hat{s}_5 \approx$ 1,0829
	$\hat{s}_6 \approx$ 1,0077	$\hat{s}_7 \approx$ 1,5903			
TMP	$REQM \approx$ 8,3059				
	$\widehat{\alpha} \approx$ 0,2248	$\widehat{\beta} \approx$ 0,0056	$\hat{\gamma} \approx$ 0,0347	$\hat{l}_1 \approx$ 8,6236	$\hat{b}_1 \approx$ 0,0161
	$\hat{s}_1 \approx$ 2,8452	$\hat{s}_2 \approx$ -0,8452	$\hat{s}_3 \approx$ 2,0463	$\hat{s}_4 \approx$ 4,2548	$\hat{s}_5 \approx$ 3,9478
	$\hat{s}_6 \approx$ 0,8481	$\hat{s}_7 \approx$ 1,0193			
Restantes causas	$REQM \approx$ 3,439				
	$\widehat{\alpha} \approx$ 0,0932	$\widehat{\beta} \approx$ 0,0215	$\hat{\gamma} \approx$ 0,0390	$\hat{l}_1 \approx$ 10,4949	$\hat{b}_1 \approx$ 0,0560
	$\hat{s}_1 \approx$ 1,3080	$\hat{s}_2 \approx$ -0,5170	$\hat{s}_3 \approx$ 0,2312	$\hat{s}_4 \approx$ 2,8820	$\hat{s}_5 \approx$ 1,5401
	$\hat{s}_6 \approx$ 1,1879	$\hat{s}_7 \approx$ 0,7190			

Sendo estes dados diários, os baixos valores dos parâmteros voltam a surgir. Estes irão refletir-se nas previsões pontuais que terão baixa variabilidade, tal como indicam as equações do modelo Holt-Winters:

DNA:

$$\begin{aligned} l_t &\approx 0,0508(X_t - s_{t-7}) + 0,9492(l_{t-1} + b_{t-1}) \\ b_t &\approx 0,0227(l_t - l_{t-1}) + 0,9773 b_{t-1} \\ s_t &\approx 0,0546(X_t - l_t) + 0,9454 s_{t-7} \\ \widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-7+[(h-1) \bmod 7]+1} \end{aligned}$$

REL:

$$\begin{aligned} l_t &\approx 0,1422(X_t - s_{t-7}) + 0,8578(l_{t-1} + b_{t-1}) \\ b_t &\approx 0,0014(l_t - l_{t-1}) + 0,9986 b_{t-1} \\ s_t &\approx 0,0187(X_t - l_t) + 0,9813 s_{t-7} \\ \widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-7+[(h-1) \bmod 7]+1} \end{aligned}$$

TMP:

$$\begin{aligned} l_t &\approx 0,2248(X_t - s_{t-7}) + 0,7752(l_{t-1} + b_{t-1}) \\ b_t &\approx 0,0056(l_t - l_{t-1}) + 0,9944 b_{t-1} \\ s_t &\approx 0,0347(X_t - l_t) + 0,9653 s_{t-7} \\ \widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-7+[(h-1) \bmod 7]+1} \end{aligned}$$

Restantes causas:

$$\begin{aligned} l_t &\approx 0,0932(X_t - s_{t-7}) + 0,9068(l_{t-1} + b_{t-1}) \\ b_t &\approx 0,0215(l_t - l_{t-1}) + 0,0227 b_{t-1} \\ s_t &\approx 0,039(X_t - l_t) + 0,961 s_{t-7} \\ \widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-7+[(h-1) \bmod 7]+1} \end{aligned}$$

O ajuste e as previsões destes modelos podem ser consultados nas Figuras C.7 e C.8 do Apêndice C.

Modelo TBATS

A metodologia TBATS é efetuada automaticamente, logo não é necessário qualquer cálculo ou estudo prévio para além da sazonalidade. Assim sendo, usando as mesmas constantes λ de Box-Cox que os na modelação via SARIMA, os modelos ajustados são:

- TBATS$(0,123\ ,\ \{5,3\},\ -\ ,\ \{7,3\})$, para a categoria **DNA**;
- TBATS$(-0,16\ ,\ \{2,2\},\ 0,909\ ,\ \{7,3\})$, para a categoria **REL**;
- TBATS$(-0,134\ ,\ \{2,2\},\ 0,829\ ,\ \{7,3\})$, para a categoria **TMP**;
- TBATS$(0,335\ ,\ \{0,0\},\ 0,876\ ,\ \{7,3\})$, para a categoria **restantes causas**.

Analisando estes modelos, conclui-se que a categoria **restantes causas** é a única série cujos erros são independentes. Nos restantes casos são aplicados processos ARMA para modelar os erros; em particular a categoria **DNA** admite erros com fortes correlações uma vez que se estimou um modelo ARMA com 5 coeficientes autorregressivos. A Tabela 5.16 mostra as estimativas dos parâmetros destes modelos.

Tabela 5.16: Parâmetros dos modelos TBATS ajustados às séries diárias marginais.

DNA	**TBATS(0,123 , {5,3} , - , {7,3})**					$\widehat{\sigma} \approx$ **0,537**		$AIC \approx$ **14188,7137**	
Parâmetro	α	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_2	θ_3
Estimativa	0,0798	-0,7232	0,7297	0,7401	-0,2264	-0,0574	0,8470	-0,6377	-0,8735
REL	**TBATS(-0,16 , {2,2} , 0,909 , {7,3})**					$\widehat{\sigma} \approx$ **0,4611**		$AIC \approx$ **13510,678**	
Parâmetro	α	β	ϕ	γ_1	ϕ_1	ϕ_2	θ_1	θ_2	
Estimativa	0,0805	-0,0075	0,9094	0,0002	-0,6430	0,2131	0,8814	0,0121	
TMP	**TBATS(-0,134 , {2,2} , 0,829 , {7,3})**					$\widehat{\sigma} \approx$ **0,5599**		$AIC \approx$ **14361,3818**	
Parâmetro	α	β	ϕ	γ_1	γ_2	ϕ_1	ϕ_2	θ_1	θ_2
Estimativa	0,1980	-0,0224	0,8291	0,0008	-0,0005	0,4402	-0,6654	-0,3449	0,6485
Restantes causas	**TBATS(0,335 , {0,0} , 0,876 , {7,3})**					$\widehat{\sigma} \approx$ **0,8254**		$AIC \approx$ **16046,8659**	
Parâmetro	α	β	ϕ						
Estimativa	0,0785	-0,0124	0,8759						

As equações dos modelos são então:

DNA:

$$
\begin{aligned}
x_t^{(0,123)} &= l_{t-1} + s_{t-7} + d_t \\
l_t &= l_{t-1} + 0{,}0798 d_t \\
d_t &= -0{,}7232 d_{t-1} + 0{,}7297 d_{t-2} + 0{,}7401 d_{t-3} - 0{,}2264 d_{t-4} - 0{,}0574 d_{t-5} + \Theta \\
\Theta &= 0{,}847 \varepsilon_{t-1} - 0{,}6377 \varepsilon_{t-2} - 0{,}8735 \varepsilon_{t-3} + \varepsilon_t \\
s_t &= \sum_{j=1}^{3} s_{j,t} \\
s_{j,t} &= s_{j,t-1} cos\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1} sen\left(\frac{2\pi j}{7}\right) \\
s^*_{j,t} &= -s_{j,t-1} sen\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1} cos\left(\frac{2\pi j}{7}\right)
\end{aligned}
$$

REL:

$$
\begin{aligned}
x_t^{(-0,16)} &= l_{t-1} + 0,9094b_{t-1} + s_{t-7} + d_t \\
l_t &= l_{t-1} + 0,9094b_{t-1} + 0{,}0805d_t \\
b_t &= 0,9094b_{t-1} - 0,0075d_t \\
d_t &= -0{,}643d_{t-1} + 0{,}2131d_{t-2} + \Theta \\
\Theta &= 0{,}8814\varepsilon_{t-1} - 0{,}0121\varepsilon_{t-2} + \varepsilon_t \\
s_t &= \sum_{j=1}^{3} s_{j,t} \\
s_{j,t} &= s_{j,t-1}cos\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1}sen\left(\frac{2\pi j}{7}\right) + 0,0002d_t \\
s^*_{j,t} &= -s_{j,t-1}sen\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1}cos\left(\frac{2\pi j}{7}\right)
\end{aligned}
$$

TMP:

$$
\begin{aligned}
x_t^{(-0,134)} &= l_{t-1} + 0,8291b_{t-1} + s_{t-7} + d_t \\
l_t &= l_{t-1} + 0,8291b_{t-1} + 0{,}198d_t \\
b_t &= 0,8291b_{t-1} - 0,0224d_t \\
d_t &= -0{,}4402d_{t-1} - 0{,}6654d_{t-2} + \Theta \\
\Theta &= 0{,}3449\varepsilon_{t-1} - 0{,}6485\varepsilon_{t-2} + \varepsilon_t \\
s_t &= \sum_{j=1}^{3} s_{j,t} \\
s_{j,t} &= s_{j,t-1}cos\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1}sen\left(\frac{2\pi j}{7}\right) + 0,0008d_t \\
s^*_{j,t} &= -s_{j,t-1}sen\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1}cos\left(\frac{2\pi j}{7}\right) - 0,0005d_t
\end{aligned}
$$

Restantes causas:

$$
\begin{aligned}
x_t^{(0,335)} &= l_{t-1} + 0,8759 b_{t-1} + s_{t-7} + \varepsilon_t \\
l_t &= l_{t-1} + 0,8759 b_{t-1} + 0{,}0785 \varepsilon_t \\
b_t &= 0,8759 b_{t-1} - 0,0124 \varepsilon_t \\
s_t &= \sum_{j=1}^{3} s_{j,t} \\
s_{j,t} &= s_{j,t-1} cos\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1} sen\left(\frac{2\pi j}{7}\right) \\
s^*_{j,t} &= -s_{j,t-1} sen\left(\frac{2\pi j}{7}\right) + s^*_{j,t-1} cos\left(\frac{2\pi j}{7}\right)
\end{aligned}
$$

Apenas os resíduos do modelo da categoria **DNA** rejeita (com 95% de confiança) a hipótese nula de autocorrelação temporal do teste de Ljung-Box, ou seja, apesar de os resíduos não serem Gaussianos, pode-se assumir que, com exceção da categoria **DNA**, estes são temporalmente independentes.

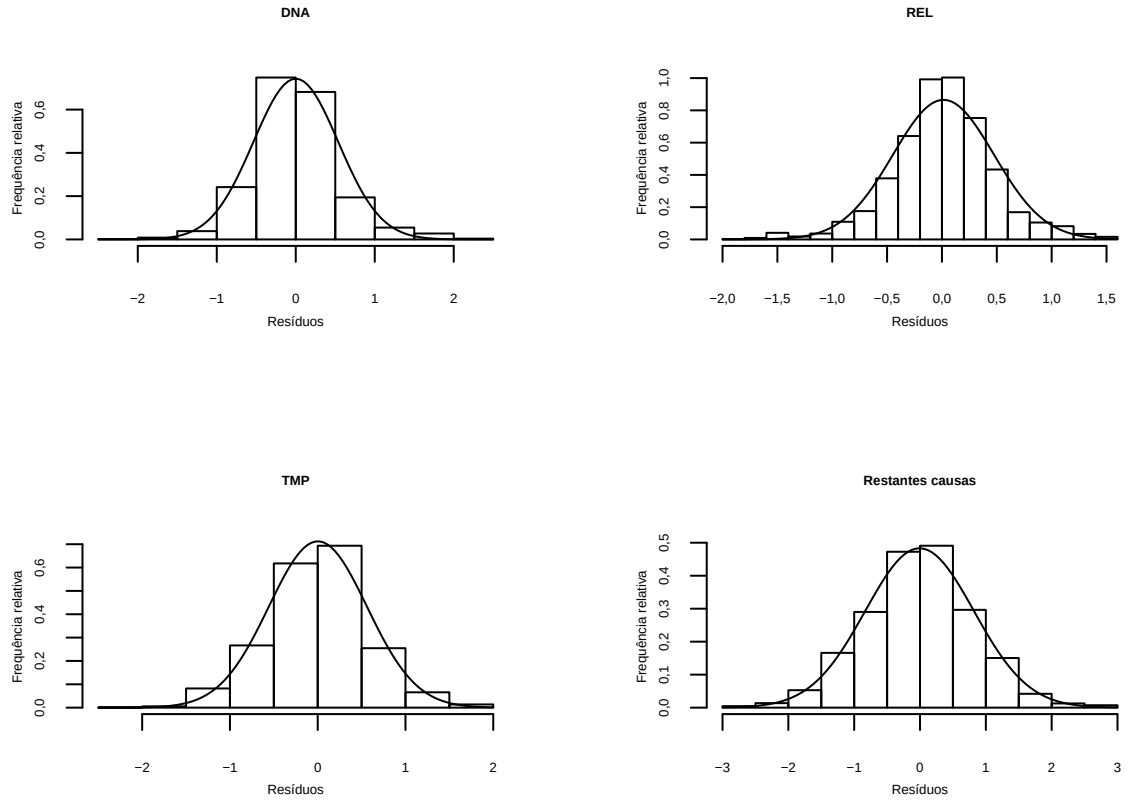


Figura 5.16: Histogramas dos resíduos dos modelos TBATS ajustados às séries marginais.

A Figura 5.16 ilustra as distribuições empíricas dos resíduos destes modelos. Partindo de uma

interpretação gráfica e tranquilizando os pressupostos de normalidade $\varepsilon_t \sim N(0, \sigma^2_{\varepsilon_t})$ necessários para a aplicação da metodologia TBATS, é aceitável assumir normalidade nestes casos, uma vez que as distribuições aparentam uma simetria forte em torno do valor nulo e a curva representativa da função de densidade empírica tende a estar ao nível das barras do histograma, o que apoia a normalidade das séries de resíduos. Assim sendo, é possível visualizar os ajustes e as previsões pontuais e intervalares (a 95% de confiança) nas Figuras C.9 e C.10 do Apêndice C.

5.3 Comparação dos métodos de previsão aplicados aos dados diários

Depois de aplicados os três métodos de modelação de séries temporais aos dados diários, resta comparar os modelos escolhidos por cada metodologia. Tal como referido na Secção 4.5, não existe nenhuma métrica consensual que possa quantificar a qualidade total de todos os modelos. São, então, utilizados nesta dissertação cinco critérios para os comparar:

- Erro quadrático médio (EQM), medida que depende da escala;
- Raiz do erro quadrático médio (REQM), medida que corresponde ao EQM na escala dos dados;
- Erro percentual absoluto médio (EPAM), medida percentual;
- Erro escalado absoluto médio (EEAM), medida escalada;
- U de Theil, medida alternativa;
- Soma das diferenças entre os valores estimados e os observados.

Os resultados destas medidas estão na Tabela 5.17 de forma a facilitar as suas comparações. Conclui-se que o modelo Holt-Winters não é eficaz na modelação destas séries, pois apenas apresenta três menores valores destes critérios de avaliação. Por outro lado, os modelos TBATS e os SARIMA apresentam qualidades semelhantes. Contudo, uma vez que o foco principal do estudo está presente na previsão e não no ajuste das séries, os modelos SARIMA são os mais indicados para este fim. Este resultado vai de encontro a Alon [Alon et al., 2001], que afirma que o método de Box-Jenkins é o mais adequado.

Estes resultados mostram, também, a vantagem que a metodologia TBATS possui em relação aos restantes modelos, já que esta foi desenvolvida propositadamente para dados com alta variabilidade, reconhecendo componentes trignométricas para modelar as sazonalidades das séries.

Tabela 5.17: Medidas de avaliação calculadas para as séries diárias, no período de treino e no período de teste, baseadas nos resultados da aplicação das três metodologias estudadas.

		Série de treino					**Série de teste**					
Categoria	**Modelo**	EQM	REQM	EPAM	EEAM	U-Theil	EQM	REQM	EPAM	EEAM	U-Theil	$\sum_{i=1}^{181}(\widehat{x}_i - x_i)$
Total	SARIMA	449,2987	21,1967	**26,4595**	**0,6544**	0,7373	1179,2000	34,3395	**50,0740**	**0,9159**	**0,9779**	-2386
	Holt-Winters	480,5283	21,9210	31,7584	0,7299	0,7636	1318,8530	36,3160	65,4806	1,0504	1,2015	4785
	TBATS	**442,8070**	**21,0430**	26,7745	0,6608	**0,7328**	**1080,3840**	**32,8692**	63,9648	0,9564	1,1454	**-600**
DNA	SARIMA	111,7706	10,5722	**28,2313**	**0,6624**	0,6173	248,1414	15,7525	**49,6437**	**1,0564**	**0,7359**	-919
	Holt-Winters	120,3412	10,9700	32,2514	0,7161	0,6443	254,2427	15,9450	58,0871	1,1229	0,8092	-5315
	TBATS	**110,3535**	**10,5049**	28,5880	0,6696	**0,6130**	**238,5400**	**15,4447**	56,0802	1,0747	0,7822	**-378**
REL	SARIMA	89,4743	9,4591	51,7080	0,6190	0,8496	154,6774	12,4369	**60,2786**	0,6290	**1,0639**	-716
	Holt-Winters	92,2227	9,6033	76,7981	0,7442	0,8788	**147,7017**	**12,1533**	64,2253	**0,6208**	1,0654	-9767
	TBATS	**89,2910**	**9,4494**	**50,7458**	**0,6175**	**0,8489**	153,9987	12,4096	61,5137	0,6271	1,0817	**-688**
TMP	SARIMA	68,8630	8,2984	**58,9639**	**0,6403**	**0,8619**	208,0338	14,4234	**67,0620**	**0,8037**	**0,9103**	-1037
	Holt-Winters	68,9886	8,3059	84,8074	0,7615	0,9429	214,6084	14,6495	108,2296	0,9021	1,1386	-9884
	TBATS	**68,2128**	**8,2591**	59,0029	0,6423	0,8641	**190,5068**	**13,8024**	126,7087	0,8868	1,2510	**-446**
Restantes causas	SARIMA	11,0775	3,3283	39,5934	0,7131	0,6465	**12,2892**	**3,5056**	**55,2404**	**0,8688**	**0,8547**	-44
	Holt-Winters	11,8265	3,4390	42,8487	0,7560	0,6665	31,8011	5,6392	94,3882	1,4264	1,4774	-9663
	TBATS	**10,9428**	**3,3080**	**39,0959**	**0,7094**	**0,6409**	12,7390	3,5692	59,9322	0,8945	0,9064	**31**

Uma vez avaliadas as previsões pontuais, resta avaliar as previsões intervalares. Teoricamente, os intervalos de previsão são calculados com 95% de confiança, i.e., 95% dos intervalos devem conter o valor observado. É preciso ter em atenção que os intervalos de previsão podem ter uma amplitude elevada, dependendo da metodologia utilizada, que é o caso dos modelos Holt-Winters. Por este motivo, a análise deve ser efetuada de uma forma global e não apenas tendo em conta as previsões pontuais nem apenas as previsões intervalares.

Na Tabela 5.18 estão expostas as taxas de cobertura dos intervalos de previsão dos três métodos de previsão. É possível verificar que os intervalos produzidos pelos modelos de Holt-Winters apresentam uma alta taxa de cobertura. Isto é resultado dos seus métodos de construção, que não necessitam dos pré-requisitos de normalidade obrigatórios para as outras duas metodologias.

Tabela 5.18: Taxas de cobertura (%) dos intervalos de previsão a 95% de confiança das séries diárias.

		SARIMA	Holt-Winters	TBATS
Categoria	**Total**	91,7	94,5	94,5
	DNA	92,3	94,5	91,7
	REL	89,5	93,9	95,0
	TMP	92,8	94,5	100
	Restantes causas	91,2	94,5	98,9
	Média	91,5	94,4	96,4

Em suma, tendo em conta a precisão pontual de previsão, o modelo TBATS parece ser mais adequado, porém, as previsões intervalares de Holt-Winters também apresentam boas taxas de cobertura.

Capítulo 6

Aplicação das metodologias de previsão aos dados mensais

No mercado das seguradoras, é sempre importante ter em conta os números de sinistros registados. No Capítulo 5 foram estudados dados diários, mas estes apresentam uma alta variabilidade. Por este motivo, de forma a efetuar um estudo mais conclusivo, a modelação no espectro mensal é fulcral. Apesar de não ter informação diária, as séries mensais tratadas neste capítulo não são tão suscetíveis a alterações derivadas de valores diários outliers, o que torna a sua modelação mais fácil.

6.1 Análise descritiva dos dados mensais

Os dados mensais estão divididos em dois conjuntos: contagens brutas dos números de sinistros e taxas de frequência mensal. Estas taxas correspondem ao quociente

$$\frac{\text{n.º de sinistros registados no mês}}{\text{n.º de habitações expostas ao risco no mês}} \tag{6.1}$$

que varia entre 0 e 1, onde o número de pessoas expostas ao risco num dado mês equivale ao número de carteiras abertas. Estes casos serão estudados em simultâneo, com o objetivo de perceber qual o melhor caso a modelar. Sublinhe-se que o foco principal desta dissertação continua a consistir no melhor modelo preditivo, ou seja, o poder preditivo terá mais peso do que a qualidade do ajuste na escolha do modelo.

O estudo começa com uma descrição descritiva dos dados mensais com o intuito de compreender o seu comportamento ao longo do tempo. O primeiro passo de uma análise de uma série temporal passa por descrever o seu histórico. Isto inclui a representação gráfica dos dados. Quando

uma série temporal é representada através do seu gráfico, é costume encontrar-se padrões. Estes padrões podem ser explicados por várias relações causa-efeito. Algumas componentes mais conhecidas são a tendência, efeito sazonal, alterações cíclias e aleatoriedade. Uma tarefa mais ambiciosa e interessante é extrapolar futuros valores com base em registos passados e, mais especificamente, calcular intervalos de previsão. Assim, a identificação destas componentes é fundamental para a seleção do modelo.

Com o intuito de perceber o comportamento dos dados mensais, a Tabela 6.1 mostra as suas estatísticas descritivas.[1] Ao contrário dos dados diários, estes apresentam muito poucos outliers, o que torna a leitura dos diagramas em caixas de bigode (Figura 6.1) mais fácil. Como esperado, o desvio padrão é superior e aponta para uma maior variabilidade durante o período observado.

Tabela 6.1: Medidas descritivas das séries mensais de contagens brutas.

	Categoria				
	DNA	**REL**	**TMP**	**Restantes causas**	**Total**
Início	01/2015	01/2015	01/2015	01/2015	01/2015
Fim	06/2021	06/2021	06/2021	06/2021	06/2021
Dimensão	78	78	78	78	78
N.º de zeros	0	0	0	0	0
Amplitude	554 - 1485	154 - 772	62 - 904	143 - 318	1013 - 2707
$Q_{0,25}$	699	210	133	221	1250
Mediana	773	259	193	235	1400
$Q_{0,75}$	909	334	265	260	1623
Média	819,32	277,29	228,81	239,45	1466,13
Desvio padrão	169,97	100,52	150,72	31	340,97
Variância	28891,34	10103,75	22716,59	961,25	116258,6
Coeficiente de variação	0,21	0,36	0,66	0,13	0,23
N.º de outliers	2	2	7	1	3

Os anos 2020 e 2021, em particular 2021, mostram um comportamento diferente dos anos anteriores: 2021 mostra uma enorme variabilidade e valores mais elevados. Isto deve-se ao contexto da pandemia da COVID-19 (Figura 6.1). Existem 4 observações outliers: janeiro de 2015 (1669 sinistros), março de 2018 (2707 sinistros), janeiro de 2020 (2138 sinistros) e dezembro de 2020 (2131 sinistros).

Antes de iniciar os processos de modelaçãodecidiu-se[2] efetuar uma transformação aos valores superiores a 2000 sinistros mensais (um "teto" de 2000 sinistros). Então, se o valor observado x_o for superior a 2000 sinistros, este sofre uma transformação $x_o = \frac{x_{o-12}+x_{o+12}}{2}$ (onde x_o é a observação outlier) que respeita a sazonalidade anual inerente aos dados (subida no final de cada

[1] A Tabela 6.1 considera outliers teóricos e não os escolhidos para serem suavizados.

[2] Escolha efetuada unanimemente entre todos os intervenientes, após serem analisados vários valores.

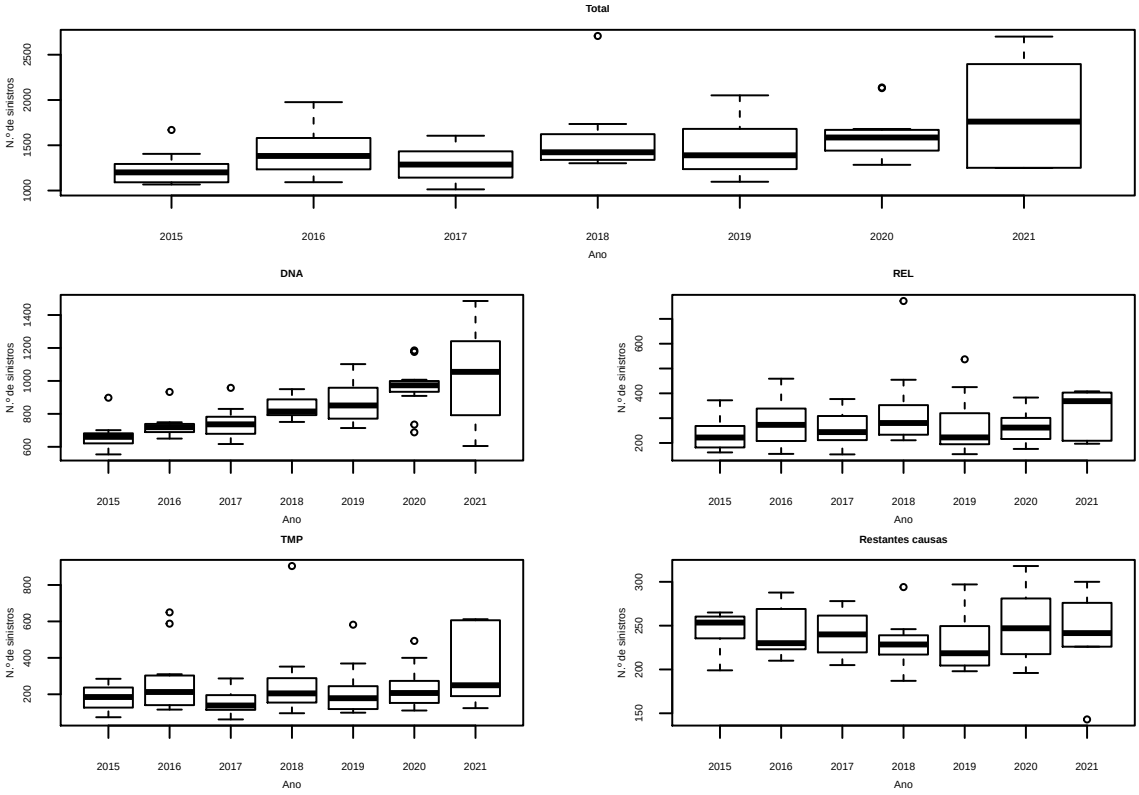


Figura 6.1: Diagramas em caixas de bigodes das séries mensais.

ano, seguida de uma descida no princípio do ano seguinte), que se manifesta nas representações gráficas das FAC e FACP (Figura 6.2). Logo, seis valores foram alterados (Figura 6.4 e Tabela 6.2).

Note-se que esta transformação foi aplicada à série diária transformada, i.e., o estudo mensal baseia-se na série diária já sem outliers. No mesmo contexto, a série correspondente à taxa de frequência, também ilustra uma sazonalidade anual (Figura 6.3), apesar de não ser tão evidente. Por este motivo, os modelos serão ajustados tendo em conta uma sazonalidade anual ($s = 12$) em ambos os casos, visíveis nos gráficos relativos às contagens brutas (Figura 6.4). Sendo a taxa de frequência um quociente, não são identificados quaisquer outliers (Figura 6.5), logo não será efetuada qualquer transformação aos dados.

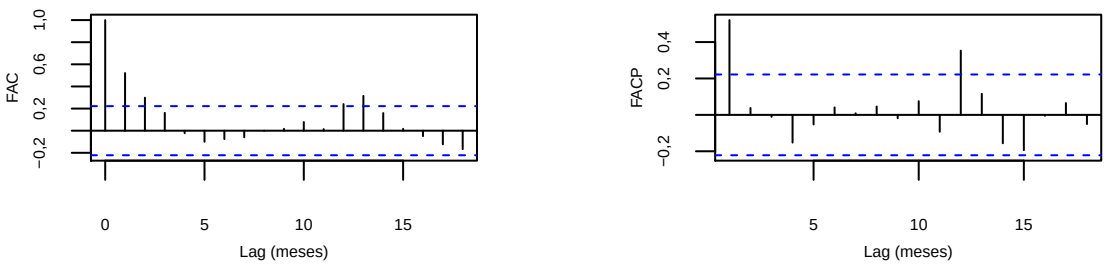


Figura 6.2: FAC e FACP da série mensal do número total de sinistros.

A série mensal correspondente à categoria **TMP** representa uma grande parte do das conta-

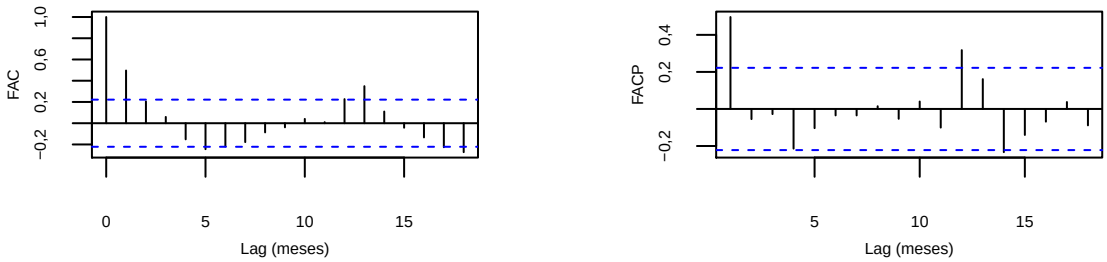


Figura 6.3: FAC e FACP da série mensal de taxa de frequência.

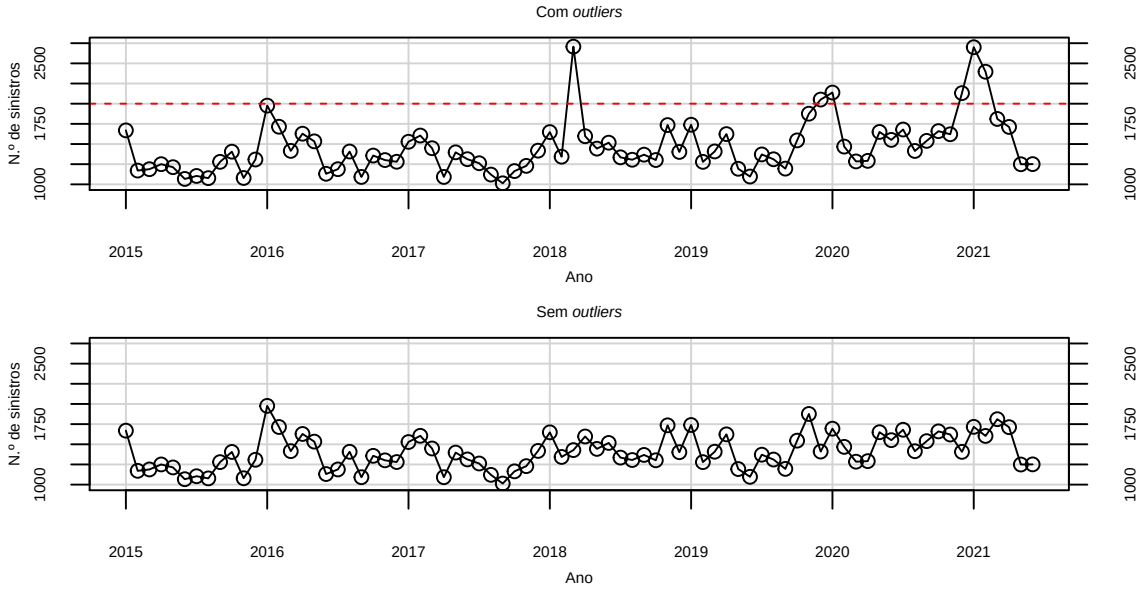


Figura 6.4: Suavização dos outliers na série mensal do número total de sinistros.

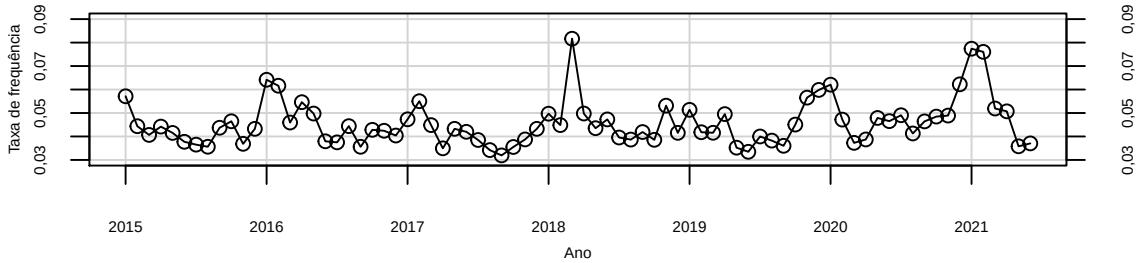


Figura 6.5: Representação gráfica da série mensal da taxa de frequência do número total de sinistros.

gens totais. Portanto seria de esperar que esta também assume valores outliers. Assim, para além da categoria **total**, a categoria **TMP** também sofre alterações nos seus outliers com a mesma transformação.

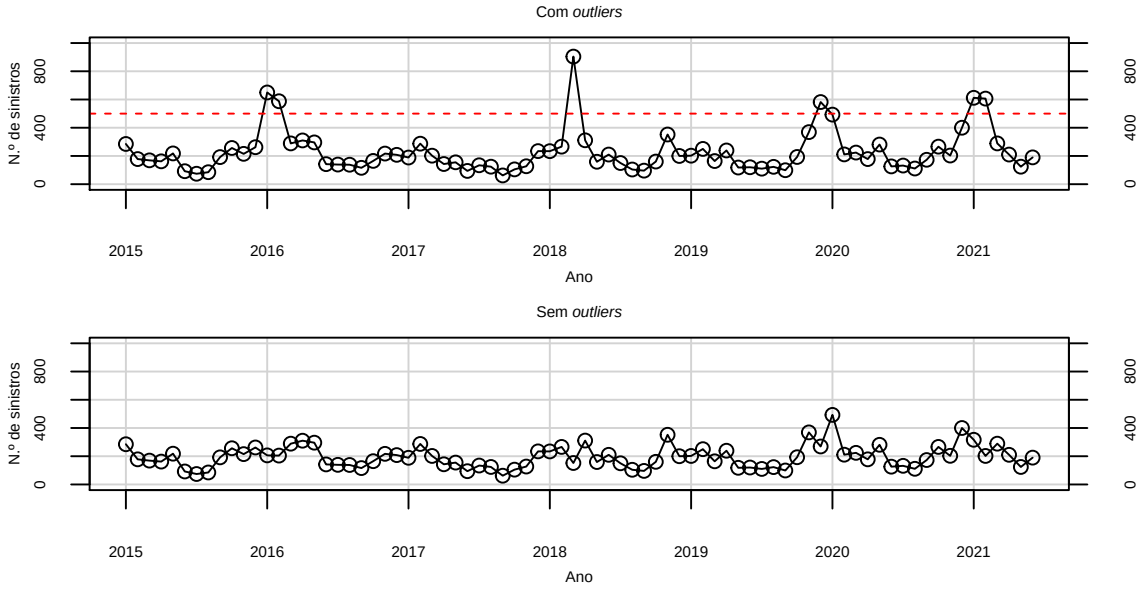


Figura 6.6: Suavização dos outliers na série mensal da categoria **TMP**.

Tabela 6.2: Outliers antes e depois da transformação.

Total		
Data	**Antes**	**Depois**
03/2018	2707	1428
12/2019	2051	1410
01/2020	2138	1694
12/2020	2131	1406
01/2021	2700	1716
02/2021	2396	1604

TMP		
Data	**Antes**	**Depois**
01/2016	650	206
02/2016	588	205
03/2018	904	152
12/2019	582	269
01/2021	612	317
02/2021	606	202

6.2 Aplicação dos métodos de previsão aos dados mensais

Uma vez efetuada uma exposição descritiva dos dados mensais, são então ajustados modelos SARIMA e Holt-Winters. A metodologia TBATS não está incluída neste capítulo, porque o critério que levou à inclusão da mesma no Capítulo 5, a alta variabilidade juntamente com possíveis sazonalidades múltiplas dos dados diários, não se manifesta nos dados mensais. Deste modo, a metodologia a aplicar nestas modelações são as mesmas que no Capítulo 5. No caso dos modelos SARIMA, a

estacionariedade tem de ser verificada (uma vez mais, através da tranformação de Box-Cox), antes de ser efetuado o ajuste do modelo. Além disso, os modelos finais só são válidos se verificarem os pressupostos já enunciados para os resíduos dos modelos.

6.2.1 Caso I: Séries mensais do número total de sinistros

Esta secção trata da modelação da série mensal **total**, enquanto que as séries mensais marginais serão modeladas na Secção 6.2.2. As séries temporais encontram-se divididas num conjunto de treino e noutro de teste, em ambos os casos (série de contagem e série de taxa de frequência). O conjunto de treino consiste nas primeiras 72 observações (desde janeiro de 2015 até dezembro de 2020) e o conjunto de teste nas últimas 6 observações (desde janeiro até junho de 2021), representando cerca de 92% e 8% dos dados, respetivamente.

Modelo SARIMA

Para se analisar a estacionariedade das séries são utilizados os testes de ADF e KPSS (com um nível de significância de 5%). Estes concluem que a série das contagens brutas totais é estacionária logo, não necessita de uma transformação de Box-Cox. Assim sendo, o modelo SARIMA será ajustado à série original.

As Tabelas 6.3 e 6.4 mostram os cinco modelo com menor AIC e BIC quando ajustados às séries de treino de ambos os casos. Foram testadas todas as combinações de parâmetros p, d, q, P, D e Q, fazendo variar p, q, P e Q entre 0 e 2 e D entre 0 e 1 (as séries já são consideradas estacionárias para a variância, logo $d = 0$).

Tabela 6.3: Ajuste dos modelos SARIMA na série mensal do número total de sinistros.

Modelo	AIC	Modelo	BIC
SARIMA$(0,0,1)(0,1,1)_{12}$	816,13	SARIMA$(0,0,1)(0,1,1)_{12}$	822,41
SARIMA$(1,0,0)(0,1,1)_{12}$	816,52	SARIMA$(1,0,0)(0,1,1)_{12}$	822,8
SARIMA$(1,0,2)(0,1,1)_{12}$	816,82	SARIMA$(1,0,0)(2,1,0)_{12}$	825,91
SARIMA$(1,0,0)(2,1,0)_{12}$	817,53	SARIMA$(0,0,1)(2,1,0)_{12}$	826,03
SARIMA$(0,0,1)(2,1,0)_{12}$	817,65	SARIMA$(1,0,1)(0,1,1)_{12}$	826,46

Tabela 6.4: Ajuste dos modelos SARIMA na série da taxa de frequência.

Modelo	AIC	Modelo	BIC
SARIMA$(2,0,2)(1,0,2)_{12}$	-487,91	SARIMA$(1,0,0)(0,0,0)_{12}$	-478,95
SARIMA$(2,0,1)(0,0,0)_{12}$	-487,51	SARIMA$(0,0,1)(0,0,0)_{12}$	-478,45
SARIMA$(2,0,1)(0,0,1)_{12}$	-486,79	SARIMA$(1,0,0)(0,0,1)_{12}$	-477,05
SARIMA$(2,0,2)(0,0,2)_{12}$	-486,27	SARIMA$(0,0,1)(0,0,1)_{12}$	-476,61
SARIMA$(1,0,0)(0,0,1)_{12}$	-486,16	SARIMA$(1,0,0)(1,0,0)_{12}$	-476,54

Tendo em conta ambos os critérios (AIC e BIC) escolheu-se o modelo SARIMA$(0,0,1)(0,1,1)_{12}$ para modelar a série das contagens brutas e o SARIMA$(1,0,0)(0,0,1)_{12}$ para modelar a série da taxa de frequência.

As características destes modelos estão apresentadas na Tabela 6.5.

Tabela 6.5: Características dos modelos SARIMA ajustados às séries mensais do número total de sinistros e taxa de frequência.

Contagens mensais	AIC $\approx 816,13$	BIC $\approx 822,41$	$\widehat{\sigma} \approx 200,16$
	Parâmetro	θ_1	μ_1
SARIMA$(0,0,1)(0,1,1)_{12}$	Estimativa	0,3972	-0,6219
	Desvio padrão	0,1225	0,1687
Taxa de frequência	AIC $\approx -486,16$	BIC $\approx -477,05$	$\widehat{\sigma} \approx 0,008$
	Parâmetro	ϕ_1	μ_1
SARIMA$(1,0,0)(0,0,1)_{12}$	Estimativa	0,3601	0,2101
	Desvio padrão	0,1124	0,1300

As Figuras 6.7 e 6.8 apresentam os conjuntos de treino das séries mensais do número total de sinistros e da taxa de frequência, tal como os valores estimados pelos seus respetivos modelos.

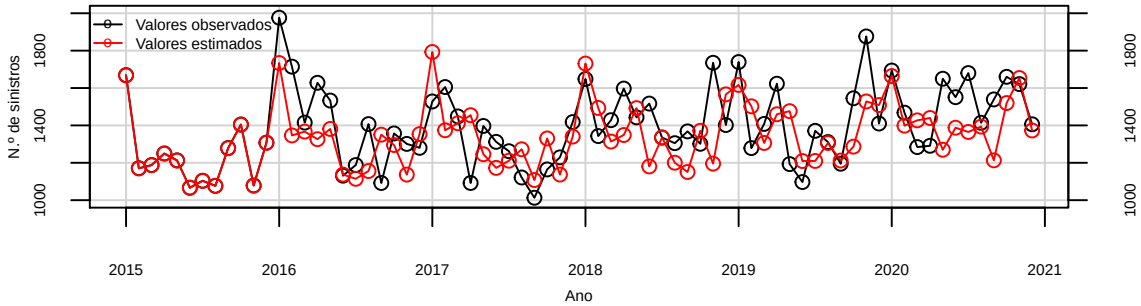


Figura 6.7: Ajuste do modelo SARIMA à série mensal do número total de sinistros.

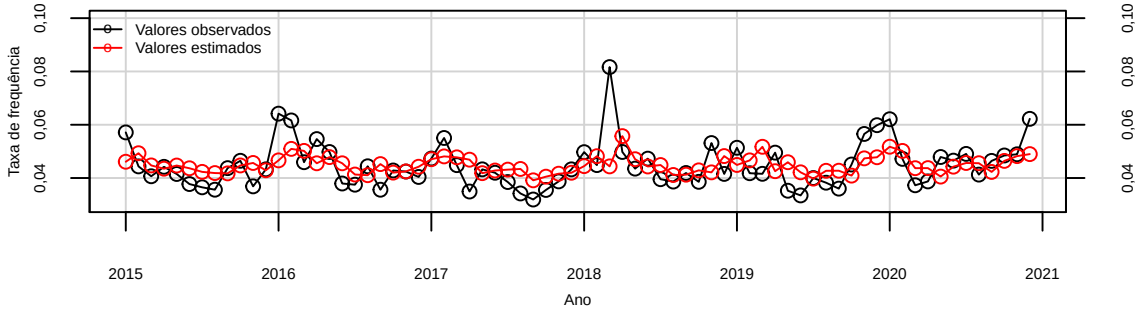


Figura 6.8: Ajuste do modelo SARIMA à série mensal da taxa de frequência do número total de sinistros.

Após a avaliação das previsões pontuais, é essencial entender a eficácia das previsões intervalares.

Teoricamente, as previsões intervalares são calculadas com 95% de confiança, ou seja, 95% dos intervalos devem conter o respetivo valor observado. Note-se que os intervalos de previsão são obtidos baseando-se nas séries de teste, que no caso mensal consistem em séries de 6 observações (6 meses). Além disso, o conjunto de teste vai de janeiro de 2021 até junho de 2021, correspondendo a meio ano da pandemia COVID-19 que por sua vez, teve um tremendo impacto em todos os ramos da sociedade, incluindo o comportamento humano e, consequentemente, no número de sinistros de habitação que claramente tem um comportamento distinto de períodos passados. Nas duas séries em estudo nesta secção, série mensal do número total de sinistros e série mensal da taxa de frequência correspondente ao número total de sinistros, são calculadas taxas de cobertura de cerca de 83% e 100%.

A validação dos modelos é efetuada relaxando os pressupostos de normalidade dos respetivos resíduos. O teste de Ljung-Box aceita a ausência de correlação temporal em ambas as séries (de contagens brutas e de taxa de frequência). Esta conclusão é apoiada pelas Figuras 6.11 e 6.12.

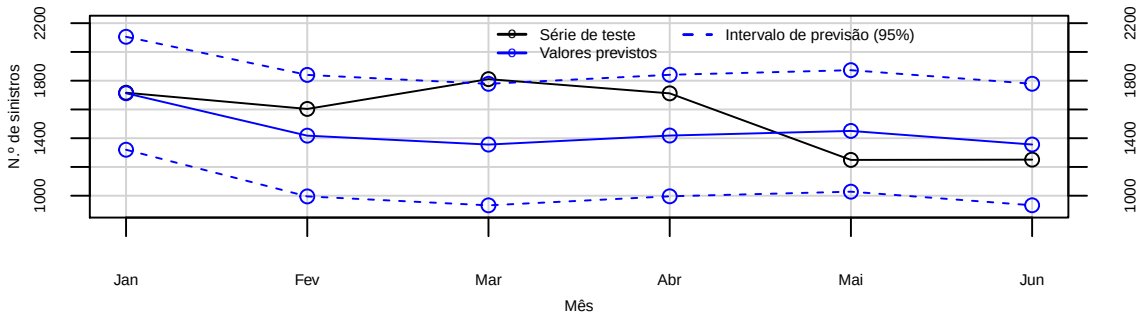


Figura 6.9: Previsões pontuais e intervalares (95% de confiança) do modelo SARIMA ajustado à série mensal do número total de sinistros.

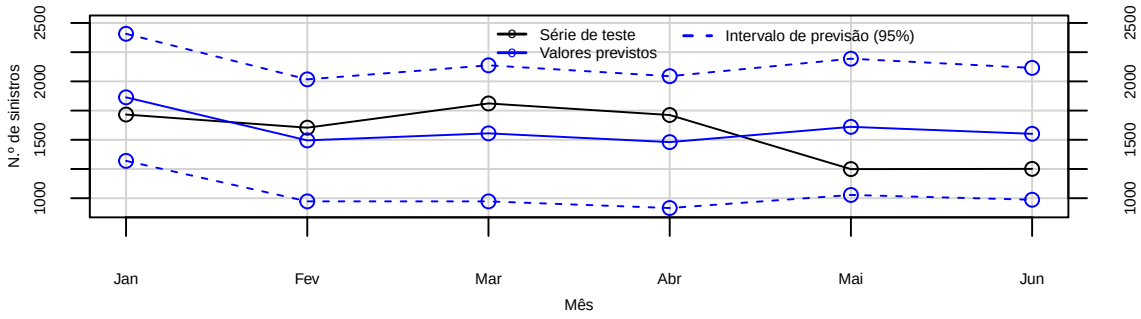


Figura 6.10: Previsões pontuais e intervalares (95% de confiança) do modelo SARIMA ajustado à série mensal da taxa de frequência do número total de sinistros.

Tabela 6.6: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as 6 observações do conjunto de teste, relativos ao modelo SARIMA ajustado à série mensal do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x}_i - x_i$
1863	(1456 ; 1969)	(1320 ; 2105)	1716	0
1495	(1141 ; 1694)	(995 ; 1840)	1604	0
1555	(1079 ; 1632)	(933 ; 1778)	1811	-407
1480	(1142 ; 1695)	(996 ; 1841)	1712	-66
1611	(1174 ; 1727)	(1028 ; 1873)	1249	346
1551	(1080 ; 1632)	(934 ; 1779)	1251	388
Taxa de cobertura	66,7%	83,3%		$\sum_{i=1}^{6} (\widehat{x}_i - x_i) = 261$

Tabela 6.7: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as 6 observações do conjunto de teste, relativos ao modelo SARIMA ajustado à série mensal da taxa de frequência do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x_i} - x_i$
1863	(1508 ; 2219)	(1320 ; 2407)	1716	147
1495	(1154 ; 1836)	(973 ; 2017)	1604	-109
1555	(1174 ; 1936)	(973 ; 2138)	1811	-256
1480	(1111 ; 1849)	(915 ; 2044)	1712	-232
1611	(1230 ; 1992)	(1028 ; 2194)	1249	362
1551	(1182 ; 1920)	(987 ; 2115)	1251	300
Taxa de cobertura	100%	100%		$\sum_{i=1}^{6} (\widehat{x_i} - x_i) = 212$

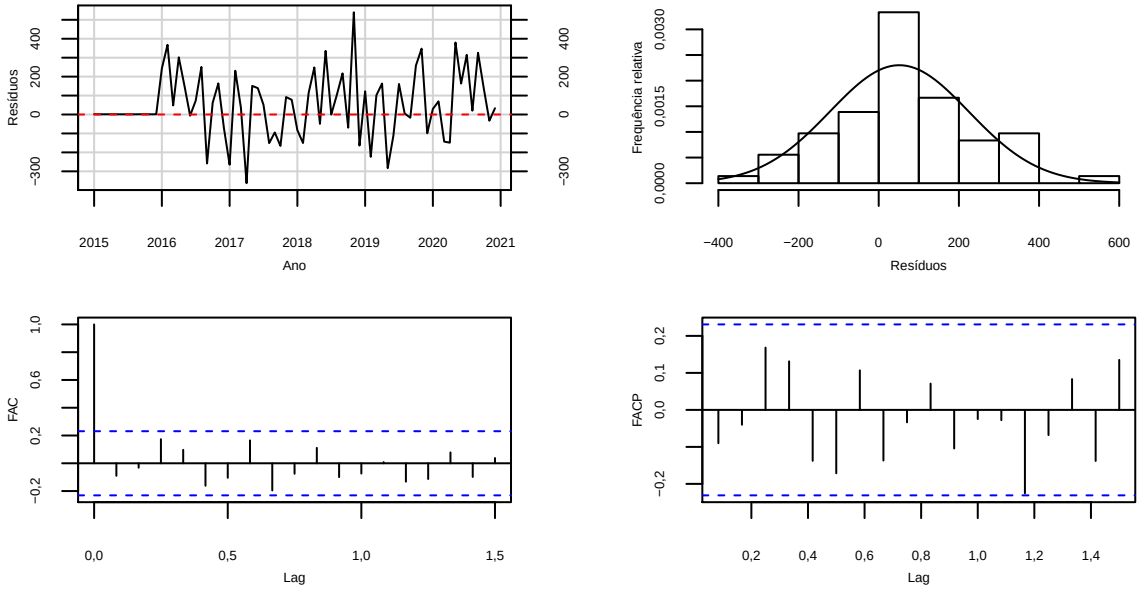


Figura 6.11: Série dos resíduos do modelo SARIMA ajustado à série mensal da categoria **total** e respetivo histograma, FAC e FACP estimadas.

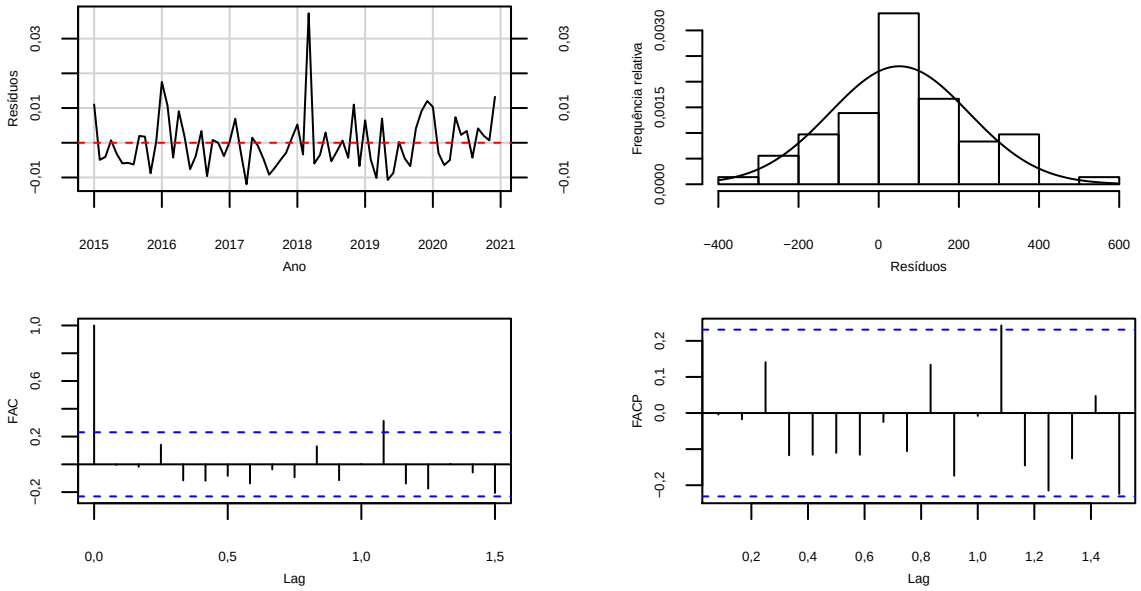


Figura 6.12: Série dos resíduos do modelo SARIMA ajustado à série mensal da taxa de frequência da categoria **total** e respetivo histograma, FAC e FACP estimadas.

Modelo Holt-Winters

O modelo Holt-Winters pode ser ajustado considerando uma decomposição das séries aditiva ou multiplicativa, logo ambos os casos são estudados nesta secção. O ajuste será efetuado no conjunto de treino, tal como na secção anterior, obtendo assim os valores iniciais dos parâmetros de alisamento com base nas primeiras 12 observações (primeiro período sazonal). Utilizam-se as medidas usuais de avaliação para comparar o poder preditivo dos modelos. O modelo Holt-Winters aditivo mostrou melhores resultados preditivos, o que vai de acordo com a oscilação presente nas séries que aparenta ser independente da tendência. No entanto, para ambas as séries, foram considerados tanto o modelo Holt-Winters aditivo como o multiplicativo. As Tabelas 6.8 e 6.9 mostram os valores das estimativas iniciais para o nível, do declive, dos fatores sazonais e das constantes de alisamento correspondente aos modelos de Holt-Winters aditivo e multiplicativo ajustados a cada série.

Tabela 6.8: Estimativas iniciais para o nível, o declive, os fatores sazonais e estimativas das constantes de alisamento, correspondentes aos modelos aditivos de Holt-Winters ajustados aos conjuntos de treino das séries mensais do número total de sinistros e taxa de frequência.

	$REQM \approx 225,98$				
Contagens	$\widehat{\alpha} \approx 0,1603$	$\widehat{\beta} \approx 0,0212$	$\widehat{\gamma} \approx 0,5974$	$\widehat{l_1} \approx 1644,5000$	$\widehat{b_1} \approx 8,4710$
	$\widehat{s_1} \approx 154,2854$	$\widehat{s_2} \approx -112,3845$	$\widehat{s_3} \approx -175,0366$	$\widehat{s_4} \approx -87,0292$	$\widehat{s_5} \approx -1,2906$
	$\widehat{s_6} \approx -110,3108$	$\widehat{s_7} \approx -15,5516$	$\widehat{s_8} \approx -200,5818$	$\widehat{s_9} \approx -177,1819$	$\widehat{s_{10}} \approx -35,7976$
	$\widehat{s_{11}} \approx 54,6681$	$\widehat{s_{12}} \approx -206,0083$			
	$REQM \approx 0,01$				
Taxa	$\widehat{\alpha} \approx 0,0839$	$\widehat{\beta} \approx 0,0128$	$\widehat{\gamma} \approx 0,3795$	$\widehat{l_1} \approx 0,0526$	$\widehat{b_1} \approx 0,0002$
de	$\widehat{s_1} \approx 0,0081$	$\widehat{s_2} \approx 0,0002$	$\widehat{s_3} \approx -0,0012$	$\widehat{s_4} \approx -0,0027$	$\widehat{s_5} \approx -0,0034$
frequência	$\widehat{s_6} \approx -0,0060$	$\widehat{s_7} \approx -0,0052$	$\widehat{s_8} \approx -0,0091$	$\widehat{s_9} \approx -0,0074$	$\widehat{s_{10}} \approx -0,0043$
	$\widehat{s_{11}} \approx -0,0005$	$\widehat{s_{12}} \approx 0,0037$			

Tabela 6.9: Estimativas iniciais para o nível, o declive, os fatores sazonais e estimativas das constantes de alisamento, correspondentes aos modelos multiplicativos de Holt-Winters ajustados aos conjuntos de treino das séries mensais do número total de sinistros e taxa de frequência.

	$REQM \approx 214,75$				
Contagens	$\widehat{\alpha} \approx 0,2195$	$\widehat{\beta} \approx 0,0246$	$\widehat{\gamma} \approx 0,6400$	$\widehat{l_1} \approx 1675,5559$	$\widehat{b_1} \approx 8,4902$
	$\widehat{s_1} \approx 1,0777$	$\widehat{s_2} \approx 0,9131$	$\widehat{s_3} \approx 0,8793$	$\widehat{s_4} \approx 0,9386$	$\widehat{s_5} \approx 0,9973$
	$\widehat{s_6} \approx 0,9177$	$\widehat{s_7} \approx 0,9774$	$\widehat{s_8} \approx 0,8595$	$\widehat{s_9} \approx 0,9684$	$\widehat{s_{10}} \approx 0,9556$
	$\widehat{s_{11}} \approx 1,0076$	$\widehat{s_{12}} \approx 0,8528$			
	$REQM \approx 0,01$				
Taxa	$\widehat{\alpha} \approx 0,1648$	$\widehat{\beta} \approx 0,0087$	$\widehat{\gamma} \approx 0,4047$	$\widehat{l_1} \approx 0,0542$	$\widehat{b_1} \approx 0,0003$
de	$\widehat{s_1} \approx 1,1512$	$\widehat{s_2} \approx 1,0033$	$\widehat{s_3} \approx 0,9740$	$\widehat{s_4} \approx 0,9515$	$\widehat{s_5} \approx 0,9393$
frequência	$\widehat{s_6} \approx 0,8817$	$\widehat{s_7} \approx 0,8948$	$\widehat{s_8} \approx 0,8136$	$\widehat{s_9} \approx 0,8447$	$\widehat{s_{10}} \approx 0,9126$
	$\widehat{s_{11}} \approx 0,9858$	$\widehat{s_{12}} \approx 1,0534$			

Tabela 6.10: Medidas de avaliação dos modelos de Holt-Winters para a série mensal do número total de sinistros.

	Modelo	EQM	REQM	EPAM	EEAM	U-Theil
Conjunto de treino	**Aditivo**	**46119,5598**	**214,7547**	**12,3648**	**0,9148**	**0,9691**
	Multiplicativo	51064,8079	225,9752	12,9967	0,9667	1,0167
Conjunto de teste	**Aditivo**	**71315,0042**	**267,0487**	**15,8089**		**1,3641**
	Multiplicativo	75164,1413	274,1608	16,1487		1,3916

Tabela 6.11: Medidas de avaliação dos modelos de Holt-Winters para a série mensal da taxa de frequência.

	Modelo	EQM	REQM	EPAM	EEAM	U-Theil
Conjunto de treino	**Aditivo**	0,0001	**0,0096**	**15,3188**	**0,8661**	**1,1034**
	Multiplicativo	0,0001	0,0097	15,6062	0,8846	1,1140
Conjunto de teste	**Aditivo**	0,0002	0,0137	**20,3246**		**1,1748**
	Multiplicativo	0,0002	**0,0134**	22,0032		1,2488

As Tabelas 6.10 e 6.11 mostram que o modelo com melhores avaliações, tanto no conjunto de treino como no conjunto de teste, é o aditivo em ambas as séries. Estes resultados vão ao encontro do referido anteriormente em relação à oscilação presente nas séries. Partindo destes resultados, as Figuras 6.13, 6.14, 6.15 e 6.16 ilustram os ajustes e as previsões, pontuais e intervalares[3] (com 95% de confiança), dos modelos de Holt-Winters aditivos ajustados a ambas as séries.

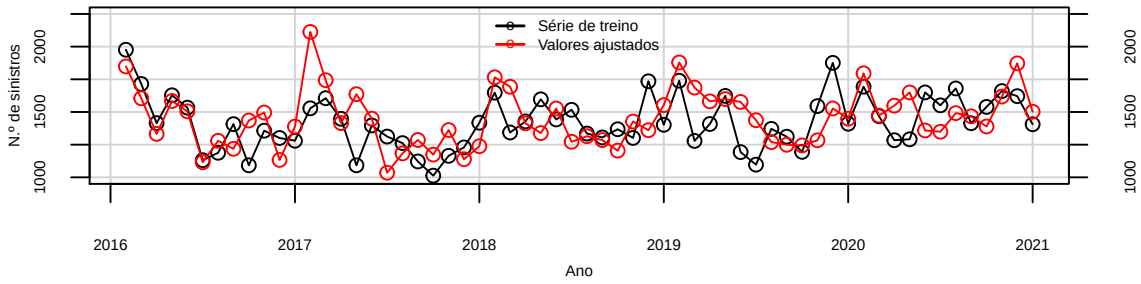


Figura 6.13: Ajuste do modelo aditivo de Holt-Winters (no período de treino) sobreposto à série mensal do número total de sinistros.

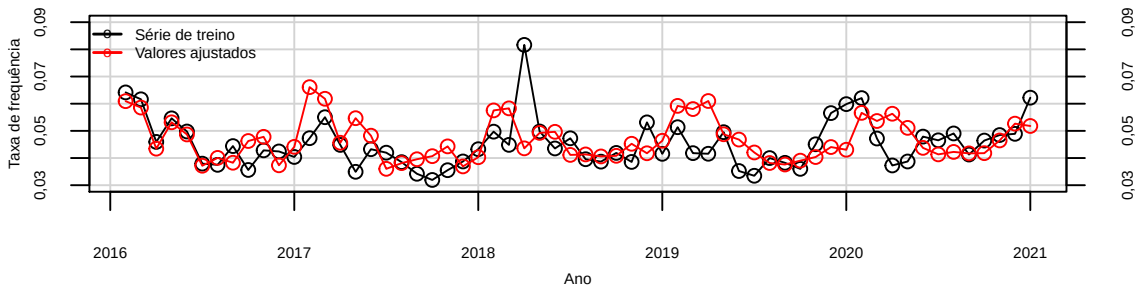


Figura 6.14: Ajuste do modelo aditivo de Holt-Winters (no período de treino) sobreposto à série mensal da taxa de frequência.

[3]Os intervalos de previsão são, uma vez mais, construídos pelo método de Bootstrap exposto na Secção 4.3.4.

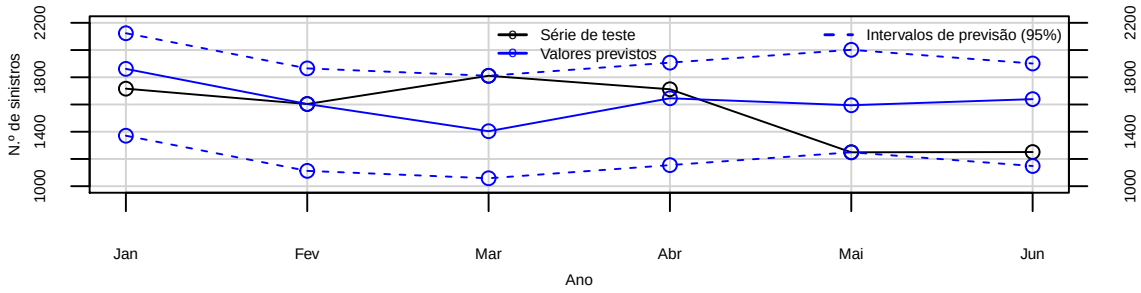


Figura 6.15: Previsões pontuais e intervalares (95% de confiança) do modelo aditivo de Holt-Winters ajustado à série mensal do número total de sinistros.

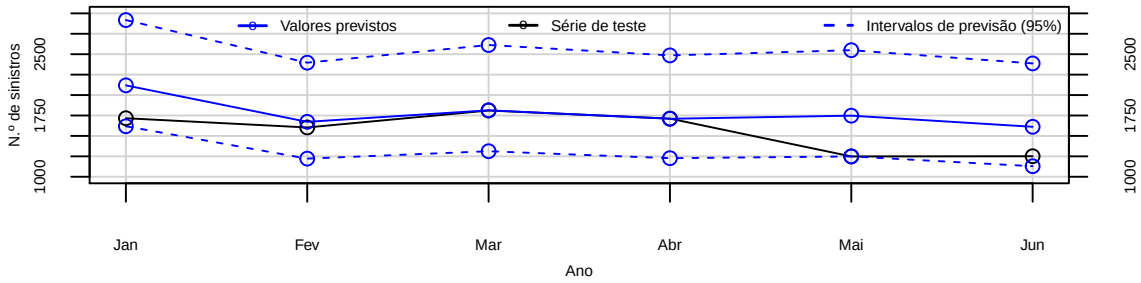


Figura 6.16: Previsões pontuais e intervalares (95% de confiança) do modelo aditivo de Holt-Winters ajustado à série mensal da taxa de frequência.

Nas Tabelas 6.12 e 6.13 estão apresentadas as previsões pontuais e intervalares (a 80% e 95% de confiança). O facto de a série de teste ser constituída apenas por 6 observações faz com que os intervalos de previsão com 80% e 95% de confiança sejam, neste caso, exatamente os mesmos, pois os quantis utilizados no cálculo dos limites destes intervalos são os mesmos em ambos os casos. Ao calcular os quantis de 95% e 80% das amostras de Bootstrap, tal como são construídos os intervalos de previsão deste método, os valores são os mesmos. Este facto não implica a baixa eficácia do método, mas sim a falta de observações na série de teste. Em contrapartida, as previsões pontuais não tendem a afastar-se muito dos valores observados, principalmente as duas primeiras (janeiro e fevereiro de 2021), onde os valores observados e as previsões pontuais coincidem.

As previsões intervalares de ambos os modelos contêm todas os valores observados, no entanto, as previsões pontuais do modelo ajustado à série de contagens brutas são melhores do que as do modelo ajustado à série da taxa de frequência.

Tabela 6.12: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as 6 observações do conjunto de teste, relativos ao modelo Holt-Winters ajustado à série mensal do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x}_i - x_i$
1716	(1371 ; 2123)	(1371 ; 2123)	1716	0
1604	(1112 ; 1865)	(1112 ; 1865)	1604	0
1404	(1058 ; 1812)	(1058 ; 1812)	1811	-407
1646	(1155 ; 1907)	(1155 ; 1907)	1712	-66
1595	(1248 ; 2002)	(1248 ; 2002)	1249	346
1639	(1148 ; 1901)	(1148 ; 1901)	1251	388
Taxa de cobertura	100%	100%		$\sum_{i=1}^{6} (\widehat{x}_i - x_i) = 261$

Tabela 6.13: Valores previstos e respetivos intervalos de previsão a 80% e 95% e valores observados, para as 6 observações do conjunto de teste, relativos ao modelo Holt-Winters ajustado à série mensal da taxa de frequência do número total de sinistros.

Valor previsto	**Intervalo de previsão a 80%**	**Intervalo de previsão a 95%**	**Valor observado**	$\widehat{x_i} - x_i$
2117	(1618 ; 2918)	(1618 ; 2918)	1716	401
1675	(1222 ; 2396)	(1222 ; 2396)	1604	71
1813	(1312 ; 2613)	(1312 ; 2613)	1811	2
1712	(1227 ; 2485)	(1227 ; 2485)	1712	0
1750	(1248 ; 2549)	(1248 ; 2549)	1249	501
1612	(1129 ; 2387)	(1129 ; 2387)	1251	361
Taxa de cobertura	100%	100%		$\sum_{i=1}^{6} (\widehat{x_i} - x_i) = 1336$

6.2.2 Caso II: Séries mensais marginais dos números de sinistros

Esta secção trata das séries mensais das categorias **DNA**, **REL**, **TMP** e **restantes causas**. Em todas estas, está definida a série que consiste nos números de sinistros observados (contagens brutas) e a série da taxa de frequência que é calculada segundo a equação 6.1, alterando apenas o numerador conforme a série.

Nota 9. A escolha da inalteração do denominador da equação 6.1 quando aplicada às séries marginais foi baseada no facto de, apesar de ser diferente, este valor não ter muita variabilidade entre estas séries. Isto é, geralmente, uma habitação que esteja exposta a um risco da categoria, e.g., **DNA**, também está exposta a um risco das restantes categorias.

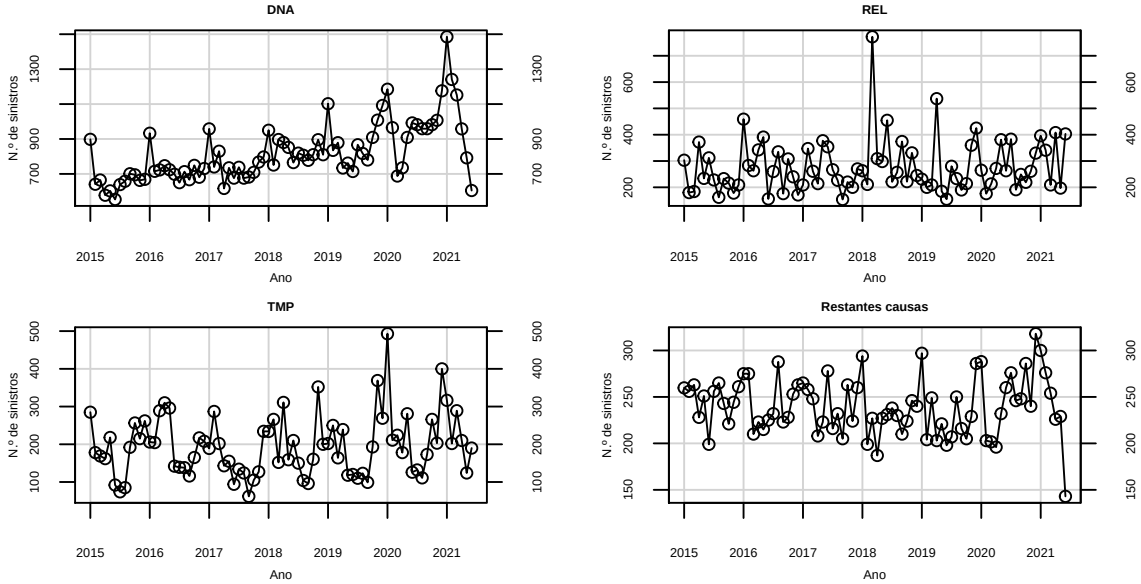


Figura 6.17: Representação gráfica das séries mensais de contagens marginais.

Em todas as categorias, o conjunto de treino e de teste são constituídos pelas primeiras 72 e as últimas 6 observações, respetivamente, tal como na Secção 6.2.1. Todos os modelos serão ajustados ao conjunto de treino e avaliados pela sua capacidade preditiva no conjunto de teste, tendo sempre em conta a Nota 8.

Modelo SARIMA

Analisando as Figuras 6.17 e 6.18, verifica-se que nenhuma série é estacionária para a variância. Assim, foram testados vários valores para λ da transformação de Box-Cox, obtendo-se $\lambda = -0,958$, $\lambda = -0,815$, $\lambda = 0,089$ e $\lambda = -0,628$ para as categorias **DNA**, **REL**, **TMP** e **restantes causas**, respetivamente. Os testes de estacionariedade ADF e KPSS aplicados, segundo a regra de Ng & Perron [Ng and Perron, 1995], às séries transformadas estão resumidos

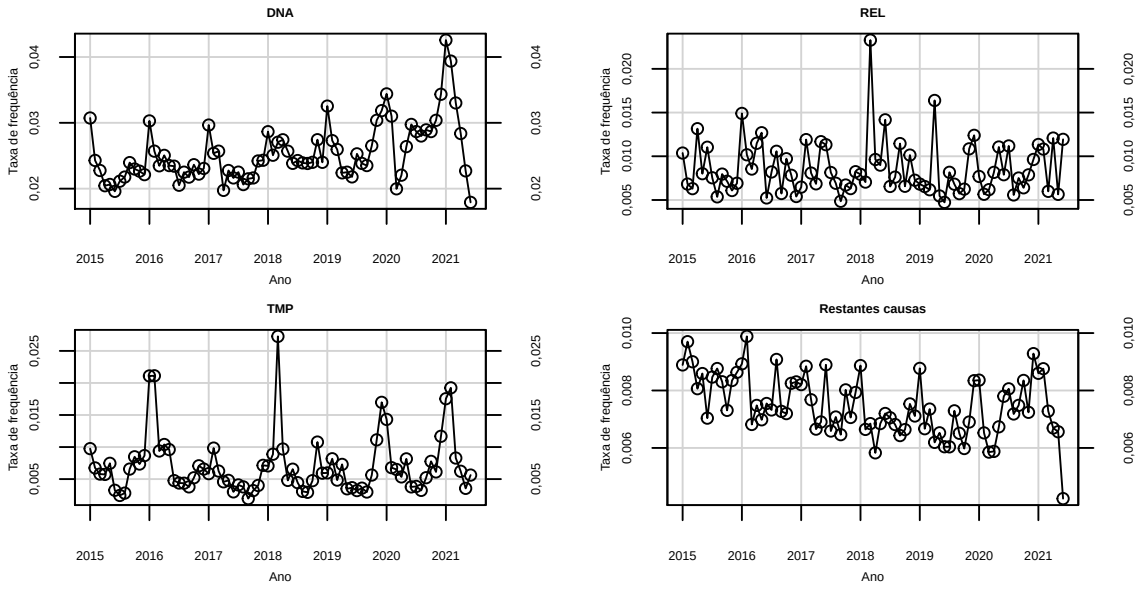


Figura 6.18: Representação gráfica das séries das taxas de frequência marginais.

nas Tabelas 6.14 e 6.15. Com nível de significância de 5%, apenas as séries correspondentes às taxas de frequência das categorias **DNA** e **restantes causas** não são consideradas estacionárias pelo teste KPSS. Contudo, uma vez que o teste ADF assume estacionariedade e para manter consistência na estrutura nesta dissertação, estes casos serão estudados como sendo estacionários.

Tabela 6.14: Resumo dos testes ADF e KPSS para as séries mensais marginais de contagens.

Teste de hipóteses	Categoria	Estatística de teste	Número de lags usado
ADF	**DNA**	5,2131	11
	REL	4,4038	10
	TMP	1,1134	10
	Restantes causas	0,8795	11
KPSS	**DNA**	0,0883	11
	REL	0,1442	10
	TMP	0,1846	10
	Restantes causas	0,1877	11

Tabela 6.15: Resumo dos testes ADF e KPSS para as séries mensais marginais de taxas de frequência.

Teste de hipóteses	Categoria	Estatística de teste	Número de lags usado
ADF	**DNA**	2,5811	11
	REL	4,1331	10
	TMP	7,8900	5
	Restantes causas	3,2653	11
KPSS	**DNA**	0,6172	11
	REL	0,1335	10
	TMP	0,0526	5
	Restantes causas	0,1794	11

Uma vez consideradas estacionárias as séries mensais marginais, o passo seguinte consiste em ajustar modelos SARIMA. Este processo é automatizado considerando todas as combinações possíveis dos parâmetros, tal como na Secção 6.2.1. As características dos modelos escolhidos, tendo em conta os menores valores de AIC e BIC, estão descritas nas Tabelas 6.16 e 6.17.

Tabela 6.16: Características dos modelos SARIMA escolhidos para modelar as séries mensais marginais de contagens, após efetuadas transformações de Box-Cox.

DNA	SARIMA$(2,0,0)(2,0,0)_{12}$	AIC $\approx -1027,8$		BIC $\approx -1014,14$	$\widehat{\sigma} \approx 0$
	Parâmetro	ϕ_1	ϕ_2	ν_1	ν_2
	Estimativa	0,5007	0,2305	0,3553	0,3306
	Erro padrão	0,1211	0,1187	0,1258	0,1427
REL	SARIMA$(1,0,1)(0,0,0)_{12}$	AIC $\approx -618,53$		BIC $\approx -609,42$	$\widehat{\sigma} \approx 0$
	Parâmetro	ϕ_1		θ_1	
	Estimativa	-0,7011		0,8908	
	Erro padrão	0,1510		0,1004	
TMP	SARIMA$(0,0,0)(0,1,1)_{12}$	AIC $\approx 111,52$		BIC $\approx 115,71$	$\widehat{\sigma} \approx 0,50$
	Parâmetro				η_1
	Estimativa				-1,0000
	Erro padrão				0,2823
Restantes causas	SARIMA$(0,0,0)(1,0,0)_{12}$	AIC $\approx -603,77$		BIC $\approx -596,94$	$\widehat{\sigma} \approx 0$
	Parâmetro			ν_1	
	Estimativa			0,4297	
	Erro padrão			0,1146	

Tabela 6.17: Características dos modelos SARIMA escolhidos para modelar as séries mensais marginais de taxas de frequência.

DNA	SARIMA$(0,0,2)(1,0,0)_{12}$	AIC $\approx -1024,2$		BIC $\approx -1012,82$	$\widehat{\sigma} \approx 0$
	Parâmetro		θ_1	θ_2	ν_1
	Estimativa		0,5577	0,4491	0,6005
	Erro padrão		0,1246	0,0884	0,1069
REL	SARIMA$(0,0,0)(0,0,1)_{12}$	AIC $\approx -616,04$		BIC $\approx -609,21$	$\widehat{\sigma} \approx 0$
	Parâmetro				η_1
	Estimativa				-0,0659
	Erro padrão				0,1399
TMP	SARIMA$(0,0,1)(0,0,0)_{12}$	AIC $\approx 138,13$		BIC $\approx 144,96$	$\widehat{\sigma} \approx 0,61$
	Parâmetro		θ_1		
	Estimativa		0,3489		
	Erro padrão		0,0900		
Restantes causas	SARIMA$(1,0,1)(1,0,0)_{12}$	AIC $\approx -600,8$		BIC $\approx -589,42$	$\widehat{\sigma} \approx 0,61$
	Parâmetro	ϕ_1	θ_1		ν_1
	Estimativa	0,2683	-0,1504		0,4204
	Erro padrão	0,6087	0,6176		0,1148

Todas as séries de contagens provêm de distribuições Gaussianas segundo os testes de Kolmogorov-Smirnov (com correção de Lilliefors) e Shapiro-Wilk (normalidade visível na Figura 6.19) e assumem que os resíduos dos seus modelos são independentes (com 95% de confiança) com base no teste de Ljung-Box. O mesmo não acontece nas séries de taxas de frequência, onde, apesar de se aceitar que os resíduos dos modelos são independentes, não são Gaussianos. A Figura 6.20 mostra que a normalidade não é evidente nestas séries.

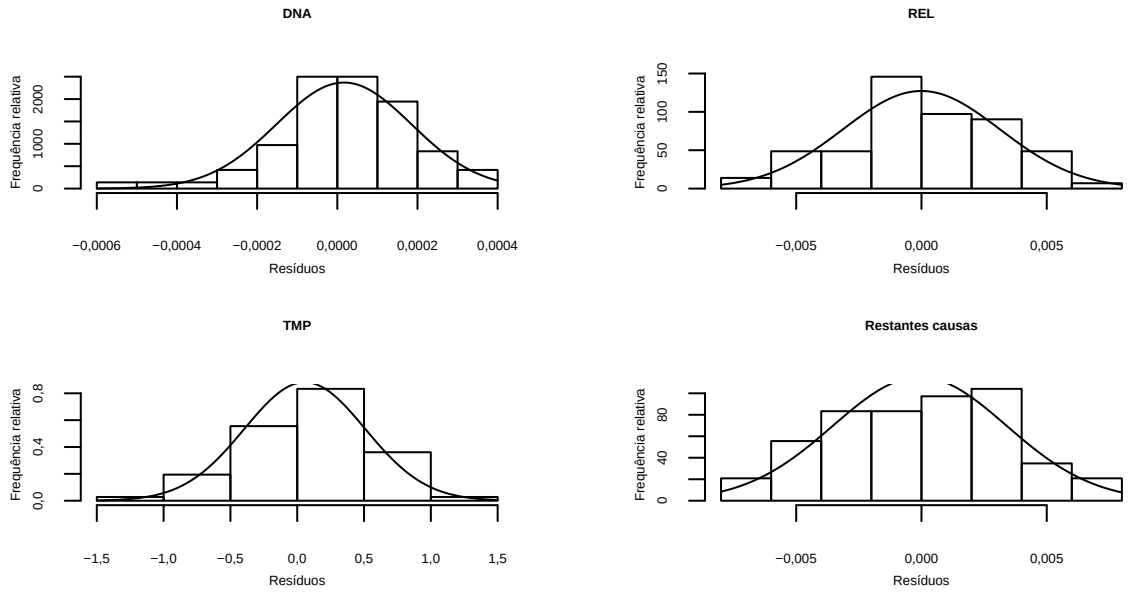


Figura 6.19: Histogramas dos resíduos dos modelos SARIMA ajustados às séries mensais de contagens brutas das categorias marginais.

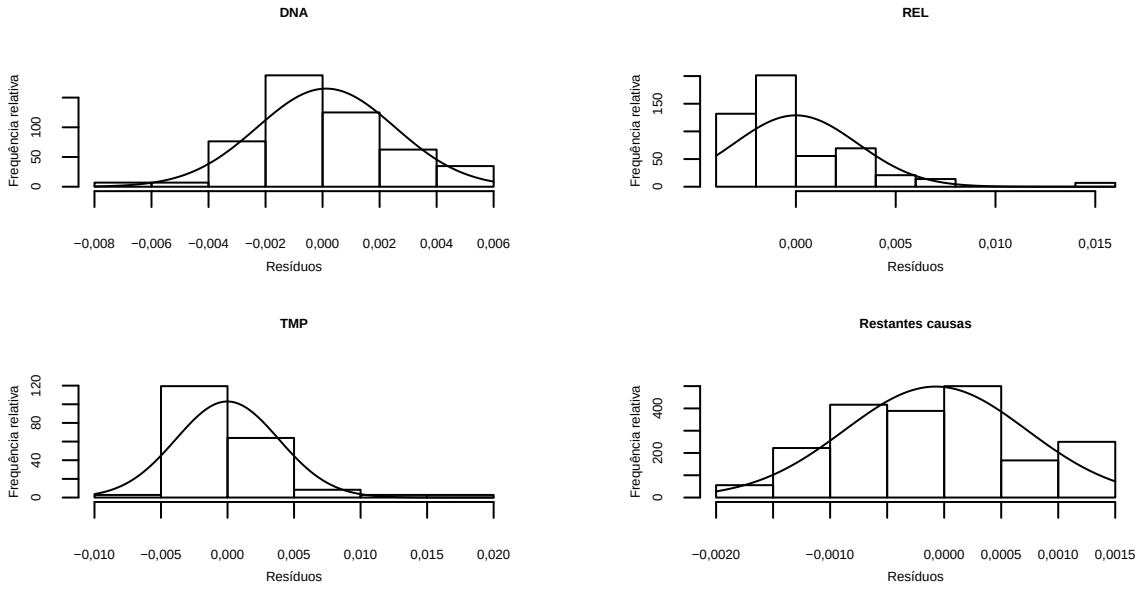


Figura 6.20: Histogramas dos resíduos dos modelos SARIMA ajustados às séries mensais de taxas de frequência das categorias marginais.

No entanto, os histogramas apresentados nas Figuras 6.19 e 6.20 mostram uma forte simetria em torno do valor nulo e as respetivas curvas das funções de densidade empíricas de cada categoria estão ao nível das suas barras, o que permite assumir normalidade dos resíduos. As representações gráficas das FAC e FACP das categorias marginais estão ilustradas nas Figuras D.1 e D.2 do Apêndice D, onde se vê que as bandas de Bartlett cobrem grande parte das autocorrelações em todos casos.

Os ajustes e previsões (pontuais e intervalares a 95% de confiança) destes modelos estão representados(as) nas Figuras D.3, D.4, D.5 e D.6 do Apêndice D.

Modelo Holt-Winters

Os gráficos das séries mensais, tanto de contagens como de taxas de frequência, não fornecem uma resposta direta quanto à versão da modelação Holt-Winters a escolher (aditiva ou multiplicativa). Por este motivo, mais uma vez, ambos os casos são ajustados e comparados para cada série. Após estes ajustes considerando sazonalidade anual ($s = 12$), os modelos aditivos são os que apresentam melhores valores das medidas de ajustamento, ou seja, não há motivo para acreditar que a variabilidade das séries aumenta em conformidade com a tendência. As Tabelas 6.18 e 6.19 mostram as estimativas iniciais para os níveis, os declives e os fatores sazonais e as estimativas das constantes de alisamento dos modelos aditivos (de contagens e de taxas de frequência).[4]

Tabela 6.18: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, correspondentes aos modelos de Holt-Winters aditivos ajustados aos conjuntos de treino da séries mensais de contagens marginais.

DNA	$REQM \approx$ 81,7473				
	$\widehat{\alpha} \approx$ 0,0266	$\widehat{\beta} \approx 0$	$\hat{\gamma} \approx$ 0,4393	$\hat{l}_1 \approx$ 1004,8879	$\hat{b}_1 \approx$ 5,8195
	$\hat{s}_1 \approx$ 223,1008	$\hat{s}_2 \approx$ -11,1616	$\hat{s}_3 \approx$ -96,0924	$\hat{s}_4 \approx$ -130,4548	$\hat{s}_5 \approx$ -49,1037
	$\hat{s}_6 \approx$ -50,0605	$\hat{s}_7 \approx$ -12,6768	$\hat{s}_8 \approx$ -43,7064	$\hat{s}_9 \approx$ -61,7872	$\hat{s}_{10} \approx$ -16,3588
	$\hat{s}_{11} \approx$ 23,8943	$\hat{s}_{12} \approx$ 102,8411			
REL	$REQM \approx$ 123,6648				
	$\widehat{\alpha} \approx$ 0,0064	$\widehat{\beta} \approx 1$	$\hat{\gamma} \approx$ 0,2553	$\hat{l}_1 \approx$ 274,4719	$\hat{b}_1 \approx$ -1,8827
	$\hat{s}_1 \approx$ 19,0422	$\hat{s}_2 \approx$ -58,2024	$\hat{s}_3 \approx$ 16,3473	$\hat{s}_4 \approx$ 47,7121	$\hat{s}_5 \approx$ 38,4100
	$\hat{s}_6 \approx$ -49,6760	$\hat{s}_7 \approx$ -2,6974	$\hat{s}_8 \approx$ -71,0474	$\hat{s}_9 \approx$ -53,0724	$\hat{s}_{10} \approx$ -64,2057
	$\hat{s}_{11} \approx$ -27,2695	$\hat{s}_{12} \approx$ -0,3987			
TMP	$REQM \approx$ 78,84290				
	$\widehat{\alpha} \approx$ 0,1244	$\widehat{\beta} \approx$ 0,0119	$\hat{\gamma} \approx$ 0,4589	$\hat{l}_1 \approx$ 261,1766	$\hat{b}_1 \approx$ 2,2901
	$\hat{s}_1 \approx$ 101,8238	$\hat{s}_2 \approx$ -2,7481	$\hat{s}_3 \approx$ -29,9951	$\hat{s}_4 \approx$ -6,9116	$\hat{s}_5 \approx$ -11,2258
	$\hat{s}_6 \approx$ -90,6642	$\hat{s}_7 \approx$ -97,8927	$\hat{s}_8 \approx$ -110,3721	$\hat{s}_9 \approx$ -91,1425	$\hat{s}_{10} \approx$ -12,5018
	$\hat{s}_{11} \approx$ 29,8829	$\hat{s}_{12} \approx$ 74,7224			
Restantes causas	$REQM \approx$ 25,3618				
	$\widehat{\alpha} \approx$ 0,0256	$\widehat{\beta} \approx 1$	$\hat{\gamma} \approx$ 0,3930	$\hat{l}_1 \approx$ 252,5801	$\hat{b}_1 \approx$ 5,1081
	$\hat{s}_1 \approx$ 55,8799	$\hat{s}_2 \approx$ -13,8828	$\hat{s}_3 \approx$ -9,2147	$\hat{s}_4 \approx$ -29,7135	$\hat{s}_5 \approx$ -5,4251
	$\hat{s}_6 \approx$ 6,3713	$\hat{s}_7 \approx$ 12,1200	$\hat{s}_8 \approx$ 13,6134	$\hat{s}_9 \approx$ -5,1876	$\hat{s}_{10} \approx$ 11,3080
	$\hat{s}_{11} \approx$ 0,6177	$\hat{s}_{12} \approx$ 47,0671			

[4] Os valores da Tabela 6.19 estão na ordem de grandeza da taxa de frequência (valores entre 0 e 1), logo o REQM não corresponde ao número de sinistros.

Tabela 6.19: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, correspondentes aos modelos de Holt-Winters aditivos ajustados aos conjuntos de treino da séries mensais de taxas de frequência marginais.

DNA	$REQM \approx$ 0,0025				
	$\widehat{\alpha} \approx$ 0,1605	$\widehat{\beta} \approx 0$	$\hat{\gamma} \approx$ 0,5255	$\hat{l}_1 \approx$ 0,0298	$\hat{b}_1 \approx$ 0,0001
	$\hat{s}_1 \approx$ 0,0061	$\hat{s}_2 \approx$ 0,0019	$\hat{s}_3 \approx$ -0,0034	$\hat{s}_4 \approx$ -0,0031	$\hat{s}_5 \approx$ -0,0010
	$\hat{s}_6 \approx$ -0,0001	$\hat{s}_7 \approx$ -0,0002	$\hat{s}_8 \approx$ -0,0012	$\hat{s}_9 \approx$ -0,0010	$\hat{s}_{10} \approx$ -0,0005
	$\hat{s}_{11} \approx$ 0,0017	$\hat{s}_{12} \approx$ 0,0031			
REL	$REQM \approx$ 0,0038				
	$\widehat{\alpha} \approx$ 0,0052	$\widehat{\beta} \approx 1$	$\hat{\gamma} \approx$ 0,2843	$\hat{l}_1 \approx$ 0,0082	$\hat{b}_1 \approx$ -0,0001
	$\hat{s}_1 \approx$ 0,0001	$\hat{s}_2 \approx$ -0,0013	$\hat{s}_3 \approx$ 0,0003	$\hat{s}_4 \approx$ 0,0016	$\hat{s}_5 \approx$ 0,0009
	$\hat{s}_6 \approx$ -0,0013	$\hat{s}_7 \approx$ -0,0003	$\hat{s}_8 \approx$ -0,0024	$\hat{s}_9 \approx$ -0,0016	$\hat{s}_{10} \approx$ -0,0022
	$\hat{s}_{11} \approx$ -0,0007	$\hat{s}_{12} \approx$ -0,0001			
TMP	$REQM \approx$ 0,0050				
	$\widehat{\alpha} \approx$ 0,1072	$\widehat{\beta} \approx$ 0,0215	$\hat{\gamma} \approx$ 0,4164	$\hat{l}_1 \approx$ 0,0089	$\hat{b}_1 \approx$ 0,0001
	$\hat{s}_1 \approx$ 0,0024	$\hat{s}_2 \approx$ 0,0009	$\hat{s}_3 \approx$ 0,0011	$\hat{s}_4 \approx$ -0,0012	$\hat{s}_5 \approx$ -0,0018
	$\hat{s}_6 \approx$ -0,0041	$\hat{s}_7 \approx$ -0,0044	$\hat{s}_8 \approx$ -0,0049	$\hat{s}_9 \approx$ -0,0044	$\hat{s}_{10} \approx$ -0,0022
	$\hat{s}_{11} \approx$ -0,0008	$\hat{s}_{12} \approx$ 0,0022			
Restantes causas	$REQM \approx$ 0,0008				
	$\widehat{\alpha} \approx$ 0,0770	$\widehat{\beta} \approx$ 0,2684	$\hat{\gamma} \approx$ 0,4692	$\hat{l}_1 \approx$ 0,0075	$\hat{b}_1 \approx$ 0,0001
	$\hat{s}_1 \approx$ 0,0016	$\hat{s}_2 \approx$ 0,0001	$\hat{s}_3 \approx$ -0,0003	$\hat{s}_4 \approx$ -0,0007	$\hat{s}_5 \approx$ -0,0002
	$\hat{s}_6 \approx$ 0,0004	$\hat{s}_7 \approx$ 0,0003	$\hat{s}_8 \approx$ 0,0003	$\hat{s}_9 \approx 0$	$\hat{s}_{10} \approx$ 0,0003
	$\hat{s}_{11} \approx$ 0,0002	$\hat{s}_{12} \approx$ 0,0014			

Os parâmetros dos modelos relativos a contagens apresentam valores distintos de 0, ao contrário dos modelos das taxas de frequência. Isto indica que estes segundos modelos não apresentam tanta variabilidade quanto os restantes. As equações dos modelos são então:

Contagens da categoria DNA:

$$
\begin{aligned}
l_t &\approx 0,0266(X_t - s_{t-12}) + 0,9734(l_{t-1} + b_{t-1}) \\
b_t &\approx b_{t-1} \\
s_t &\approx 0,4393(X_t - l_t) + 0,5607 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Taxa de frequência da categoria DNA:

$$
\begin{aligned}
l_t &\approx 0,1605(X_t - s_{t-12}) + 0,8395(l_{t-1} + b_{t-1}) \\
b_t &\approx b_{t-1} \\
s_t &\approx 0,5255(X_t - l_t) + 0,4745 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Contagens da categoria REL:

$$
\begin{aligned}
l_t &\approx 0,0064(X_t - s_{t-12}) + 0,9936(l_{t-1} + b_{t-1}) \\
b_t &\approx l_t - l_{t-1} \\
s_t &\approx 0,2553(X_t - l_t) + 0,7447 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Taxa de frequência da categoria REL:

$$
\begin{aligned}
l_t &\approx 0,0052(X_t - s_{t-12}) + 0,9948(l_{t-1} + b_{t-1}) \\
b_t &\approx l_t - l_{t-1} \\
s_t &\approx 0,2843(X_t - l_t) + 0,7157 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Contagens da categoria TMP:

$$
\begin{aligned}
l_t &\approx 0,1244(X_t - s_{t-12}) + 0,8756(l_{t-1} + b_{t-1}) \\
b_t &\approx 0,0119(l_t - l_{t-1}) + 0,9881 b_{t-1} \\
s_t &\approx 0,4589(X_t - l_t) + 0,5411 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Taxa de frequência da categoria TMP:

$$
\begin{aligned}
l_t &\approx 0,1072(X_t - s_{t-12}) + 0,8928(l_{t-1} + b_{t-1}) \\
b_t &\approx 0,0215(l_t - l_{t-1}) + 0,9785 b_{t-1} \\
s_t &\approx 0,4164(X_t - l_t) + 0,5836 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Contagens da categoria restantes causas:

$$
\begin{aligned}
l_t &\approx 0,0256(X_t - s_{t-12}) + 0,9744(l_{t-1} + b_{t-1}) \\
b_t &\approx l_t - l_{t-1} \\
s_t &\approx 0,393(X_t - l_t) + 0,607 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Taxa de frequência da categoria restantes causas:

$$
\begin{aligned}
l_t &\approx 0,0266(X_t - s_{t-12}) + 0,9734(l_{t-1} + b_{t-1}) \\
b_t &\approx b_{t-1} \\
s_t &\approx 0,4393(X_t - l_t) + 0,5607 s_{t-12} \\
\widehat{X}_{t+h} &\approx l_t + hb_t + s_{t-12+[(h-1) \bmod 12]+1}
\end{aligned}
$$

Os ajustes e previsões pontuais e intervalares (a 95% de confiança) podem ser visualizadas nas Figuras D.7, D.8, D.9 e D.10 do Apêndice D.

6.3 Comparação dos métodos de previsão aplicados aos dados mensais

Após a aplicação dos métodos de modelação aos dados mensais, resta apenas avaliá-los consoante as métricas descritas na Secção 5.3: EQM, REQM, EPAM, EEAM e U de Theil. Para comparar as séries de contagem com as de taxas de frequência, é necessário primeiro transformar os valores ajustados de taxas em contagens. Tal não aconteceu anteriormente mas, para fins de avaliação final, as Tabelas 6.22 e 6.23 mostram os valores das medidas de avaliação todos na mesma unidade (número de sinistros).

Na série do número total de sinistros, nota-se uma clara vantagem na modelação das contagens brutas em relação à modelação da respetiva taxa de frequência. No entanto, o oposto verifica-se nas séries das categorias marginais, com exceção da categoria **TMP**. Isto deve-se, maioritariamente, ao facto de grande parte do comportamento da série da categoria **total** ser justificado pela série da categoria **TMP**. Para além disso, curiosamente, as séries de contagens são, em regra geral, modeladas com mais eficácia por modelos de Holt-Winters e as séries de taxas de frequência, modeladas com mais eficácia por modelos SARIMA. No caso da categoria **restantes causas**, a conclusão não é tão simples, pois, só por si é uma série com baixos valores e representa várias categorias de sinistros (onde cada uma pode ter as suas idiossincrasias), ao contrário das restantes séries.

Não menos importantes que as previsões pontuais, são as intervalares. Estas podem ser avaliadas quanto às suas taxas de cobertura apresentadas nas Tabelas 6.20 e 6.21.

Tabela 6.20: Taxas de cobertura (%) dos intervalos de previsão a 95% de confiança das séries mensais de contagens.

		SARIMA	Holt-Winters
Categoria	**Total**	83,3	66,7
	DNA	83,3	66,7
	REL	100	66,7
	TMP	100	66,7
	Restantes causas	83,3	66,7
	Média	90	66,7

Tabela 6.21: Taxas de cobertura (%) dos intervalos de previsão a 95% de confiança das séries mensais de taxas de frequência.

		SARIMA	Holt-Winters
Categoria	**Total**	100	66,7
	DNA	33,3	66,7
	REL	100	66,7
	TMP	100	66,7
	Restantes causas	83,3	66,7
	Média	83,3	66,7

As taxas de cobertura dos modelos SARIMA apresentam melhores resultados, apesar de não serem muito próximos da cobertura teórica, 95%. O modelo de Holt-Winters mantém-se constante nos 66,7% de taxa de cobertura em todas as séries, o que significa que todos os seus intervalos cobrem 4 das 6 observações dos conjuntos de teste. Pela análise intervalar, não é possível inferir que a modelação das contagens brutas é superior à modelação das taxas de frequência.

Tabela 6.22: Medidas de avaliação calculadas para as séries mensais de contagens, no período de treino e no período de teste, baseadas nos resultados da aplicação das metodologias SARIMA e Holt-Winters.

		Série de treino					**Série de teste**				
Categoria	**Modelo**	EQM	REQM	EPAM	EEAM	U-Theil	EQM	REQM	EPAM	U-Theil	$\sum_{i=1}^{6}(\widehat{x}_i - x_i)$
Total	SARIMA	**32273,5600**	**179,6484**	**9,2620**	**0,6865**	**0,7646**	**36726,0700**	**191,6405**	**10,7347**	**0,8959**	-632
	Holt-Winters	46119,5600	214,7547	12,3648	0,9148	0,9691	63930,5000	252,8448	14,2552	1,3524	**406**
DNA	SARIMA	659463,4000	812,0735	99,8662	9,3913	7,2727	1161746	1077,8430	99,8904	4,8990	-651
	Holt-Winters	**6682,6200**	**81,7473**	**7,6722**	**0,6897**	**0,7052**	**77978,3000**	**279,2460**	**27,9606**	**1,9186**	**580**
REL	SARIMA	83937,4700	289,7196	99,5101	2,5706	2,2406	113287,8000	336,5826	99,5905	1,9522	-465
	Holt-Winters	**15292,9700**	**123,6648**	**33,9573**	**0,8290**	**0,9910**	**13874,1000**	**117,7884**	**37,9106**	**0,5618**	**271**
TMP	SARIMA	41453,4300	203,6012	96,0879	2,8274	2,4724	50387,0100	224,4705	96,6888	2,5180	-117
	Holt-Winters	**6216,2080**	**78,8429**	**35,2402**	**0,9536**	**1,0073**	**3342,1850**	**57,8116**	**22,6928**	**0,8425**	**-58**
Restantes causas	SARIMA	57482,5900	239,7553	99,3477	10,0732	6,7583	58378,6900	241,6168	99,3151	4,8137	**-25**
	Holt-Winters	**643,2234**	**25,3619**	**8,1969**	**0,7999**	**0,6908**	**3242,0160**	**56,9387**	**20,8202**	**1,4665**	114

Tabela 6.23: Medidas de avaliação calculadas para as séries mensais de taxas de frequência, no período de treino e no período de teste, baseadas nos resultados da aplicação das metodologias SARIMA e Holt-Winters.

		Série de treino					**Série de teste**				
Categoria	**Modelo**	EQM	REQM	EPAM	EEAM	U-Theil	EQM	REQM	EPAM	U-Theil	$\sum_{i=1}^{6}(\widehat{x}_i - x_i)$
Total	SARIMA	**37320,4500**	**193,1850**	**11,7850**	**0,8259**	**0,8141**	**34198,3000**	**184,9278**	**11,3650**	**0,9315**	**213**
	Holt-Winters	66926,6600	258,7019	14,4356	1,0629	1,1640	91323,3200	302,1975	16,1894	1,3294	1333
DNA	SARIMA	**6119,5150**	**78,2273**	7,4531	0,7001	**0,6724**	**60943**	**246,8664**	**23,1162**	**1,2618**	**-676**
	Holt-Winters	6900,7190	83,0706	**7,3993**	**0,6769**	0,7035	64517,7800	254,0035	26,7630	1,8112	-6233
REL	SARIMA	**9801,5590**	**99,0028**	**27,2957**	**0,6930**	**0,8034**	**10001,8600**	**100,0093**	**32,9029**	**0,5805**	**-202**
	Holt-Winters	15653,0400	125,1121	33,7549	0,8264	1,0078	15604,9500	124,9198	39,0561	0,5887	-1954
TMP	SARIMA	**6740,1890**	**82,0987**	**42,6489**	**1,0145**	**1,1013**	**3906,4080**	**62,5013**	**26,2582**	**0,6932**	**189**
	Holt-Winters	20459,5700	143,0370	52,2334	1,5098	1,4690	7784,0540	88,2273	41,9856	1,0004	-1331
Restantes causas	SARIMA	**655,8402**	**25,6094**	9,0421	0,9003	**0,7186**	**2765,3000**	**52,5861**	21,7200	**1,2717**	**123**
	Holt-Winters	691,6101	26,2985	**8,5509**	**0,8381**	0,7194	3297,0970	57,4204	**20,9665**	1,4682	-1428

Capítulo 7

Conclusões

O estudo apresentado nesta dissertação dedicou-se à análise de séries temporais de números de sinistros habitação. Originalmente, os dados apresentavam registos desde 1995, contudo este trabalho focou-se apenas no período de janeiro de 2015 até junho de 2021, dividindo-os em 5 categorias (**total**, **DNA**, **REL**, **TMP** e **restantes causas**). O objetivo principal deste projeto foi criar modelos de previsão que pudessem ser utilizados na prática, de forma a auxiliar na tomada de decisões e no planeamento estratégico dentro da empresa Ageas. Para tal foi realizado um estudo comparativo entre três metodologias de análise de séries temporais para os dados diários: metodologia clássica de Box-Jenkins, SARIMA, métodos de alisamento exponencial de Holt-Winters e uma mais recente, desenvolvida por Robin Hyndman, TBATS. Também foi efetuado um estudo no espectro mensal, mas este não incluiu modelos TBATS, pois não havia necessidade, uma vez que a metodologia TBATS foi criada propositadamente para casos como o das séries diárias (sazonalidades múltiplas e alta variabilidade das séries). De modo a avaliarem-se os desempenhos dos modelos, foram comparados os respetivos valores de cinco métricas de avaliação distintas (EQM, REQM, EPAM, EEAM e estatística U-Theil). Foram também construídos e avaliados pelas suas taxas de cobertura, intervalos de previsão para cada metodologia.

Da análise diária, não foi possível concluir de forma clara qual a metodologia que mais se destacou. Contudo, os modelos TBATS apresentaram, em geral, melhor ajuste às séries de treino e alguns menores valores das medidas de avaliação nos conjuntos de teste. Estes foram, também, os modelos cujos intervalos de previsão apresentaram maiores taxas de cobertura, apesar de os intervalos de previsão de Holt-Winters também terem exibido ótimos resultados (o que já se esperava, tendo em conta o método de construção dos mesmos). No entanto, é importante salientar que estes estudos foram realizados relaxando alguns pressupostos para a aplicação das metodologias (aplicação a dados reais); fazendo, assim, com que os modelos TBATS sejam os mais apropriados para este tipo de dados de séries temporais. É mais eficiente modelar a série dos números totais de sinistros do que modelar a soma de todas as séries marginais. Esta conclusão pode ser retirada

através da última coluna da Tabela 5.17 onde se vê que o valor correspondente à série **total** é inferior à soma de todos os restantes (dentro da última coluna da tabela).

Da análise mensal, pode-se concluir que as categorias **total** e **TMP** apresentaram melhores valores de ajuste e previsão quando modeladas pelas suas taxas de frequência, ao contrário das restantes séries. Notou-se uma forte correlação entre estas duas categorias, de onde se pode inferir que grande parte dos sinistros registados são derivados de tempestades. O método de alisamento exponencial de Holt-Winters realçou-se nesta abordagem uma vez que, em geral, as séries de contagens apresentaram melhores valores das medidas de avaliação quando modeladas por esta metodologia. Não obstante, a teoria de Box-Jenkins mostrou-se, também, eficaz na modelação das séries das taxas de frequência e nas previsões intervalares. Porém, a categoria **restantes causas** engloba várias subcategorias menos relevantes para o estudo, o que torna a sua modelação mais complicada, pois cada subcategoria tem o seu próprio processo gerador. Mais uma vez, conclui-se que é mais eficiente modelar a série dos números totais de sinistros do que modelar a soma de todas as séries marginais, pelo mesmo motivo explicado nas conclusões da análise diária.

Em suma, a metodologia TBATS mostrou a sua utilidade na modelação de séries com comportamentos mais voláteis, sobrepondo-se à metolodiga clássia SARIMA e ao alisamento exponencial de Holt-Winters que não mostraram tanta capacidade para descrever os comportamentos das séries diárias. No entanto, de entre as três metodologias, os modelos de Holt-Winters, por serem modelos não paramétricos não exigindo tantos pressupostos, foi o método com melhores previsões intervalares em termos diários e mensais.

7.1 Trabalho Futuro

Com o estudo terminado, é sempre possível atualizar as séries acrescentando os novos números de sinistros registados, originando novos conjuntos de treino e de teste e obtendo melhores resultados. Ficam também algumas abordagens e investigações por fazer que não encaixaram no âmbito desta dissertação, nomeadamente modelos de regressão linear múltipla que têm bastante utilidade neste setor, uma vez que os números de sinistros dependem fortemente de muitos fatores externos, tais como: pluviosidade, temperatura, vento, humidade, região geográfica, etc.. Como se concluiu que as tempestades são a causa mais agravante para os números registados de sinistros, todas estas variáveis, consequentemente, devem ter um enorme peso nos processos de modelação temporal. O factor da região geográfica também pode ser explorado com ferramentas de análise espacial, uma vez que zonas mais ventosas ou chuvosas estão mais sujeitas a maiores registos de sinistros, tornando viável um estudo de modelação espacial.

Machine Learning é também uma área que ultimamente se tem destacado na qual se pretende obter previsões de séries temporais, apresentando ferramentas com grande utilidade, entre as quais

se realçam as redes neuronais artificiais. Contudo, estes métodos apresentam certas limitações, apesar de terem a enorme vantagem de serem métodos que vão treinando ao longo do tempo, logo obtendo melhores resultados. Isto tem motivado estudos recentes a escolherem, também, métodos paramétricos em vez de não paramétricos, que podem vir a ter sucesso no contexto de previsões de séries temporais.

Bibliografia

[Aburto and Weber, 2007] Aburto, L. and Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. Applied Soft Computing, 7(1):136–144.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6):716–723.

[Alon et al., 2001] Alon, I., Qi, M., and Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. Journal of retailing and consumer services, 8(3):147–156.

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

[Anderson, 1976] Anderson, O. D. (1976). Time series analysis and forecasting: the Box-Jenkins approach. Butterworths, first edition.

[Arca et al., 2006] Arca, B., Spano, D., Snyder, R., Fiori, M., and Duce, P. (2006). Short term forecasting of reference evapotranspiration using limited area models and time series techniques. In 27th Conference on Agricultural and Forest Meteorology.

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

[Box et al., 2016] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2016). Time series analysis: forecasting and control. John Wiley & Sons, fifth edition.

[Brown, 1960] Brown, R. G. (1960). Statistical forecasting for inventory control. Journal of the Royal Statistical Society.

[Brożyna et al., 2018] Brożyna, J., Mentel, G., Szetela, B., and Strielkowski, W. (2018). Multiseasonality in the tbats model using demand for electric energy as a case study. Economic Computation & Economic Cybernetics Studies & Research, 52(1).

[Caiado, 2011] Caiado, J. (2011). Métodos de previsão em gestão-com aplicações em excel. Edições Sílabo, Lisboa.

[Chatfield, 2000] Chatfield, C. (2000). Time-series forecasting. Chapman and Hall/CRC, first edition.

[Chatfield, 2003] Chatfield, C. (2003). The analysis of time series: an introduction. Chapman and hall/CRC, sixth edition.

[Chen et al., 1996] Chen, C., Davis, R. A., and Brockwell, P. J. (1996). Order determination for multivariate autoregressive processes using resampling methods. Journal of multivariate analysis, 57(2):175–190.

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

[Chu and Zhang, 2003] Chu, C.-W. and Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of production economics, 86(3):217–231.

[Clemen, 1989] Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International journal of forecasting, 5(4):559–583.

[Clevele and Terpenning, 1982] Clevele, W. S. and Terpenning, I. J. (1982). Graphical methods for seasonal adjustment. Journal of the American Statistical Association, 77(377):52–62.

[Cordeiro, 2011] Cordeiro, C. M. H. (2011). Métodos de reamostragem em modelos de previsão. PhD thesis, Universidade de Lisboa.

[Cowpertwait and Metcalfe, 2009] Cowpertwait, P. S. and Metcalfe, A. V. (2009). Introductory time series with R. Springer Science & Business Media, first edition.

[da Veiga et al., 2016] da Veiga, C. P., da Veiga, C. R. P., Puchalski, W., dos Santos Coelho, L., and Tortato, U. (2016). Demand forecasting based on natural computing approaches applied to the foodstuff retail segment. Journal of Retailing and Consumer Services, 31:174–181.

[De Gooijer and Hyndman, 2006] De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. International journal of forecasting, 22(3):443–473.

[De Livera et al., 2011] De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American statistical association, 106(496):1513–1527.

[Dickey and Fuller, 1979] Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, 74(366a):427–431.

[Efron, 1992] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics, volume II, pages 569–593. Springer.

[Enders, 2015] Enders, W. (2015). Applied econometric time series fourth edition. New York (US): University of Alabama.

[Gardner Jr and McKenzie, 1985] Gardner Jr, E. S. and McKenzie, E. (1985). Forecasting trends in time series. Management science, 31(10):1237–1246.

[Harvey and Fernandes, 1989] Harvey, A. and Fernandes, C. (1989). Time series models for insurance claims. Journal of the Institute of Actuaries, 116(3):513–528.

[Holt, 1957] Holt, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages. ONR Memorandum, 52(52):5–10.

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

[Hyndman and Athanasopoulos, 2013] Hyndman, R. J. and Athanasopoulos, G. (2013). Forecasting: principles and practice. OTexts, first edition.

[Hyndman et al., 2006] Hyndman, R. J. et al. (2006). Another look at forecast-accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting, 4(4):43–46.

[Hyndman and Koehler, 2006] Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4):679–688.

[Jebb et al., 2015] Jebb, A. T., Tay, L., Wang, W., and Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. Frontiers in psychology, 6:727.

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

[Kuvulmaz et al., 2005] Kuvulmaz, J., Usanmaz, S., and Engin, S. N. (2005). Time-series forecasting by means of linear and nonlinear models. In Mexican International Conference on Artificial Intelligence, pages 504–513. Springer.

[Kwiatkowski et al., 1992] Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of econometrics, 54(1-3):159–178.

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

[Lima, 2018] Lima, S. M. R. (2018). Métodos de previsão de séries temporais: uma aplicação a dados do segmento do retalho. Master's thesis, Universidade do Minho.

[Murteira et al., 1993] Murteira, B., Müller, D., and Turkman, K. F. (1993). Análise de sucessões cronológicas. McGraw-Hill.

[Naim et al., 2018] Naim, I., Mahara, T., and Idrisi, A. R. (2018). Effective short-term forecasting for daily time series with complex seasonal patterns. Procedia computer science, 132:1832–1841.

[Ng and Perron, 1995] Ng, S. and Perron, P. (1995). Unit root tests in arma models with data-dependent methods for the selection of the truncation lag. Journal of the American Statistical Association, 90(429):268–281.

[Pan, 2013] Pan, Y. (2013). Predicting aggregate retail sales using hybrid arima. In Proceedings of the 7th Global Business and Social Science Research Conference. Métodos de previsao de séries temporais–uma aplicaçao a dados do segmento do Retalho.

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

[Phillips and Perron, 1988] Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75(2):335–346.

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

[Ramos et al., 2015] Ramos, P., Santos, N., and Rebelo, R. (2015). Performance of state space and arima models for consumer retail sales forecasting. Robotics and computer-integrated manufacturing, 34:151–163.

[Said and Dickey, 1984] Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 71(3):599–607.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, pages 461–464.

[Schwert, 2002] Schwert, G. W. (2002). Tests for unit roots: A monte carlo investigation. Journal of Business & Economic Statistics, 20(1):5–17.

[Shu et al., 2014] Shu, M.-H., Hung, W., Nguyen, T.-L., Hsu, B., and Lu, C. (2014). Forecasting with fourier residual modified arima model-an empirical case of inbound tourism demand in new zealand. WSEAS Transactions on Mathematics, 13(1):12–21.

[Shumway et al., 2000] Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). Time series analysis and its applications, volume 3. Springer, fourth edition.

[Silva, 2013] Silva, J. I. M. d. (2013). A metodologia bootstrap associada ao método de holt-winters na previsão de séries temporais. Master's thesis, Universidade do Minho.

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

[Tong, 1990] Tong, H. (1990). Non-linear time series: a dynamical system approach. Oxford university press, first edition.

[Tsay, 2000] Tsay, R. S. (2000). Time series and forecasting: Brief history and future research. Journal of the American Statistical Association, 95(450):638–643.

[Udny Yule, 1927] Udny Yule, G. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London Series A, 226:267–298.

[Wang and Cai, 2009] Wang, D. and Cai, X. (2009). Irrigation scheduling—role of weather forecasting and farmers' behavior. Journal of water resources planning and management, 135(5):364–372.

[Wang, 2006] Wang, S. (2006). Exponential smoothing for forecasting and Bayesian validation of computer models. PhD thesis, Georgia Institute of Technology.

[West and Harrison, 1989] West, M. and Harrison, J. (1989). Subjective intervention in formal models. Journal of Forecasting, 8(1):33–53.

[Wheelwright et al., 1998] Wheelwright, S., Makridakis, S., and Hyndman, R. J. (1998). Forecasting: methods and applications. John Wiley & Sons, third edition.

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

[Wold, 1938] Wold, H. (1938). A study in the analysis of stationary time series. PhD thesis, Almqvist & Wiksell.

[Yokuma and Armstrong, 1995] Yokuma, J. T. and Armstrong, J. S. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. International Journal of Forecasting, 11(4):591–597.

[Yule, 1926] Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. Journal of the royal statistical society, 89(1):1–63.

[Zhang, 2003] Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. Neurocomputing, 50:159–175.

[Zhang and Qi, 2005] Zhang, G. P. and Qi, M. (2005). Neural network forecasting for seasonal and trend time series. European journal of operational research, 160(2):501–514.

Apêndice A

Comportamentos teóricos das FAC e FACP de modelos de Box-Jenkins

Tabela A.1: Comportamentos teóricos das FAC e FACP de modelos de Box-Jenkins.

Modelo	**FAC**	**FACP**
AR(p)	Decrescimento exponencial para zero	Queda abrupta para zero a partir do $lag = p + 1$
MA(q)	Queda abrupta para zero a partir do $lag = p + 1$	Decrescimento exponencial para zero
ARMA(p, q)	Decrescimento exponencial para zero	Decrescimento exponencial para zero
SAR$(P)_s$	Decrescimento exponencial para zero com periodicidade s	Queda abrupta para zero a partir do $lag = (P + 1)_s$
SMA$(Q)_s$	Queda abrupta para zero a partir do $lag = (Q + 1)_s$	Decrescimento exponencial para zero com periodicidade s
SARMA$(P, Q)_s$	Decrescimento exponencial para zero com periodicidade s	Decrescimento exponencial para zero com periodicidade s
SARMA$(p, q)(P, Q)_s$	Decrescimento exponencial para zero	Decrescimento exponencial para zero

Apêndice B

Medidas descritivas dos dados diários

Tabela B.1: Medidas descritivas dos números diários de sinistros no período observado.

	Categoria				
	DNA	**REL**	**TMP**	**Restantes causas**	**TOTAL**
Início	01/01/1995	01/01/1995	01/01/1995	01/01/1995	01/01/1995
Fim	30/06/2021	30/06/2021	30/06/2021	30/06/2021	30/06/2021
Dimensão	9678	9678	9678	9678	9678
N.º de zeros	6995	7208	7305	318	309
Amplitude	0 - 118	0 - 372	0 - 2225	0 - 2702	0 - 2703
$Q_{0,25}$	0	0	0	3	4
Mediana	0	0	0	7	13
$Q_{0,75}$	6	1	0	14	35
Média	6,77	2,25	2,27	13,24	24,54
Desvio Padrão	13,39	8,58	26,43	34,93	49,80
Variância	179,37	73,53	698,71	1219,82	2480,32
Coeficiente de Variação	1,98	3,81	11,66	2,64	2,03
N.º de outliers	2009	2125	2373	1021	363

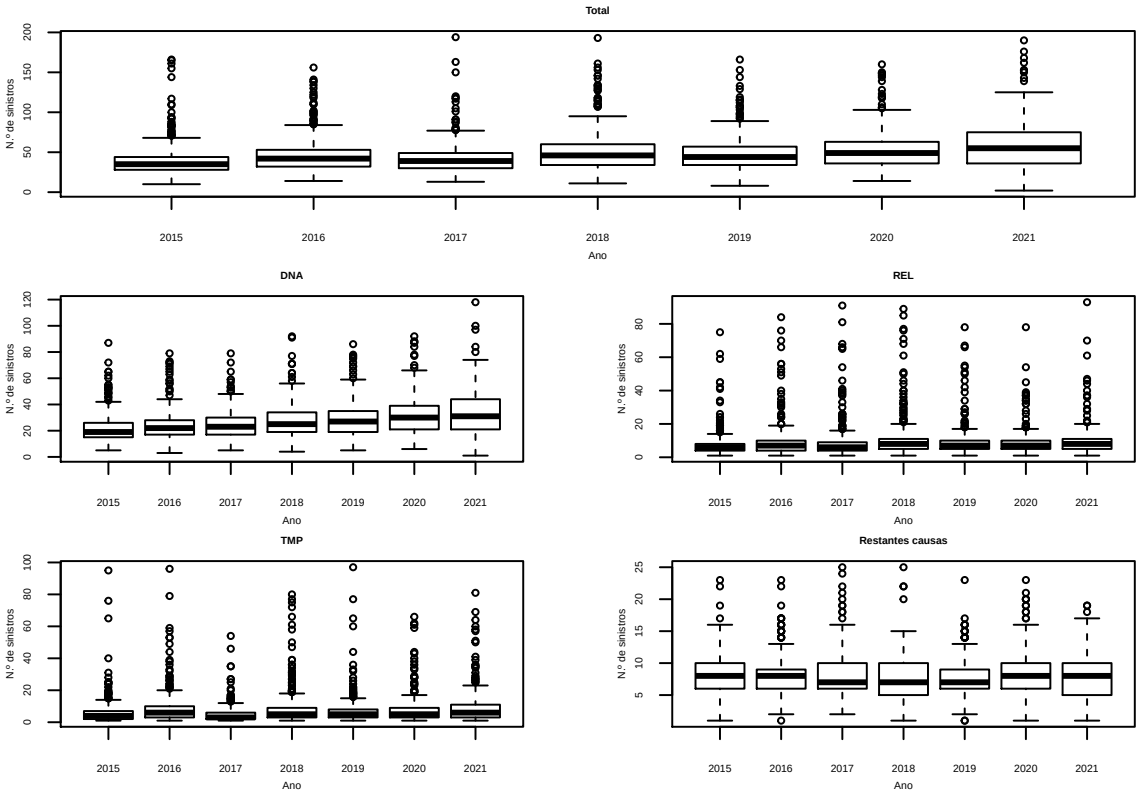


Figura B.1: Diagramas em caixas de bigodes das séries diárias com outliers.

Apêndice C

Aplicação dos métodos de previsão aos dados diários

Tabela C.1: Características do modelo TBATS considerando sazonalidade complexa (semanal e anual simultaneamente) ajustado à série diária do número total de sinistros após uma transformação de Box-Cox com λ = -0,184.

	TBATS(1 , {3,2} , 0,931 , {7,3} , {365,25 , 4})						$\widehat{\sigma} =$ **0,1795**		$AIC =$ **9391,9278**	
Parâmetro	α	$\gamma_1^{(7)}$	$\gamma_1^{(365,25)}$	$\gamma_2^{(7)}$	$\gamma_2^{(365,25)}$	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2
Estimativa	0,0383	0,0003	-0,0001	-0,0002	-0,0002	0,2208	-0,8175	0,1581	0,0051	0,8686

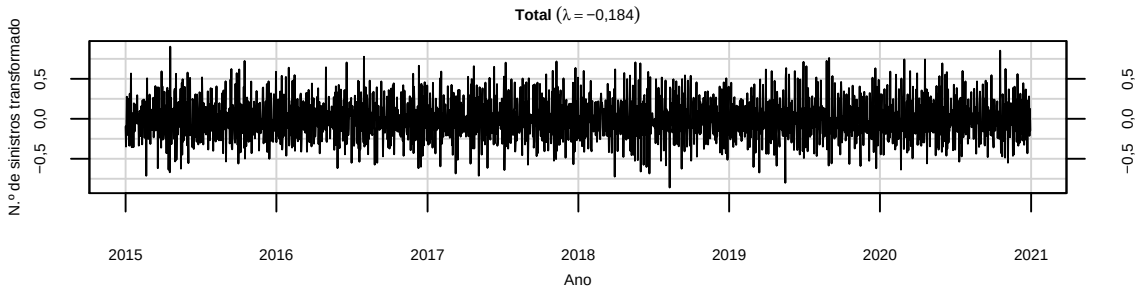


Figura C.1: Série diária da categoria **total** após aplicada uma transformação de Box-Cox seguida de uma diferenciação de 1.ª ordem ($d = 1$).

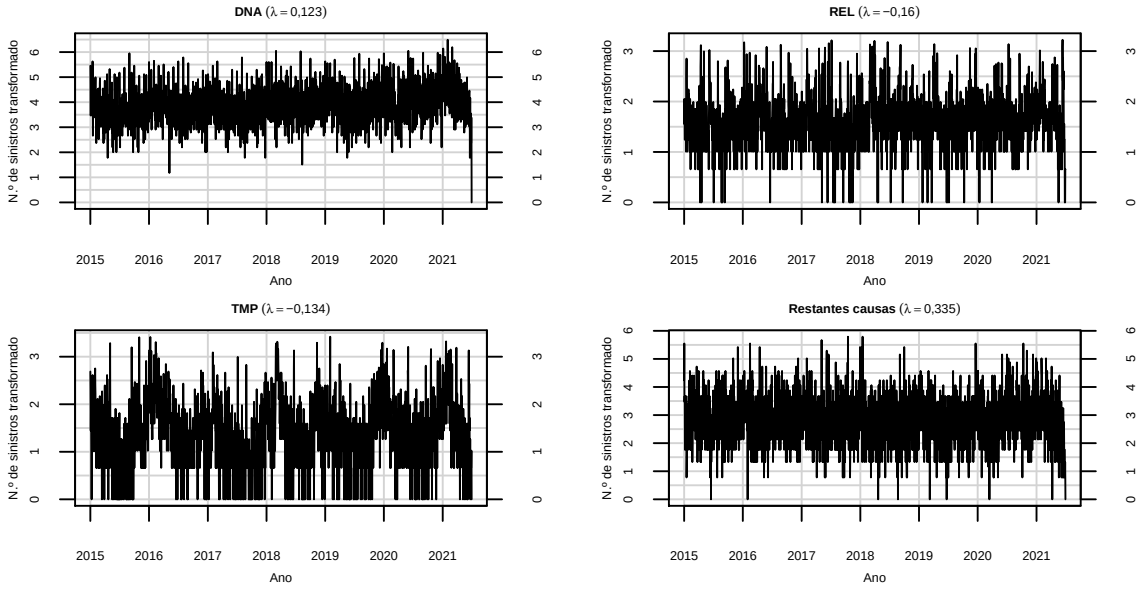


Figura C.2: Representação gráfica das séries diárias marginais após as transformações de Box-Cox.

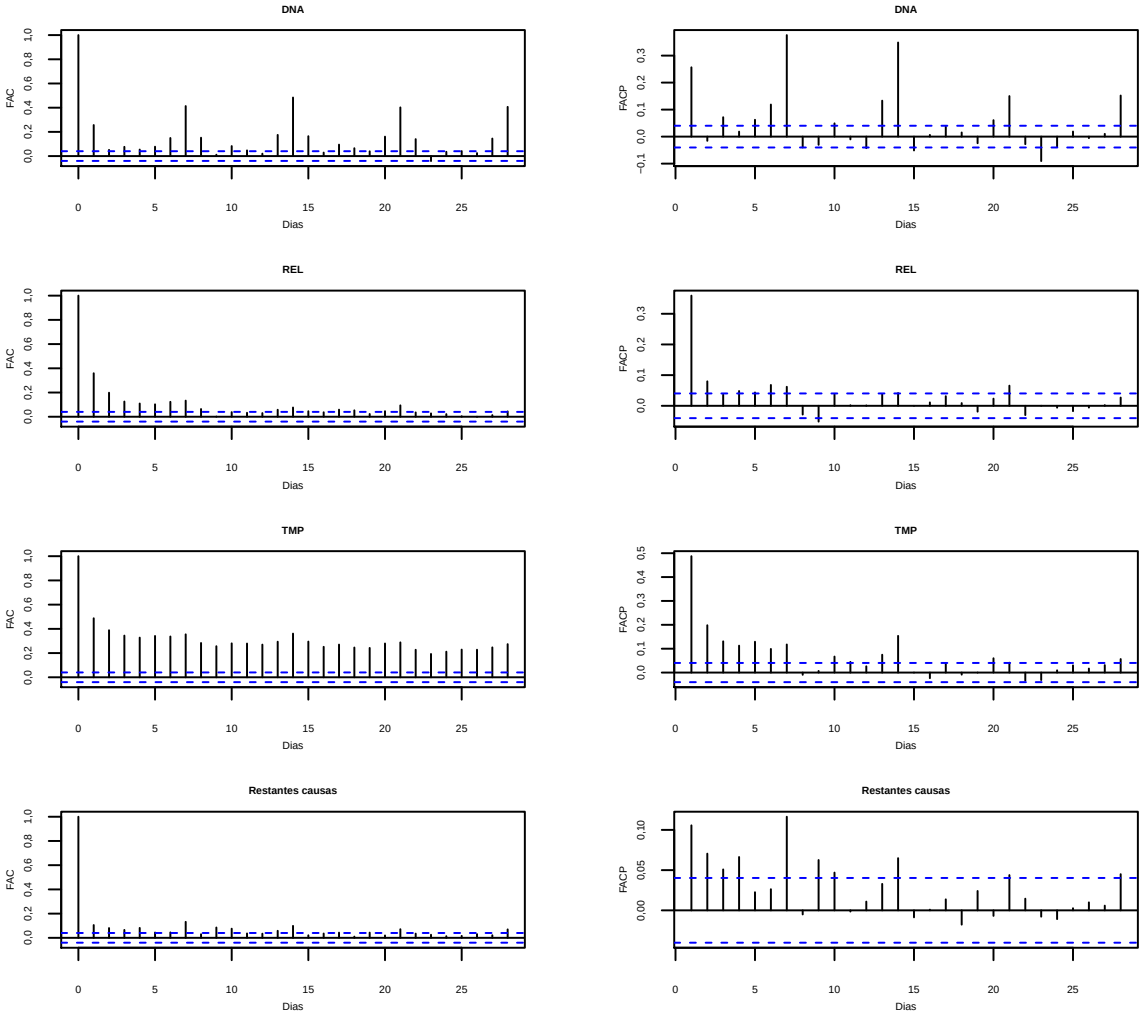


Figura C.3: FAC e FACP das séries diárias marginais após a transformação de Box-Cox.

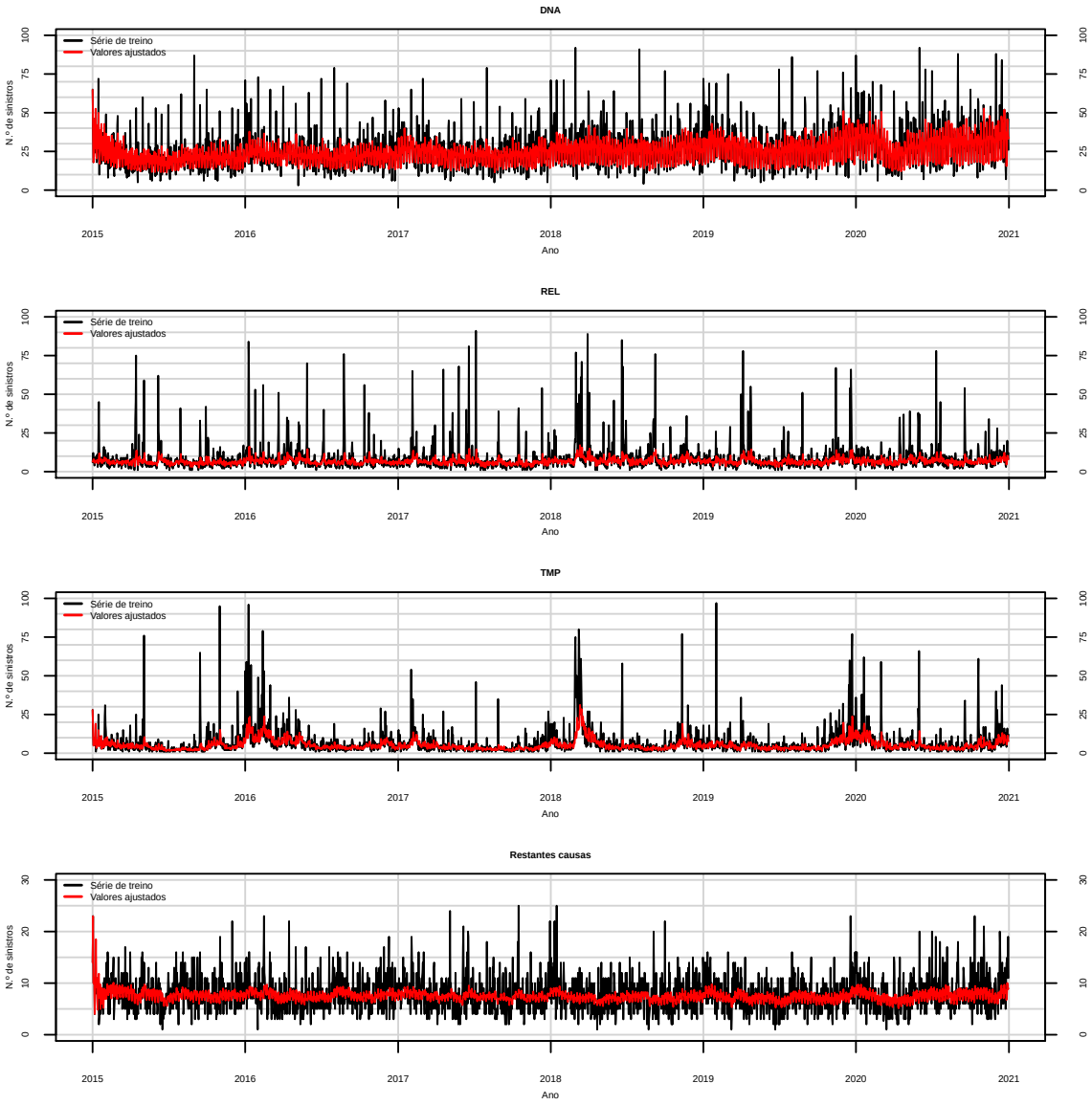


Figura C.4: Ajuste dos modelos SARIMA das séries diárias marginais.

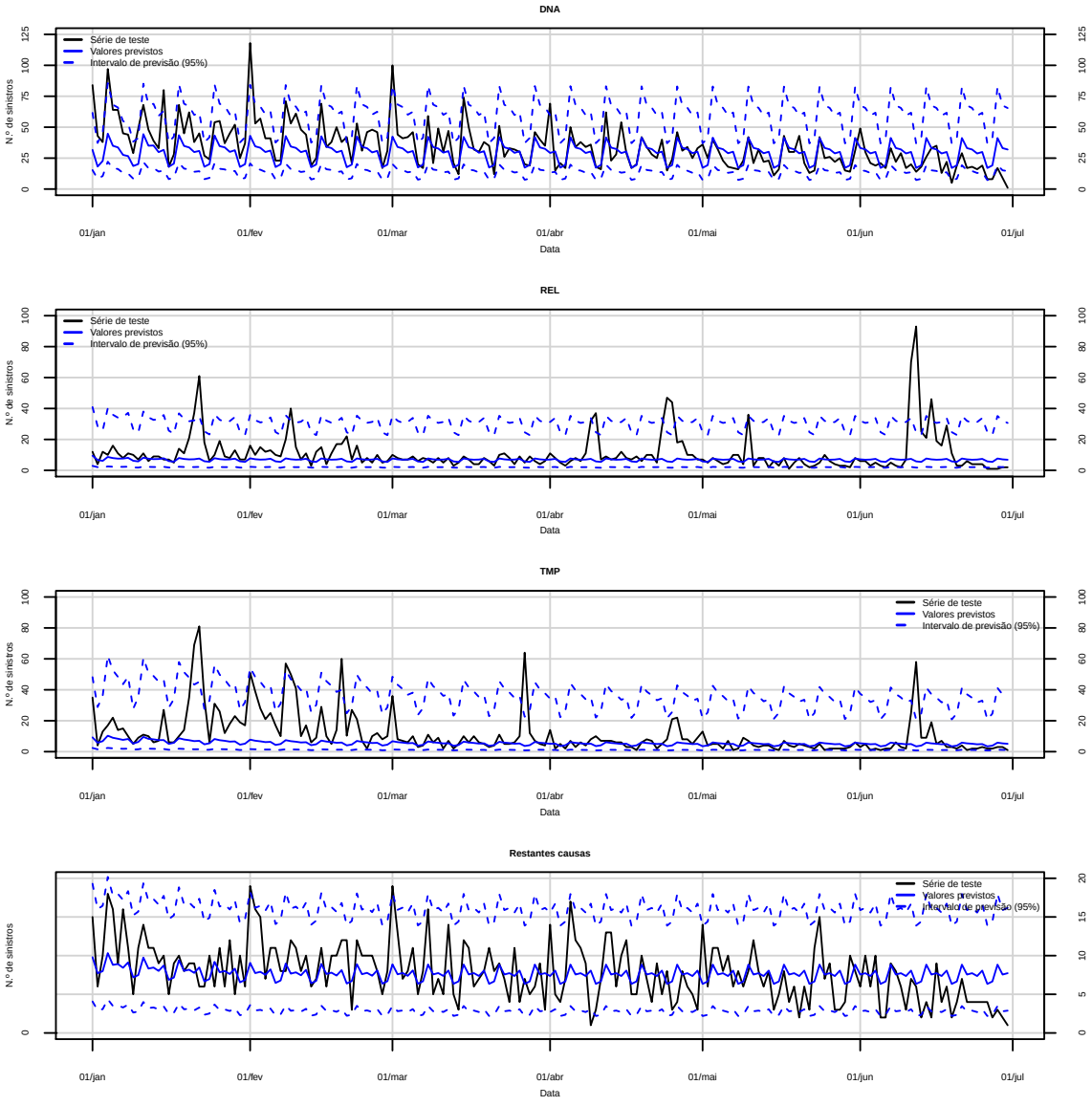


Figura C.5: Previsões pontuais e intervalares (a 95% de confiança) dos modelos SARIMA ajustados às séries diárias marginais.

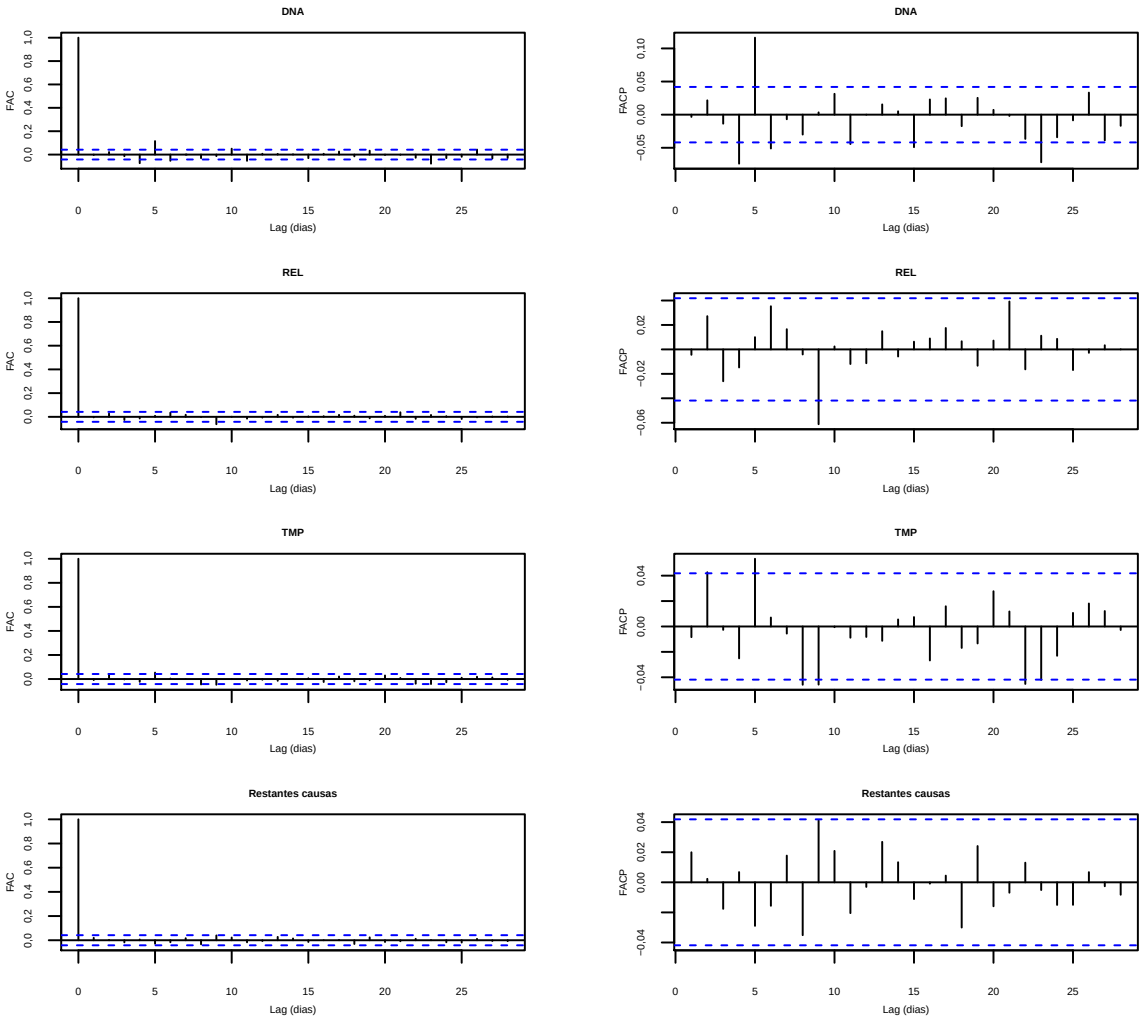


Figura C.6: FAC e FACP dos resíduos dos modelos SARIMA ajustados às séries diárias marginais.

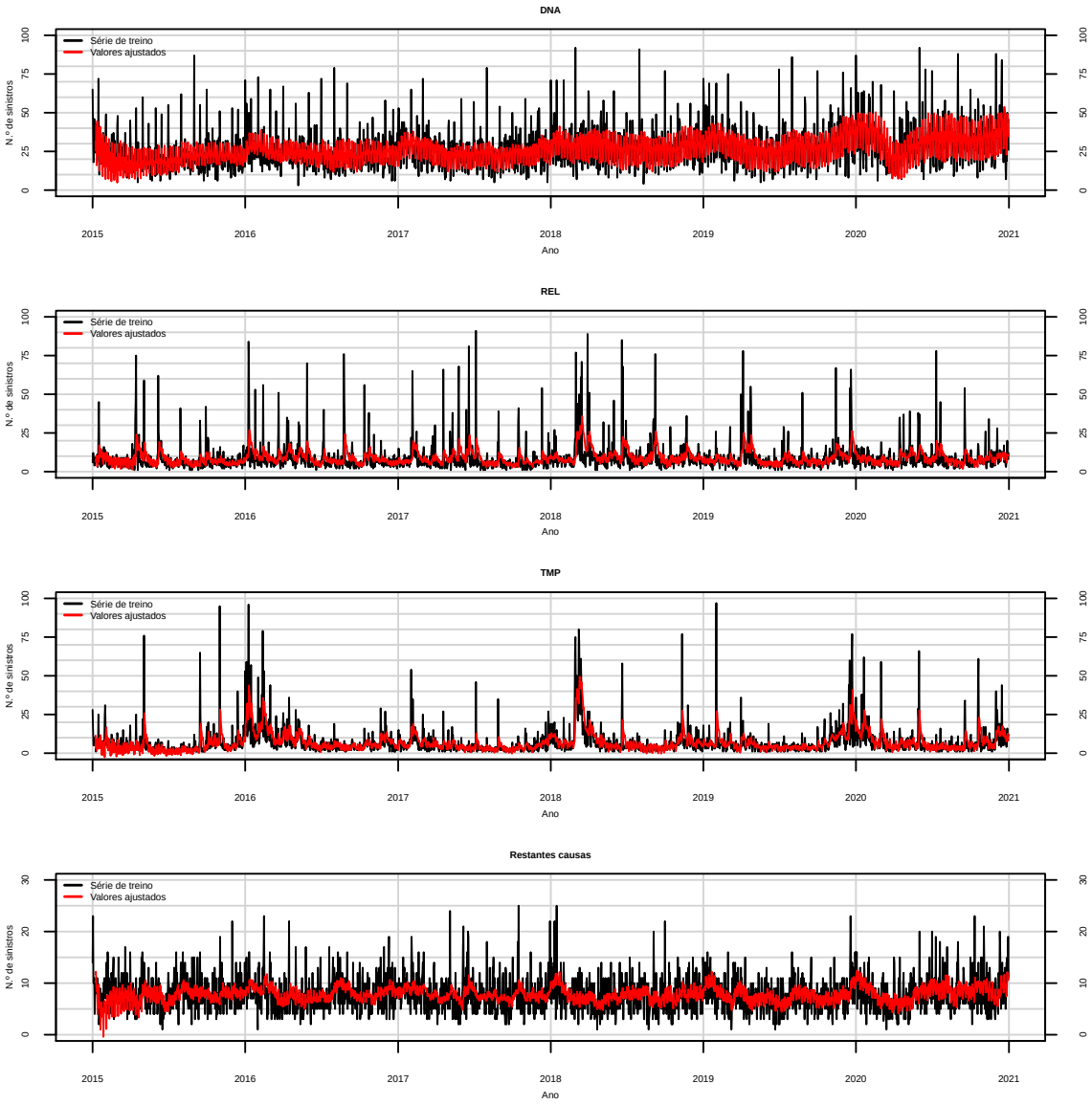


Figura C.7: Ajuste dos modelos Holt-Winters das séries diárias marginais.

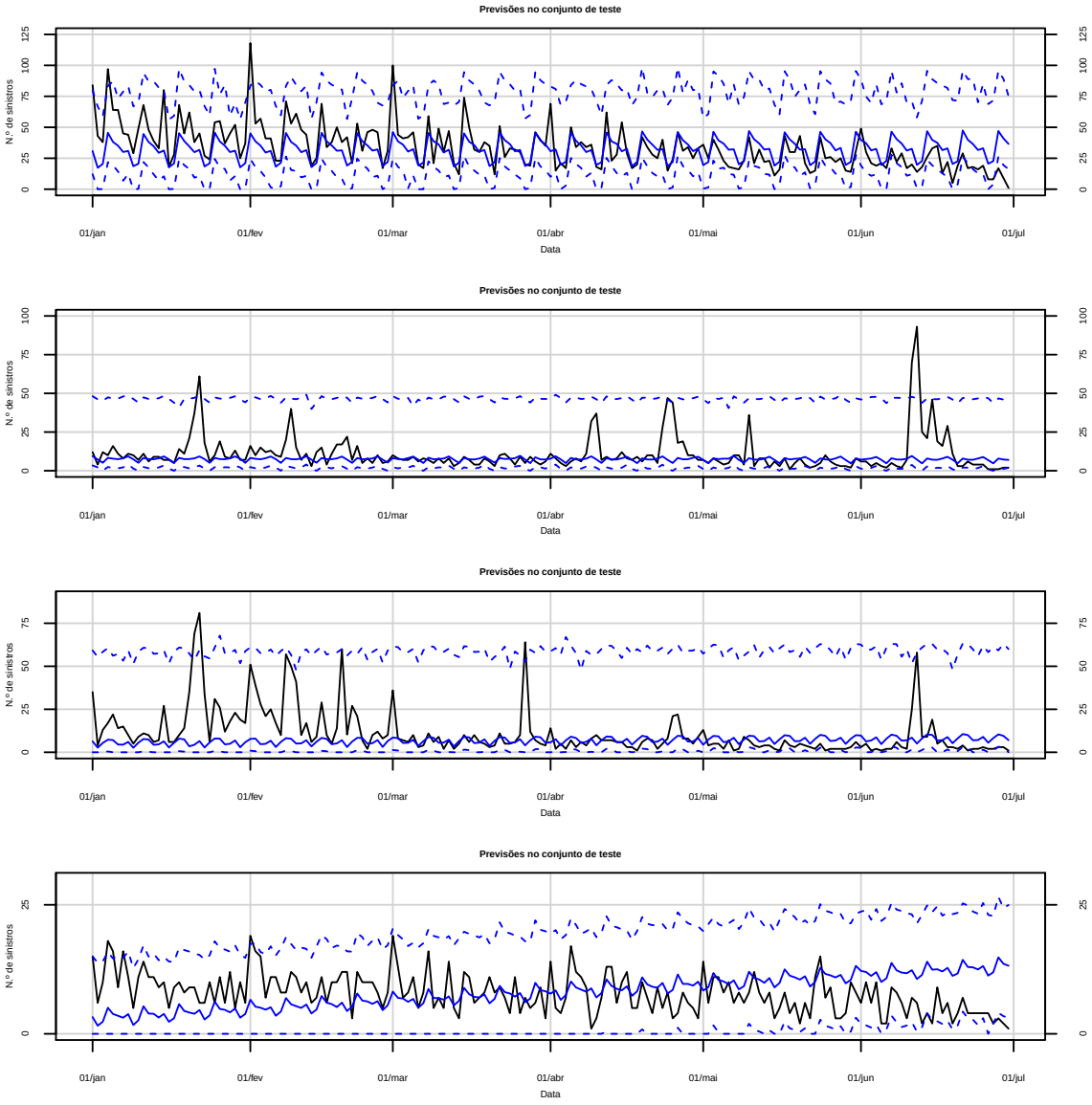


Figura C.8: Previsões pontuais e intervalares (a 95% de confiança) dos modelos Holt-Winters ajustados às séries diárias marginais.

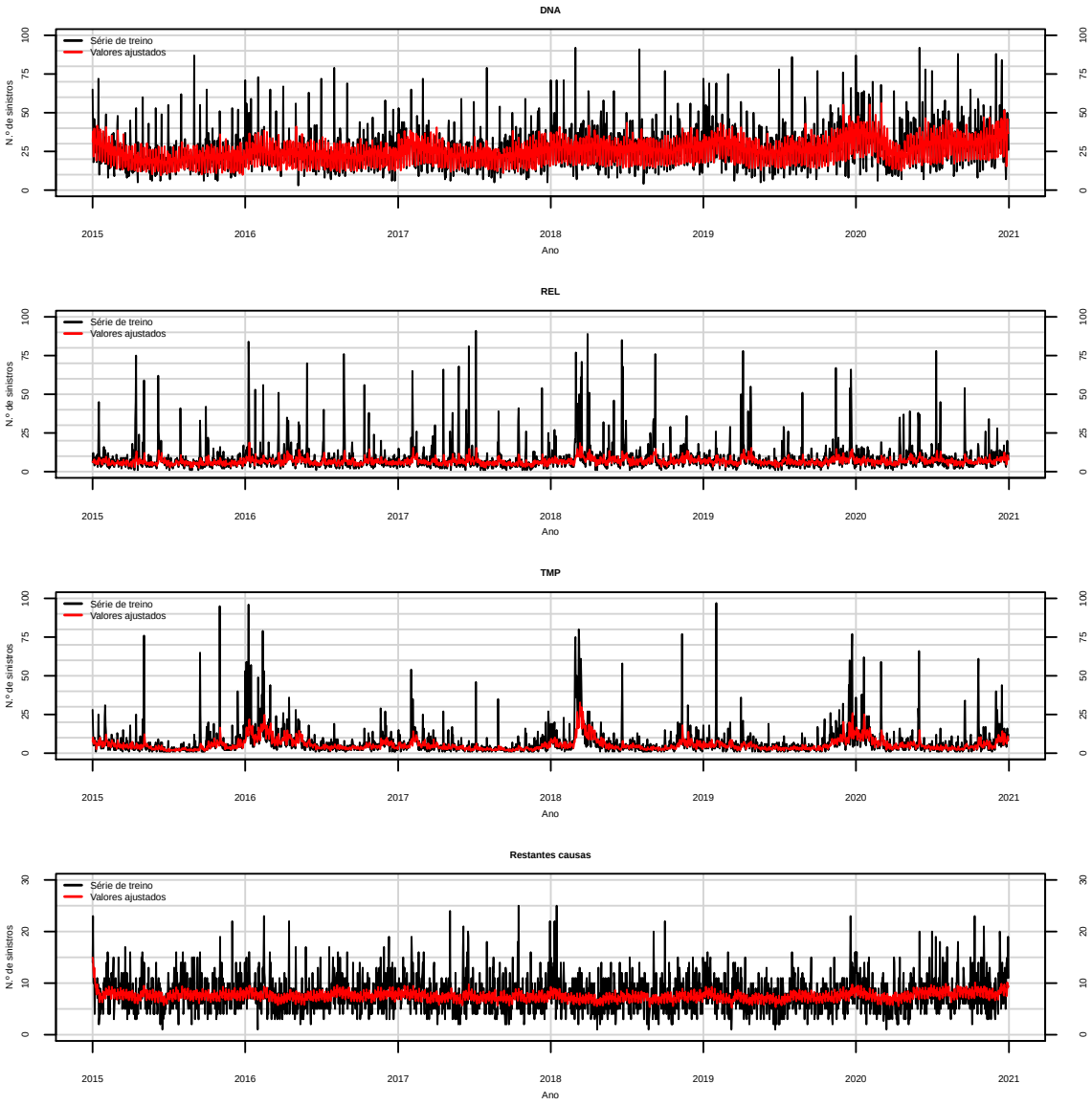


Figura C.9: Ajuste dos modelos TBATS das séries diárias marginais.

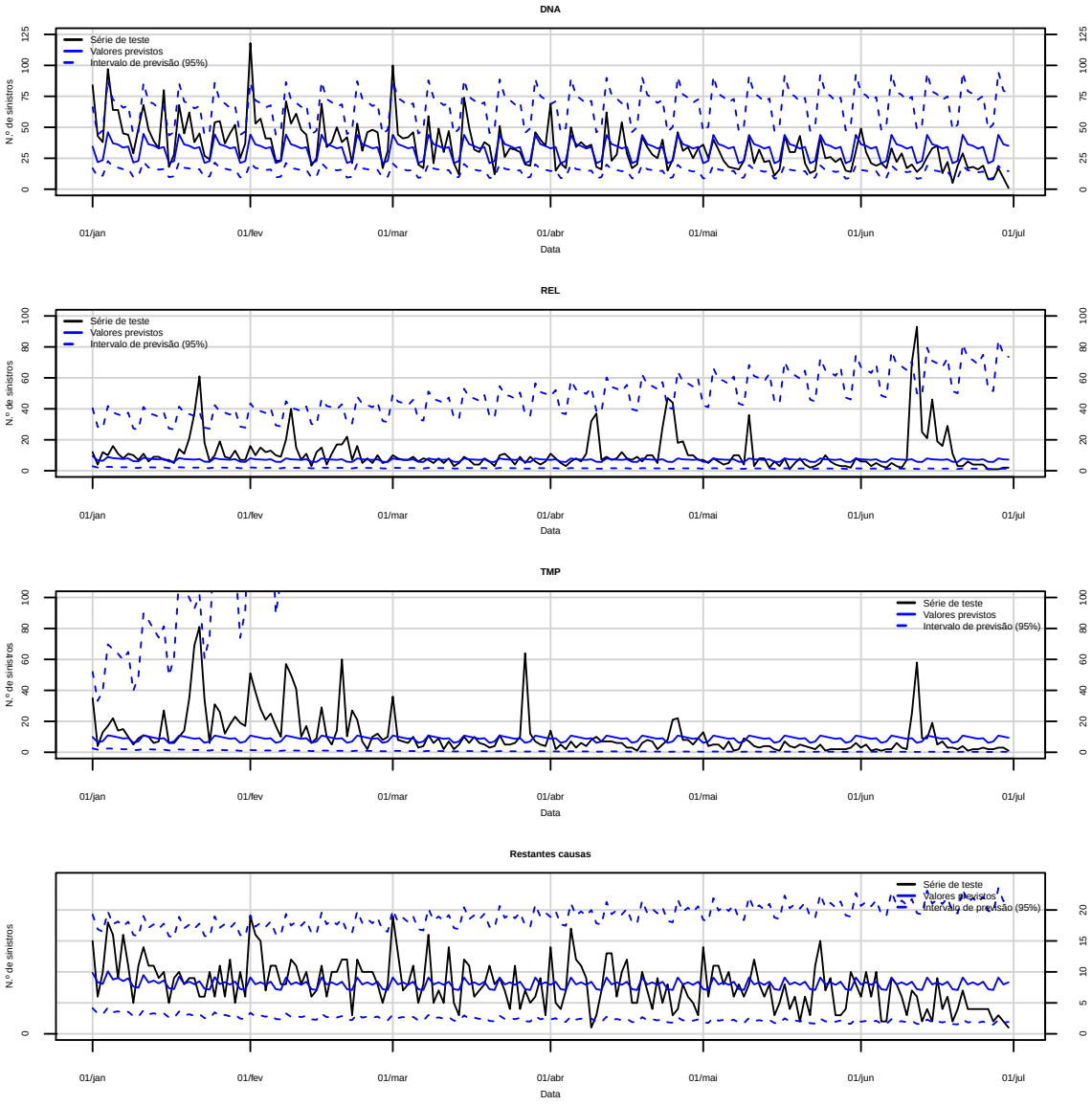


Figura C.10: Previsões pontuais e intervalares (a 95% de confiança) dos modelos TBATS ajustados às séries diárias marginais.

Apêndice D

Aplicação dos métodos de previsão aos dados mensais

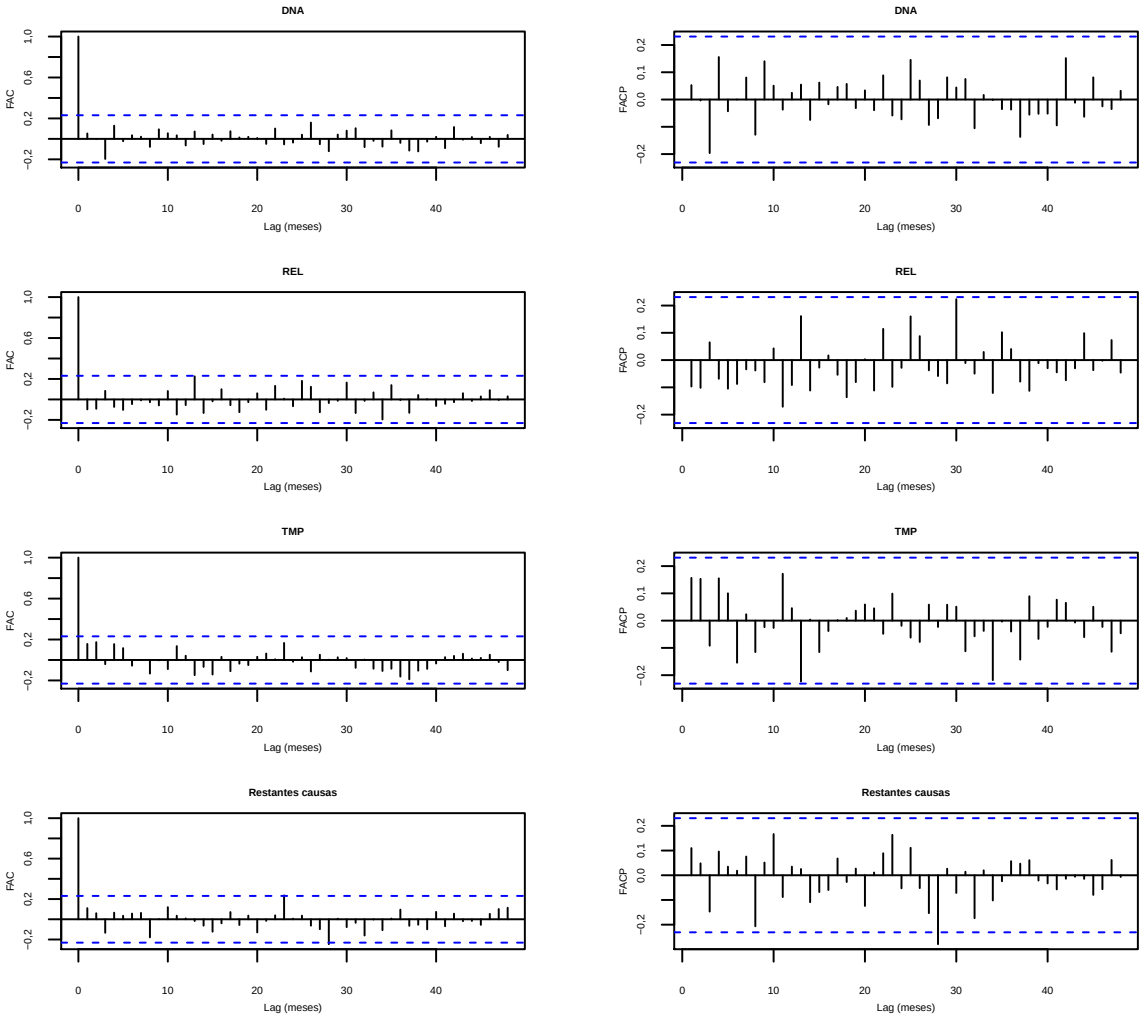


Figura D.1: FAC e FACP dos resíduos dos modelos SARIMA ajustados às séries mensais de contagens marginais.

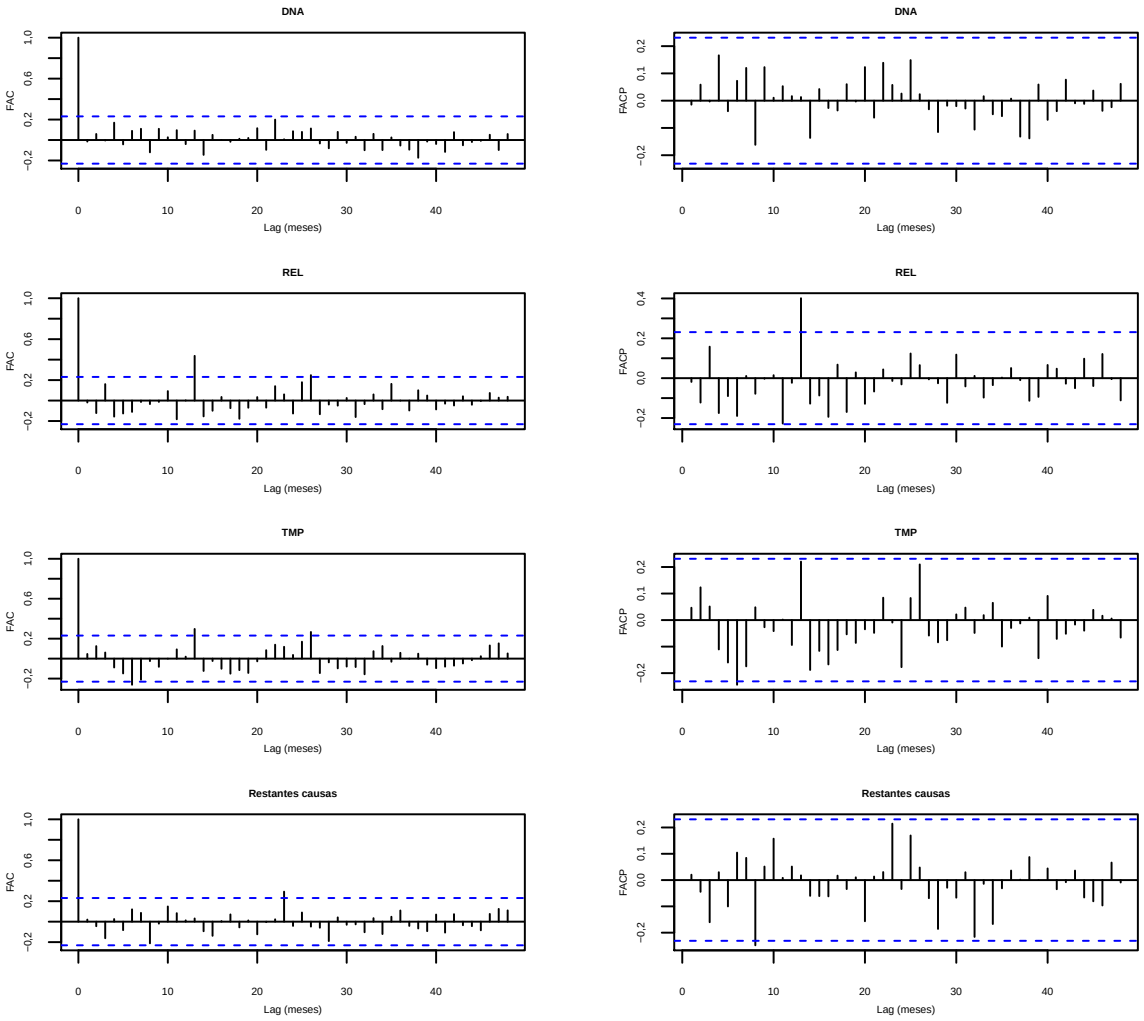


Figura D.2: FAC e FACP dos resíduos dos modelos SARIMA ajustados às séries mensais de taxas de frequência marginais.

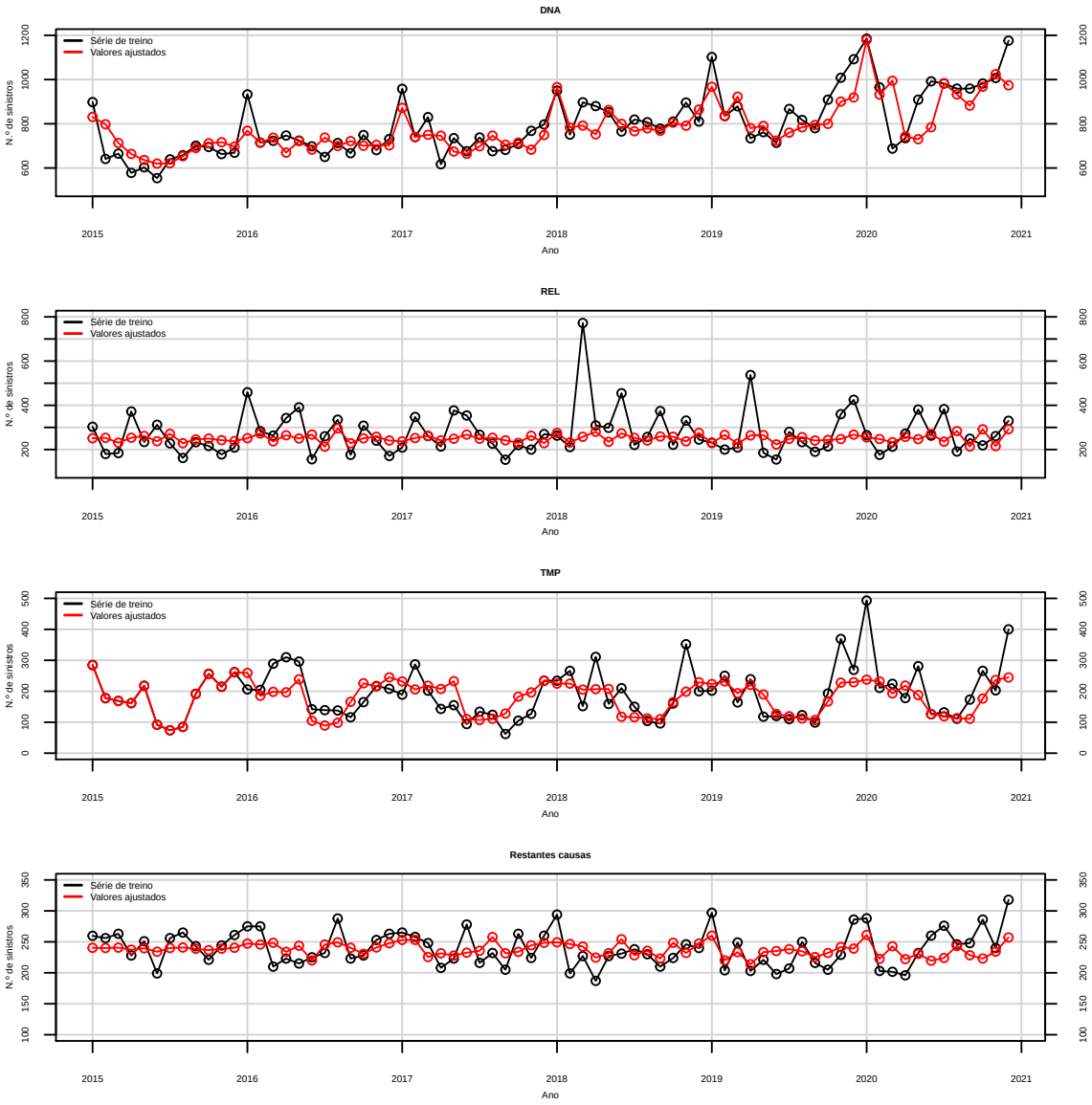


Figura D.3: Ajuste dos modelos SARIMA das séries mensais de contagens marginais.

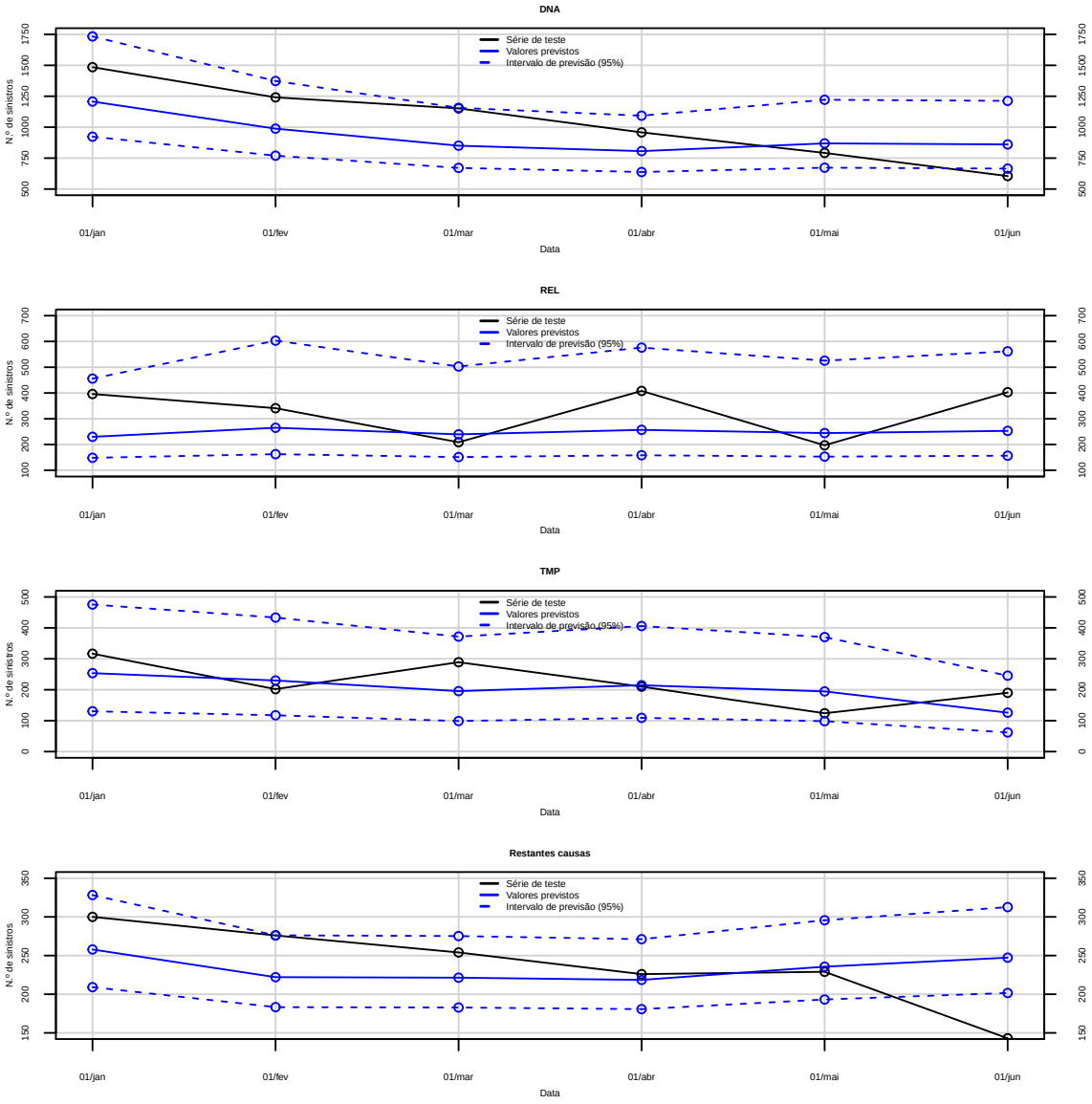


Figura D.4: Previsões pontuais e intervalares (a 95% de confiança) dos modelos SARIMA ajustados às séries mensais de contagens marginais.

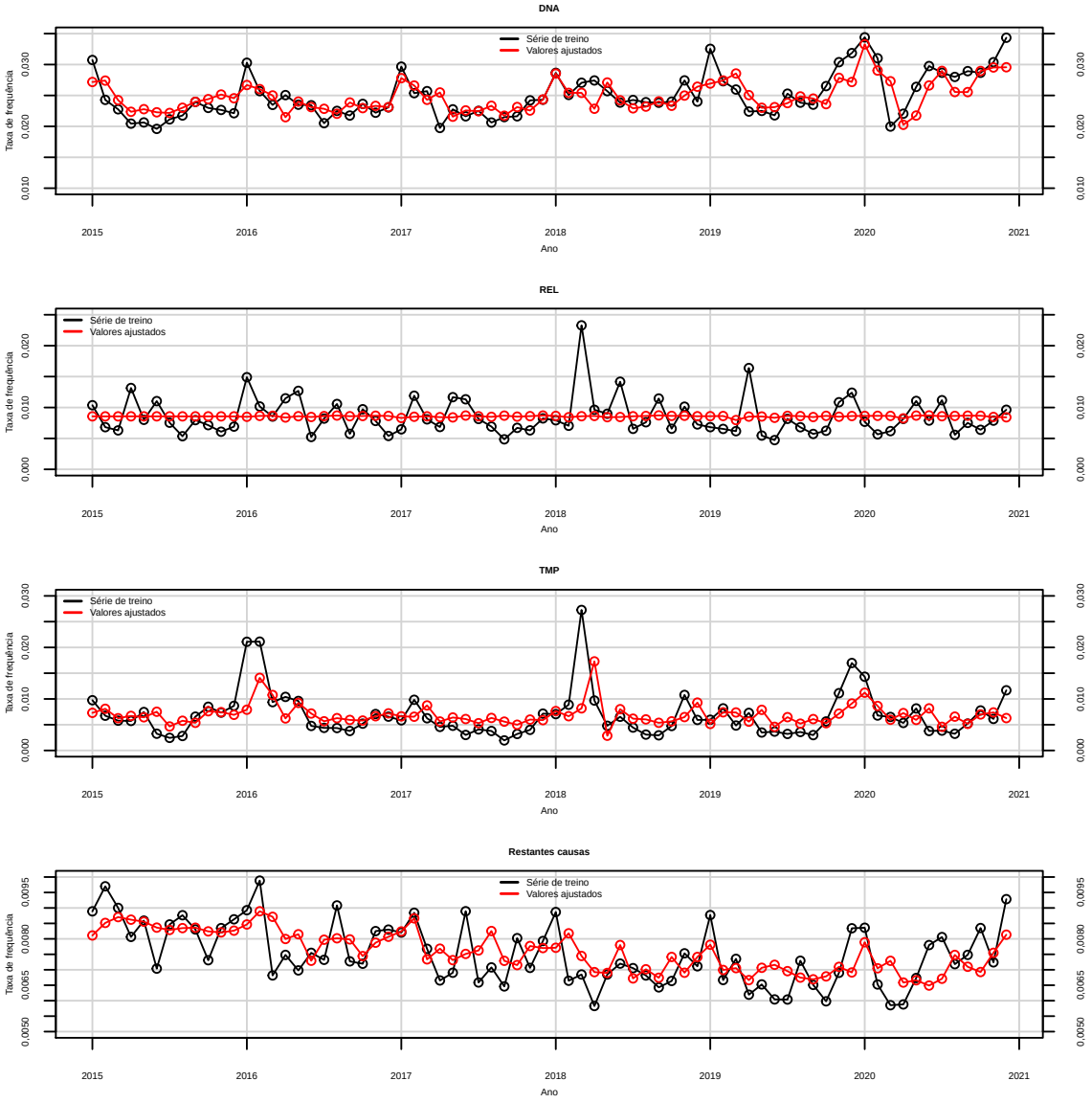


Figura D.5: Ajuste dos modelos SARIMA das séries mensais de taxas de frequência marginais.

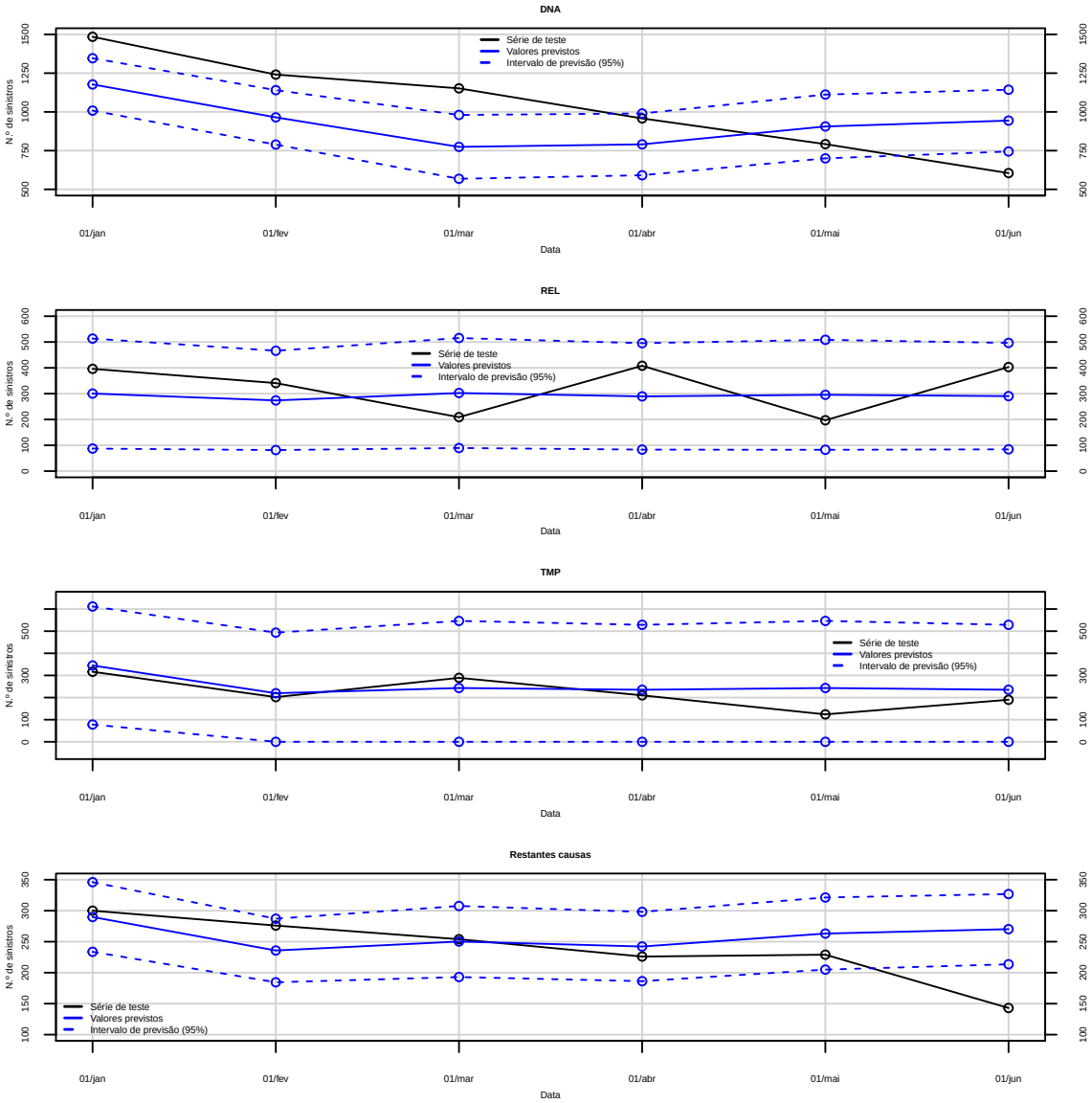


Figura D.6: Previsões pontuais e intervalares (a 95% de confiança) dos modelos SARIMA ajustados às séries mensais de taxas de frequência marginais.

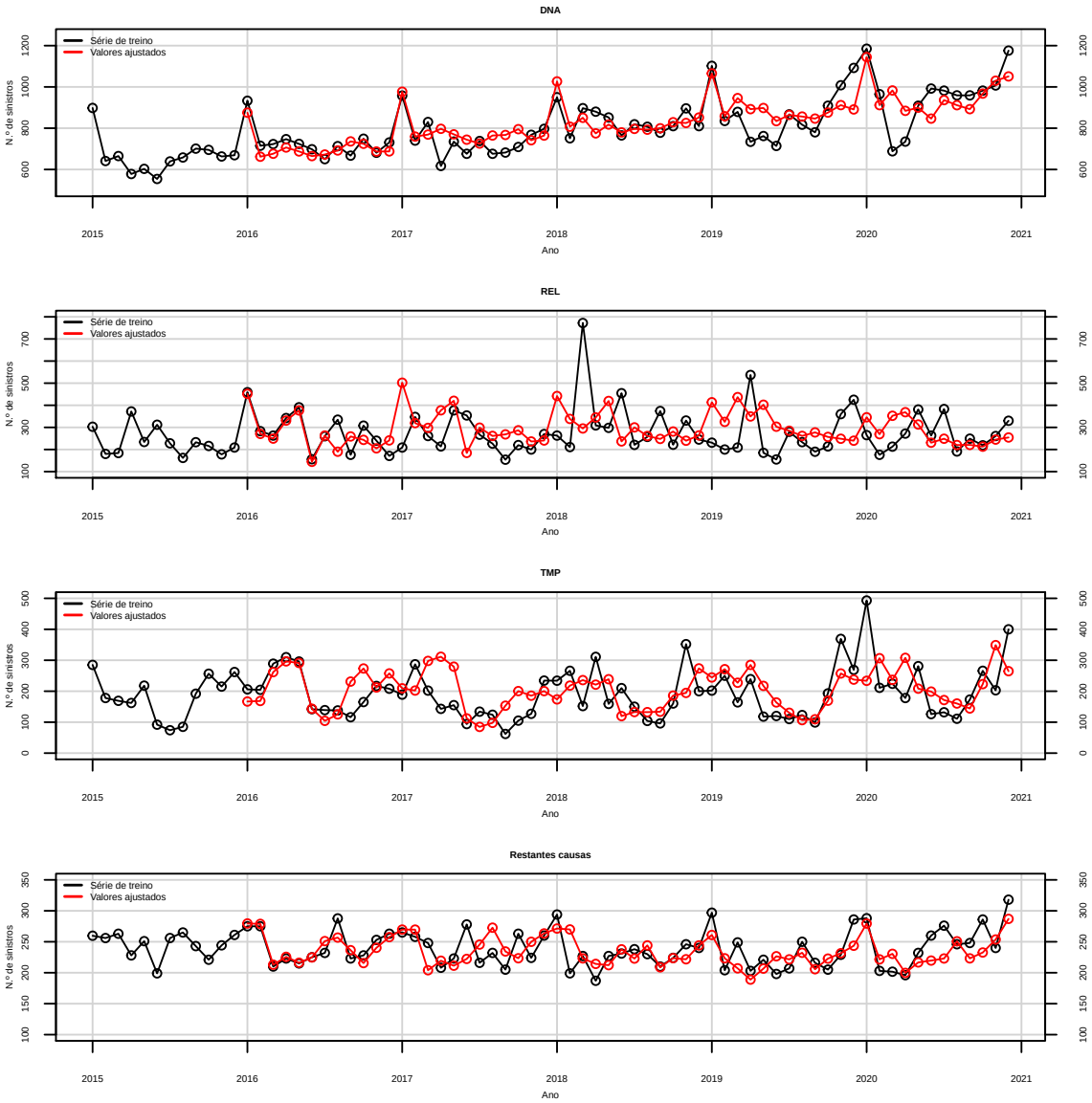


Figura D.7: Ajuste dos modelos Holt-Winters às séries mensais de contagens marginais.

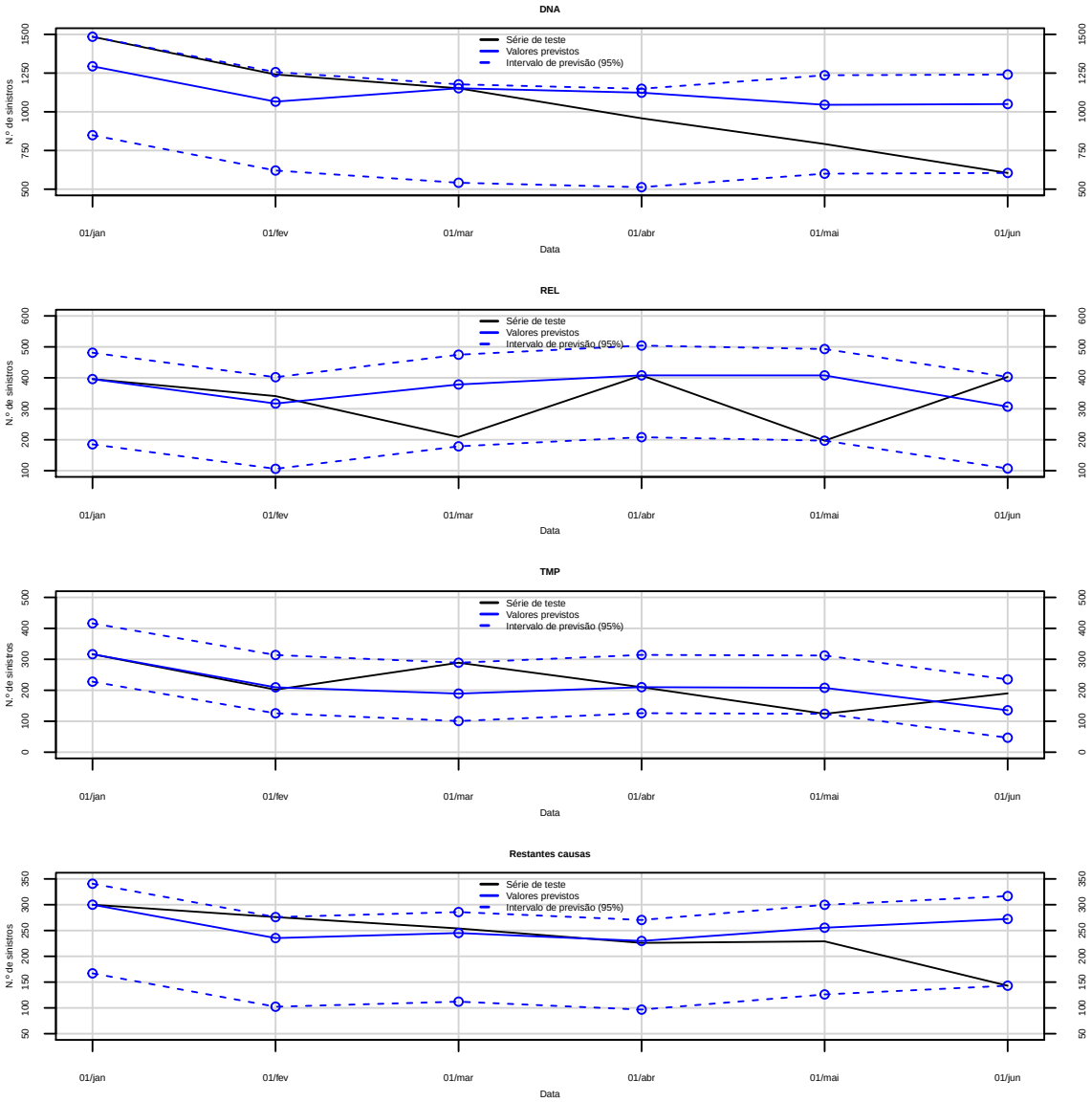


Figura D.8: Previsões pontuais e intervalares (a 95% de confiança) dos modelos Holt-Winters ajustados às séries mensais de contagens marginais.

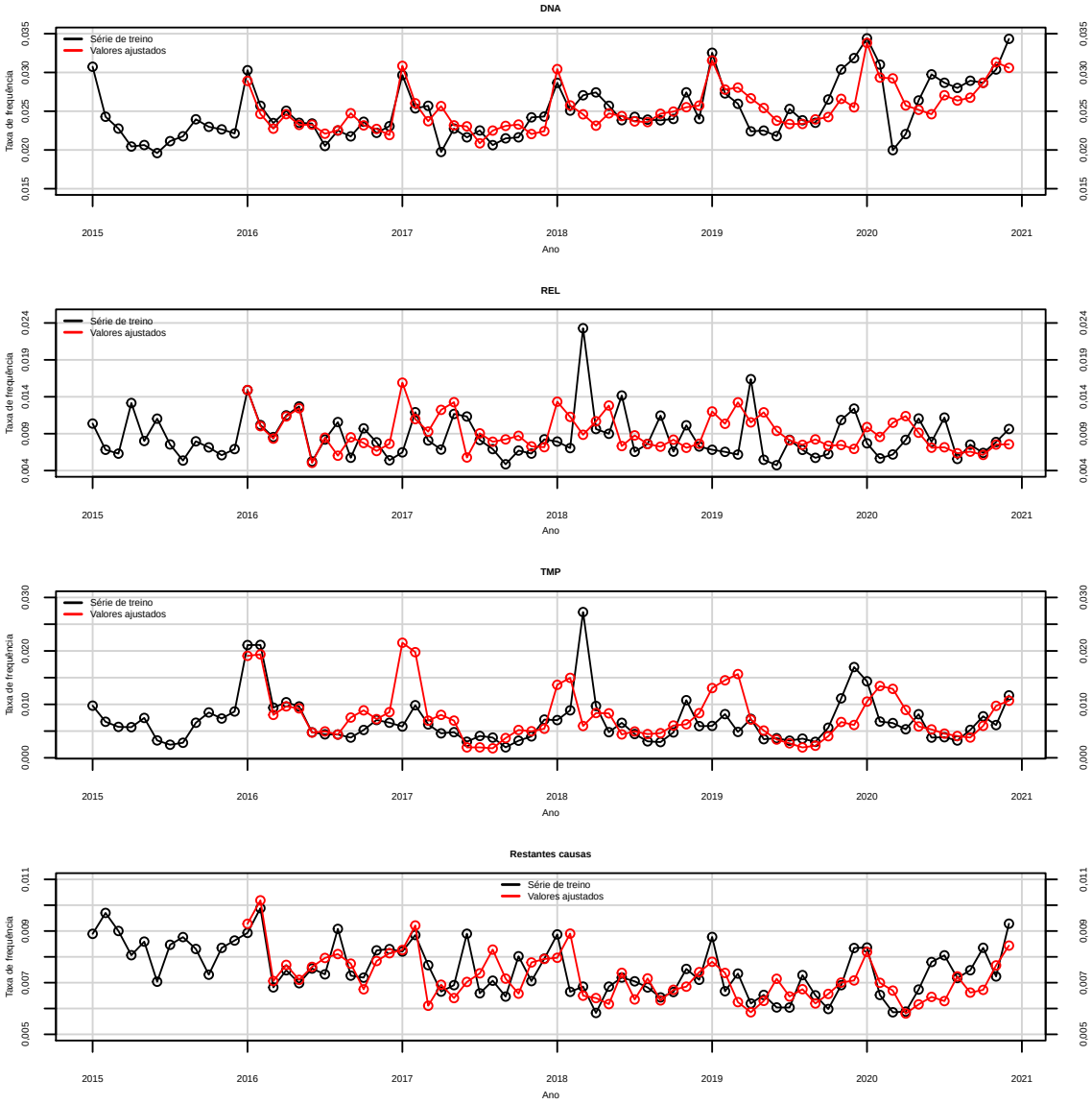


Figura D.9: Ajuste dos modelos Holt-Winters às séries mensais de taxas de frequência marginais.

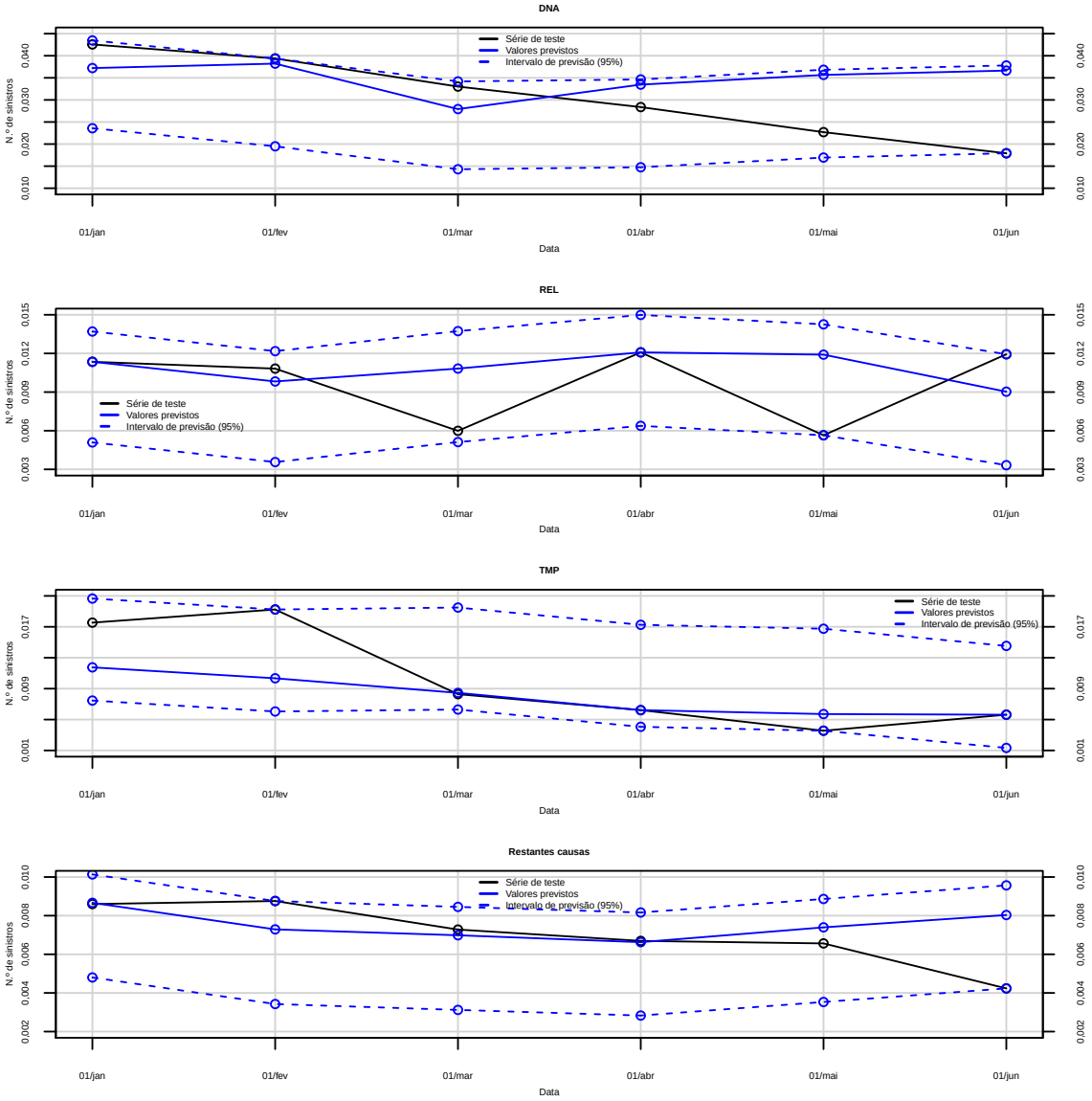


Figura D.10: Previsões pontuais e intervalares (a 95% de confiança) dos modelos Holt-Winters ajustados às séries mensais de taxas de frequência marginais.