



**Métodos de previsão de séries temporais
— uma aplicação a dados do segmento do Retalho**

Susana Machado Rodrigues Lima

UMinho | 2018



Universidade do Minho

Escola de Ciências

Susana Machado Rodrigues Lima

**Métodos de previsão de séries temporais
— uma aplicação a dados do segmento do
Retalho**

outubro de 2018



Universidade do Minho

Escola de Ciências

Susana Machado Rodrigues Lima

**Métodos de previsão de séries temporais
— uma aplicação a dados do segmento do
Retailho**

Tese de Mestrado
Mestrado em Estatística

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

outubro de 2018

DECLARAÇÃO

Nome: Susana Machado Rodrigues Lima

Endereço eletrónico: susanamrlima@gmail.com

Telefone: 933360490

Número do Cartão de Cidadão: 14857098

Título da dissertação: Métodos de previsão de séries temporais – uma aplicação a dados do segmento do Retalho

Orientadores: Professora Doutora Arminda Manuela Andrade Pereira Gonçalves

Professor Doutor Marco André da Silva Costa

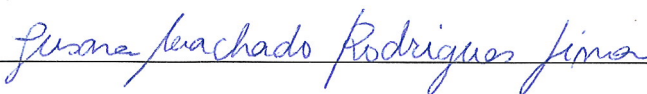
Ano de conclusão: 2018

Designação do Mestrado: Mestrado em Estatística

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA DISSERTAÇÃO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.

Universidade do Minho, 31 de outubro de 2018.

Assinatura:

 _____

Agradecimentos

“Quem caminha sozinho pode até chegar mais rápido, mas aquele que vai acompanhado, com certeza vai mais longe.”
(Clarice Lispector, s.d.)

Obrigada à Professora Doutora Arminda Manuela Gonçalves e ao Professor Doutor Marco Costa, pela generosa partilha de conhecimentos, pela constante disponibilidade, pelas palavras de incentivo e, essencialmente, pela amizade. O meu muito obrigada por me terem acompanhado neste percurso.

Obrigada à minha família, por ser o meu porto de abrigo e me apoiar em todas as minhas decisões. Obrigada por me ajudarem a crescer e me incentivarem todos os dias a ser uma pessoa melhor. A minha vida não estaria completa sem vocês.

Obrigada ao Hugo, meu companheiro de aventuras, por nunca me ter deixado desistir durante esta longa caminhada. Obrigada por me fazeres sempre ver o lado bom e, principalmente, por estares sempre do meu lado.

Obrigada às minhas colegas e amigas de Mestrado, por me terem acompanhado nestes dois anos e os terem tornado ainda mais especiais.

Obrigada aos meus amigos, por, longe ou perto, estarem sempre a torcer por mim. Vocês estão sempre presentes no meu coração.

E por fim, obrigada a todos aqueles que, durante todos estes anos, passaram pela minha vida, por, de alguma forma, terem contribuído para o meu crescimento pessoal e profissional.

Resumo

Os métodos de previsão constituem uma ferramenta fundamental nos processos de tomada de decisão em muitas áreas, como em Economia, Gestão e Finanças ou em Ambiente.

Na área da Economia, nomeadamente no segmento do retalho, a previsão de vendas suporta grande parte das decisões de planeamento estratégico de qualquer empresa e deve ser o mais precisa possível para garantir a rentabilidade da empresa.

Neste trabalho, é estudado um conjunto de séries temporais do segmento do retalho, relativas à variável índice de volume de negócios (TOVT), recolhidas nas bases de dados do Eurostat: séries temporais de Portugal, Alemanha, Espanha, França, Itália, Países Baixos e Reino Unido. Como muitas outras séries económicas, as vendas no segmento do retalho apresentam fortes tendências e padrões sazonais. A escolha do melhor processo para modelar e prever esses padrões tem sido um problema recorrente na análise de séries temporais.

O principal objetivo deste trabalho consiste em avaliar a precisão de vários métodos de previsão, na área da modelação de séries temporais, aplicados a dados do segmento do retalho, de forma a identificar as metodologias mais adequadas. Desta forma, é proposta a comparação da precisão entre os modelos ARMA e respetivas extensões, os modelos de decomposição clássica associados a modelos de regressão linear múltipla e os métodos de alisamento exponencial (Holt-Winters). A escolha destes métodos deve-se à sua capacidade de modelar tendências e flutuações sazonais presentes em dados económicos, particularmente em dados do segmento do retalho.

Para avaliar a capacidade preditiva das três metodologias adotadas, são utilizadas diversas medidas de avaliação, nomeadamente o EQM, o REQM, o EPAM, o EEAM e a estatística U de Theil.

Os resultados deste estudo comparativo sugerem que os modelos de regressão linear múltipla não são os mais adequados para prever índices de vendas no retalho, enquanto os modelos SARIMA são identificados como os mais precisos. Apesar de não serem considerados como os mais apropriados, os modelos de Holt-Winters também constituem uma alternativa viável.

Palavras-chave: Retalho; Séries temporais; Previsão; SARIMA; Regressão linear múltipla; Holt-Winters; Precisão.

Abstract

Forecasting methods are a key tool in decision-making processes in many areas, such as economics, management and finance, or environment.

In the area of economics, particularly in the retail segment, sales forecasting supports most of the strategic planning decisions of any retail business, and must be as accurate as possible to ensure the company's profitability.

In this work we study a set of retail time series in the retail business, related to the variable index of turnover (TOVT), which was collected in the Eurostat retail databases: time series of Portugal, Germany, Spain, France, Italy, the Netherlands and United Kingdom. Like many other economic time series, retail sales present strong trends and seasonal patterns. How to best model and forecast these patterns has been a long-standing issue in time series analysis.

The main propose of this work is to evaluate the accuracy of various forecasting methods, in the area of time series modelling, applied to data from the retail segment in order to identify the most appropriate methodologies. Thus, it is proposed to compare the accuracy of the ARMA models and their extensions, the classical decomposition time series associated with multiple linear regression models, and the exponential smoothing methods (Holt-Winters). These methods are chosen because of their ability to model trend and seasonal fluctuations present in economic data, particularly in retail sales data.

In order to evaluate the predictive capacity of the three methodologies adopted, several evaluation measures are used, namely MSE, RMSE, MAPE, MASE and U-Theil statistic.

The results of this comparative study suggest that multiple linear regression models are not the most appropriate to forecast retail sales indexes, while SARIMA models are identified as the most accurate ones. Although not considered the most appropriate, Holt-Winters models are also a viable alternative.

Keywords: Retail sales; Time series models; Forecasting; SARIMA; Multiple linear regression; Holt-Winters; Forecast accuracy.

Conteúdo

1	Introdução	1
1.1	Dados e motivação	1
1.2	Objetivos e estrutura do trabalho	2
2	Revisão de literatura	5
3	Séries temporais	11
3.1	Conceito de série temporal	11
3.1.1	Objetivos da análise de uma série temporal	12
3.1.2	Componentes de uma série temporal	13
3.2	Processos estocásticos	15
3.2.1	Processos estocásticos estacionários	16
3.2.2	Processos estocásticos não estacionários	22
4	Modelos e métodos de previsão	33
4.1	Modelos SARIMA e a metodologia Box-Jenkins	34
4.1.1	Modelos de processos estacionários	35
4.1.2	Modelos de processos não estacionários	42
4.1.3	Metodologia Box-Jenkins	45
4.1.4	Análise dos resíduos	49
4.1.5	Seleção de modelos	51
4.1.6	Previsão pontual e intervalar	53

4.2	Métodos de alisamento exponencial	54
4.2.1	Método de alisamento exponencial simples	57
4.2.2	Método linear de Holt	58
4.2.3	Método de Holt-Winters	59
4.2.4	Outros métodos de alisamento exponencial	61
4.2.5	Intervalos de previsão	62
4.3	Modelos de regressão linear	63
4.3.1	Conceitos fundamentais	64
4.3.2	Modelos de regressão para a tendência	71
4.3.3	Séries temporais interrompidas	71
4.3.4	Modelos de regressão para a sazonalidade	73
4.3.5	Modelos de regressão com erros correlacionados	74
4.4	Medidas de avaliação	76
4.4.1	Medidas de avaliação utilizadas	79
5	Aplicação dos métodos de previsão a dados do segmento do retalho	83
5.1	Apresentação da base de dados	88
5.2	Análise descritiva	89
5.3	Aplicação dos métodos de previsão	94
5.3.1	Caso I: Comércio a retalho em Portugal	98
5.3.2	Caso II: <i>E-commerce</i> em Portugal	118
5.3.3	Análise global	134
5.4	Comparação dos métodos de previsão	139
6	Conclusão	145
6.1	Sugestões para trabalho futuro	146
A	Representação gráfica (logaritmo)	157

B	Medidas descritivas dos dados	159
C	Aplicação dos métodos de previsão	163
D	Funções no ambiente R	187

Lista de Figuras

3.1	Simulação de um ruído branco e respetivas FAC e FACP empíricas. . .	21
3.2	Duas realizações de séries temporais não estacionárias.	22
3.3	Simulação de um modelo de passeio aleatório e as suas diferenças de 1. ^a ordem.	26
3.4	Simulação de um modelo de passeio aleatório com <i>drift</i> e as suas diferenças de 1. ^a ordem.	27
4.1	Simulação de um processo autorregressivo de equação $(1 - 0,6B + 0,3B^2)Y_t = \epsilon_t$, e respetivas FAC e FACP empíricas.	37
4.2	Simulação de um processo de médias móveis de equação $Y_t = (1 - 0,7B + 0,1B^2)\epsilon_t$, e respetivas FAC e FACP empíricas.	38
4.3	Simulação de um processo autorregressivo e de médias móveis, ARMA(2,2), de equação $(1 + 0,5B + 0,3B^2)Y_t = (1 - 0,7B + 0,1B^2)\epsilon_t$, e respetivas FAC e FACP empíricas.	39
4.4	Simulação de um processo autorregressivo e de médias móveis sazonal, SARMA(1,1) ₁₂ , de equação $(1 - 0,65B^{12})Y_t = (1 - 0,25B^{12})\epsilon_t$, e respetivas FAC e FACP empíricas.	41
4.5	Simulação de um processo multiplicativo, SARMA(1,0)(1,0) ₁₂ , de equação $(1 - 0,7B)(1 + 0,25B^{12})Y_t = \epsilon_t$, e respetivas FAC e FACP empíricas.	42

4.6	Simulação de um processo autorregressivo e de médias móveis integrado, ARIMA(2, 1, 1), de equação $(1 + 0,2B + 0,3B^2)(1 - B)Y_t = (1 + 0,7B)\epsilon_t$, e respectivas FAC e FACP empíricas.	43
4.7	Simulação de um processo autorregressivo e de médias móveis integrado sazonal, SARIMA(2, 1, 1)(1, 1, 1) ₁₂ , de equação $(1 - 1,2B + 0,8B^2)(1 - 0,3B^{12})(1 - B)(1 - B^{12})Y_t = (1 + 0,4B)(1 + 0,7B^{12})\epsilon_t$, e respectivas FAC e FACP empíricas.	45
4.8	Esquema da metodologia Box-Jenkins.	46
5.1	Distribuição do volume de negócios do comércio a retalho pelos 28 Estados-Membros da União Europeia (em biliões de euros), no ano de 2016.	86
5.2	Representação gráfica das séries do comércio a retalho, com exceção de veículos automóveis e motociclos, para os sete países em estudo, no período observado.	90
5.3	Representação gráfica das séries da venda a retalho por correspondência ou via <i>internet</i> , para os sete países em estudo, no período observado.	92
5.4	Diagramas em caixa de bigodes para a série do comércio a retalho em Portugal, dividida pelos doze meses do ano civil, no período observado.	93
5.5	Diagramas em caixa de bigodes para a série da venda a retalho por correspondência ou via <i>internet</i> em Portugal, dividida pelos doze meses do ano civil, no período observado.	94
5.6	Série dos logaritmos do comércio a retalho em Portugal, após diferenciação de 1. ^a ordem ($d = 1$), e respectivas FAC e FACP estimadas.	99
5.7	Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após diferenciação de 1. ^a ordem e ajustamento da parte sazonal, e respectivas FAC e FACP estimadas.	100

5.8	Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.	102
5.9	Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo SARIMA, sobrepostas à série do comércio a retalho em Portugal.	103
5.10	Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de Holt-Winters, sobrepostas à série do comércio a retalho em Portugal.	106
5.11	Série dos logaritmos do comércio a retalho em Portugal e respetiva componente tendência, obtida por médias móveis (usando o comando <i>decompose</i> no ambiente R).	107
5.12	Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após ajustamento de uma tendência linear, e respetivas FAC e FACP estimadas.	108
5.13	FAC e FACP obtidas para as séries dos resíduos, resultantes do ajustamento dos dois modelos de regressão (completos) à série dos logaritmos do comércio a retalho em Portugal.	110
5.14	Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após ajustamento do modelo de regressão, e respetivo histograma, FAC e FACP estimadas.	116
5.15	Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de regressão linear, sobrepostas à série do comércio a retalho em Portugal.	117

5.16	Série dos logaritmos do <i>e-commerce</i> em Portugal, após diferenciação de 1. ^a ordem ($d = 1$), e respectivas FAC e FACP estimadas.	119
5.17	Série dos resíduos para a série dos logaritmos do <i>e-commerce</i> em Portugal, após diferenciação de 1. ^a ordem e ajustamento da parte sazonal, e respectivas FAC e FACP estimadas.	120
5.18	Série dos resíduos para a série dos logaritmos do <i>e-commerce</i> em Portugal, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.	122
5.19	Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo SARIMA, sobrepostas à série do <i>e-commerce</i> em Portugal.	123
5.20	Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de Holt-Winters, sobrepostas à série do <i>e-commerce</i> em Portugal.	125
5.21	Série dos logaritmos das vendas por correspondência ou via <i>internet</i> em Portugal e respetiva componente tendência, obtida por médias móveis (usando o comando <i>decompose</i> no ambiente R).	126
5.22	FAC e FACP obtidas para as séries dos resíduos, resultantes do ajustamento dos três modelos de regressão (completos) à série dos logaritmos do <i>e-commerce</i> em Portugal.	127
5.23	Série dos resíduos para a série dos logaritmos do <i>e-commerce</i> em Portugal, após ajustamento do modelo de regressão, e respetivo histograma, FAC e FACP estimadas.	132

5.24	Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de regressão linear, sobrepostas à série do <i>e-commerce</i> em Portugal.	133
A.1	Representação gráfica das séries dos logaritmos do comércio a retalho, com exceção de veículos automóveis e motociclos, para os sete países em estudo, no período observado.	157
A.2	Representação gráfica das séries dos logaritmos da venda a retalho por correspondência ou via <i>internet</i> , para os sete países em estudo, no período observado.	158

Lista de Tabelas

3.1	Transformações usuais de Box-Cox.	24
4.1	Padrões teóricos das FAC e FACP dos modelos SARMA(p, q)(P, Q) $_s$	48
4.2	Classificação dos métodos de alisamento exponencial proposta por Pegels (1969).	55
4.3	Classificação atual dos métodos de alisamento exponencial.	55
4.4	Notação para o alisamento exponencial.	56
4.5	Equações do método de Holt-Winters.	60
4.6	Inicialização do método de Holt-Winters.	60
4.7	Equações do método de Holt-Winters com tendência aditiva amortecida.	62
4.8	Tabela de análise de variância (ANOVA).	68
5.1	Taxa de crescimento real do PIB (%) na União Europeia e em alguns Estados-Membros, nos anos de 2007 a 2017.	84
5.2	Volume de negócios do comércio a retalho (em biliões de euros) na União Europeia e em alguns Estados-Membros, nos anos de 2007 a 2016.	87
5.3	Volume de negócios do <i>e-commerce</i> no comércio a retalho (% de variação anual) na União Europeia e em alguns Estados-Membros, nos anos de 2009 a 2016.	88

5.4	Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos do comércio a retalho em Portugal.	100
5.5	Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos do comércio a retalho em Portugal.	101
5.6	Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos do comércio a retalho em Portugal.	101
5.7	Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método multiplicativo de Holt-Winters à série do comércio a retalho em Portugal.	104
5.8	Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método aditivo de Holt-Winters à série do comércio a retalho em Portugal.	104
5.9	Medidas de avaliação dos modelos de Holt-Winters para a série do comércio a retalho em Portugal.	104
5.10	Erro quadrático médio a h -passos para o modelo de Holt-Winters escolhido para a série do comércio a retalho em Portugal.	105
5.11	Teste à normalidade dos erros de previsão a h -passos para o modelo de Holt-Winters escolhido para a série do comércio a retalho em Portugal.	106
5.12	Resultados do ajustamento do modelo completo, com um evento, à série dos logaritmos do comércio a retalho em Portugal.	111
5.13	Resultados do ajustamento do modelo completo, com dois eventos, à série dos logaritmos do comércio a retalho em Portugal.	112

5.14	Resultados do ajustamento do modelo final reduzido, com um evento, à série dos logaritmos do comércio a retalho em Portugal.	114
5.15	Resultados do ajustamento do modelo final reduzido, com dois eventos, à série dos logaritmos do comércio a retalho em Portugal.	115
5.16	Medidas de avaliação dos modelos de regressão para a série do comércio a retalho em Portugal.	116
5.17	Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos do <i>e-commerce</i> em Portugal.	120
5.18	Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos do <i>e-commerce</i> em Portugal.	121
5.19	Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos do <i>e-commerce</i> em Portugal.	121
5.20	Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método multiplicativo de Holt-Winters à série do <i>e-commerce</i> em Portugal.	123
5.21	Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método aditivo de Holt-Winters à série do <i>e-commerce</i> em Portugal.	124
5.22	Medidas de avaliação dos modelos de Holt-Winters para a série do <i>e-commerce</i> em Portugal.	124
5.23	Erro quadrático médio a h -passos para o modelo de Holt-Winters escolhido para a série do <i>e-commerce</i> em Portugal.	124
5.24	Resultados da aplicação dos testes de estacionariedade às séries dos resíduos dos três modelos de regressão ajustados.	126

5.25	Resultados do ajustamento do modelo final reduzido, com tendência linear, à série dos logaritmos do <i>e-commerce</i> em Portugal.	129
5.26	Resultados do ajustamento do modelo final reduzido, com um evento, à série dos logaritmos do <i>e-commerce</i> em Portugal.	130
5.27	Resultados do ajustamento do modelo final reduzido, com dois eventos, à série dos logaritmos do <i>e-commerce</i> em Portugal.	131
5.28	Medidas de avaliação dos modelos de regressão para a série do <i>e-commerce</i> em Portugal.	131
5.29	Estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters às séries em estudo.	135
5.30	Estimativas dos parâmetros dos modelos SARIMA ajustados às séries dos logaritmos do comércio a retalho.	136
5.31	Estimativas dos parâmetros dos modelos SARIMA ajustados às séries dos logaritmos do <i>e-commerce</i>	136
5.32	Estrutura dos modelos de regressão ajustados às séries dos logaritmos do comércio a retalho.	138
5.33	Estrutura dos modelos de regressão ajustados às séries dos logaritmos do <i>e-commerce</i>	139
5.34	Medidas de avaliação calculadas para as séries do comércio a retalho, no período de treino (janeiro de 2000 a dezembro de 2016) e no período de teste (janeiro de 2017 a fevereiro de 2018), com base nos resultados obtidos na aplicação dos três métodos de previsão em estudo.	141
5.35	Medidas de avaliação calculadas para as séries do <i>e-commerce</i> , no período de treino (janeiro de 2000 a dezembro de 2016) e no período de teste (janeiro de 2017 a fevereiro de 2018), com base nos resultados obtidos na aplicação dos três métodos de previsão em estudo.	142
5.36	Taxas de cobertura (%) efetiva dos intervalos de previsão a 95%.	144

B.1	Medidas descritivas dos índices de volume de negócios, registados no comércio a retalho (com exceção de veículos automóveis e motociclos).	159
B.2	Medidas descritivas dos índices de volume de negócios, registados na venda a retalho por correspondência ou via <i>internet</i> .	160
B.3	Médias anuais e médias de meses homólogos dos índices de volume de negócios, registados no comércio a retalho, com exceção de veículos automóveis e motociclos.	161
B.4	Médias anuais e médias de meses homólogos dos índices de volume de negócios, registados na venda a retalho por correspondência ou via <i>internet</i> .	162
C.1	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho na Alemanha.	163
C.2	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em Espanha.	163
C.3	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em França.	164
C.4	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em Itália.	164
C.5	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho nos Países Baixos.	164
C.6	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em Portugal.	164
C.7	Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho no Reino Unido.	165
C.8	Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho na Alemanha.	165

C.9 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em Espanha.	165
C.10 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em França.	166
C.11 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em Itália.	166
C.12 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho nos Países Baixos.	166
C.13 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em Portugal.	167
C.14 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho no Reino Unido.	167
C.15 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho na Alemanha.	168
C.16 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em Espanha.	169
C.17 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em França.	170
C.18 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em Itália.	171

C.19 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho nos Países Baixos.	172
C.20 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em Portugal.	173
C.21 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho no Reino Unido.	174
C.22 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> na Alemanha.	175
C.23 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> em Espanha.	175
C.24 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> em França.	175
C.25 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> em Itália.	175
C.26 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> nos Países Baixos.	176
C.27 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> em Portugal.	176
C.28 Resultados do ajustamento do modelo SARIMA à série dos logaritmos do <i>e-commerce</i> no Reino Unido.	176
C.29 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> na Alemanha.	176
C.30 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> em Espanha.	177

C.31 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> em França.	177
C.32 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> em Itália.	177
C.33 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> nos Países Baixos. . .	178
C.34 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> em Portugal.	178
C.35 Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do <i>e-commerce</i> no Reino Unido. . . .	178
C.36 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> na Alemanha.	179
C.37 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> em Espanha.	180
C.38 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> em França.	181
C.39 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> em Itália.	182
C.40 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> nos Países Baixos.	183
C.41 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> em Portugal.	184

C.42 Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do <i>e-commerce</i> no Reino Unido.	185
D.1 Principais funções utilizadas no ambiente R.	189

Lista de abreviaturas

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

AIC – *Akaike Information Criterion* (em português, Critério de Informação de Akaike)

ANOVA – *Analysis of Variance* (em português, Análise de Variância)

AR – *Autoregressive* (em português, Autorregressivo)

ARCH – *Autoregressive Conditional Heteroscedastic* (em português, Autorregressivo Condicionalmente Heterocedástico)

ARFIMA – *Autoregressive Fractionally Integrated Moving Average* (em português, Autorregressivo e de Médias Móveis Fracionalmente Integrado)

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

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

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

DE – Alemanha

DF – Dickey-Fuller

EAM – Erro Absoluto Médio

EAMN – Erro Absoluto Médio *Naïve*

EAMNs – Erro Absoluto Médio *Naïve* Sazonal

EEAM – Erro Escalado Absoluto Médio

EPAM – Erro Percentual Absoluto Médio

EPAMS – Erro Percentual Absoluto Médio Simétrico

EQM – Erro Quadrático Médio

ES – Espanha

EU-27 – União Europeia (constituída por 27 países, antes da adesão da Croácia)

EU-28 – União Europeia (constituída por 28 países)

Eurostat – *European Statistical Office* (em português, Gabinete de Estatísticas da União Europeia)

FAC – Função de Autocorrelação

FACP – Função de Autocorrelação Parcial

FR – França

GDP – *Gross Domestic Product* (equivalente ao PIB português)

INE – Instituto Nacional de Estatística

IT – Itália

IVA – Imposto sobre o Valor Acrescentado

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 Scaled Error* (em português, Erro Escalado Absoluto Médio)

MQE – Média de Quadrados Residual

MQR – Média de Quadrados da Regressão

MQT – Média de Quadrados Total

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

NA – Não Aplicável

NL – Países Baixos

PIB – Produto Interno Bruto

PORDATA – Base de Dados de Portugal Contemporâneo

PP – Phillips-Perron

PT – Portugal

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)

SMA – *Seasonal Moving Average* (em português, Médias Móveis Sazonal)

SQE – Soma de Quadrados Residual

SQR – Soma de Quadrados da Regressão

SQT – Soma de Quadrados Total

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

TOVT – *Index of Turnover - Total* (em português, Índice de Volume de Negócios)

UK – Reino Unido

VAR – *Vector Autoregressive* (em português, Autorregressivo Vetorial)

VIF – *Variance Inflation Factor* (em português, Fator de Inflação da Variância)

Capítulo 1

Introdução

Os métodos de previsão constituem uma ferramenta fundamental nos processos de tomada de decisão em diversas áreas, como em Economia, Gestão e Finanças, ou em Ambiente, entre outras.

Existem vários métodos de previsão suportados por diferentes metodologias estatísticas, desde os mais complexos, que envolvem diferentes parâmetros, aos métodos mais simples e de fácil compreensão. Estes métodos podem ser baseados em modelos de séries temporais designados por modelos de Box-Jenkins, modelos de decomposição clássica associados a modelos de regressão linear múltipla e métodos de alisamento exponencial, entre outros. Uma descrição mais exaustiva dos vários métodos de previsão pode ser encontrada em De Gooijer & Hyndman (2006).

No entanto, com a existência de múltiplas abordagens, a opção por apenas uma delas torna-se complicada. As medidas de avaliação, baseadas na análise da precisão das previsões, são, em muitos dos casos, o principal critério de seleção de um método de previsão. Através destas medidas é possível avaliar a eficácia de um modelo ou método de previsão, mostrando o quanto ele é capaz de descrever a série temporal em estudo e de obter previsões que venham a revelar-se o mais precisas possível.

1.1 Dados e motivação

O comércio a retalho é um setor dinâmico e complexo, cuja estrutura reflete as características culturais da sociedade em que se insere e, por isso, de forma a avaliar o desenvolvimento de uma economia nacional, revela-se fundamental o estudo do mesmo. De facto, de acordo com o Eurostat (*European Statistical Office*), além de ter representado quase um terço do volume de negócios do setor do comércio em 2015, o comércio a retalho deteve 58,4% das empresas comerciais e empregou 8,7%

da população ativa da União Europeia, o que corresponde a aproximadamente 18,8 milhões de postos de trabalho. No mesmo ano, em Portugal, o segmento do retalho no consumo privado representou 34,4%, pelo que se considera que o estudo das variáveis económicas associadas a esta área é imprescindível e bastante útil, tanto para caracterizar o passado recente como antecipar tendências.

A introdução da *internet* no quotidiano da população europeia e o seu constante desenvolvimento deram origem a novas práticas de consumo, em particular, ao comércio eletrónico (*e-commerce*). De facto, embora a maioria das transações ainda ocorra *offline*, o retalho *online* ganhou mais e mais destaque nos últimos anos, sendo esta evolução visível por toda a União Europeia.

Neste estudo são, assim, analisados dados relativos a variáveis (índices) económicas associadas ao segmento do comércio a retalho, mais especificamente, ao comércio a retalho generalizado e ao *e-commerce*. Os dados dizem respeito a sete países distintos – Portugal (PT), Espanha (ES), França (FR), Alemanha (DE), Itália (IT), Países Baixos (NL) e Reino Unido (UK) –, que, à exceção de Portugal, representam os países da Europa com maiores relações comerciais com Portugal. A variável utilizada corresponde ao índice de volume de negócios (*index of turnover - total*, TOVT), cujo objetivo é mostrar a evolução do mercado de bens e serviços, recolhida mensalmente no período de janeiro de 2000 a fevereiro de 2018, nos sete países referidos e para os dois tipos de comércio a retalho.

De referir ainda que na aplicação das metodologias aos dados se recorre ao ambiente R (R Core Team, 2017), sendo, essencialmente, utilizadas as *packages stats*, *forecast* e *Metrics*. Além disso, são também implementadas algumas funções que permitem a construção de intervalos de previsão para o método de Holt-Winters.

1.2 Objetivos e estrutura do trabalho

O objetivo principal deste trabalho é identificar os modelos/métodos de previsão mais adequados às variáveis recolhidas, procedendo-se, para isso, ao seu ajustamento e comparação em termos da sua capacidade explicativa e preditiva.

Com esse propósito, no Capítulo 2 é feita uma breve revisão da literatura, relacionada com os métodos de previsão de séries temporais. São também apresentados alguns exemplos de aplicação destes métodos em diferentes áreas de investigação, nomeadamente na área em estudo – o segmento do retalho – e, além disso, é abordado o problema da escolha entre os diversos métodos de previsão, sendo identificados alguns critérios possíveis de avaliação.

Nos Capítulos 3 e 4 descrevem-se os conteúdos teóricos relacionados com os métodos de previsão de séries temporais, mais concretamente, os conceitos fundamentais sobre séries temporais, os métodos que são aplicados aos dados e as medidas de avaliação utilizadas para selecionar o método mais adequado ao estudo realizado.

A aplicação das metodologias aos dados é apresentada no Capítulo 5. Inicialmente, é feita uma breve análise do setor do comércio, tanto a nível europeu como nacional, sendo dada uma maior relevância ao comércio a retalho e ao *e-commerce*. Depois de apresentados os dados em estudo, realiza-se uma análise exploratória dos mesmos, seguindo-se a aplicação das três metodologias selecionadas. Os casos nacionais são descritos com mais pormenor, enquanto que para os restantes é desenvolvida uma análise mais global dos resultados. Por fim, é realizado um estudo comparativo da capacidade preditiva dos diferentes métodos aplicados, através das medidas de avaliação apresentadas. As taxas de cobertura dos intervalos de previsão são também calculadas, representando um indicador adicional do desempenho das metodologias.

As principais conclusões do trabalho desenvolvido e sobre os resultados obtidos são descritas no Capítulo 6, assim como algumas sugestões de investigação para trabalho futuro.

Capítulo 1. Introdução

Capítulo 2

Revisão de literatura

Apesar de existirem, ainda durante o século XIX, algumas tentativas (determinísticas) de estudar séries temporais, é já no século XX que Yule (1927) estabelece a noção de estocasticidade em séries temporais, postulando que cada série pode ser considerada como uma realização de um processo estocástico (De Gooijer & Hyndman, 2006). Em 1970, baseados nos trabalhos de Yule (1926) e Wold (1938), Box & Jenkins (1970) desenvolvem uma abordagem prática e sistemática para a construção de modelos autorregressivos e de médias móveis integrados (*autorregressive integrated moving average*, ARIMA), que não é mais do que um processo de modelação iterativo em três fases: identificação, estimação e validação (Zhang, 2003; Tsay, 2000). Também durante o século XX, mais precisamente nos anos 50 e 60, surgem, com os trabalhos de Brown (1959, 1963), Holt (1957) e Winters (1960), os modelos de alisamento exponencial, usualmente designados por modelos de Holt-Winters.

No entanto, a linearidade assumida por estes modelos – tanto os ARIMA como os de Holt-Winters –, apesar de ser uma suposição útil e uma ferramenta poderosa em muitos casos, começa a mostrar-se, no final dos anos 70 e início dos anos 80, insuficiente em muitas aplicações reais (De Gooijer & Hyndman, 2006). Assim, de forma a ultrapassar a restrição de linearidade imposta pelos modelos lineares e a possibilitar a explicação de alguns padrões não lineares, começam a surgir outros modelos – não lineares –, tais como os modelos bilineares e os modelos autorregressivos com limiares (*threshold autoregressive*, TAR) (consultar Tong, 1990, para mais informação). Na área da Econometria, Engle (1982) desenvolve o modelo autorregressivo condicionalmente heterocedástico (*autoregressive conditional heterocedastic*, ARCH), que se mostra muito útil para lidar com a volatilidade, muito comum em séries económicas (Tsay, 2000). Mais recentemente, as redes neuronais artificiais têm sido sugeridas como uma alternativa para a previsão de séries temporais, uma vez que os modelos

anteriores são desenvolvidos para padrões não lineares específicos e nem sempre são capazes de modelar outros tipos de não linearidade (Zhang, 2003).

Quanto à existência de outros modelos, devem-se ressaltar aqueles que satisfazem a necessidade de considerar conjuntamente várias séries temporais relacionadas entre si, i.e., os modelos autorregressivos vetoriais (*vector autorregressive*, VAR). Além desses, surgem, também por necessidade, os modelos autorregressivos e de médias móveis fracionalmente integrados (*autoregressive fractionally integrated moving average*, ARFIMA), que permitem lidar com séries que possuam funções de autocorrelação empíricas com decaimentos mais lentos do que nos modelos ARIMA, ou seja, com séries de memória longa ou longa dependência (Tsay, 2000). Uma revisão mais aprofundada sobre os métodos de previsão de séries temporais desenvolvidos nos últimos 25 anos pode ser consultada em De Gooijer & Hyndman (2006).

Na literatura encontram-se vários exemplos de aplicação dos diferentes métodos de previsão de séries temporais. Por exemplo, na área da Economia, estes métodos são muito aplicados a variáveis macroeconômicas (e.g., Stock & Watson, 1999; Marcellino et al., 2003), em particular ao *gross domestic product*, GDP (e.g., Wang, 2016; Dritsaki, 2015), de forma a avaliar o crescimento econômico de um determinado país ou de uma determinada região. Ainda dentro desta área, encontram-se exemplos relacionados com o turismo (e.g., Song & Li, 2008; Chu, 2009), as taxas de câmbio (e.g., Sfetsos & Siriopoulos, 2005; Pacelli, 2012; Santos et al., 2007), o retalho (Alon et al., 2001; Chu & Zhang, 2003; Veiga et al., 2016; Ramos et al., 2015), etc. Outra área em que os métodos de previsão de séries temporais são amplamente utilizados é a das Ciências do Ambiente, muitas vezes para avaliar e prever aspectos relacionados com o aquecimento global (e.g., Ye et al., 2013; Bolin et al., 2014). Também na área da Medicina existem diversas aplicações, nomeadamente no que respeita a prever internamentos de emergência (e.g., Vile et al., 2016; Aboagye-Sarfo et al., 2015), no estudo de algumas doenças (e.g., Crabtree et al., 1990) ou até de causas de morte (e.g., Kim et al., 2004).

No que respeita a exemplos de estudos comparativos dentro do segmento do retalho, Alon et al. (2001) comparam um método não linear – o modelo de redes neuronais artificiais – com três métodos tradicionais de previsão – o modelo de Holt-Winters, o modelo ARIMA de Box-Jenkins e o modelo de regressão múltipla. Nesse estudo utilizam a série das vendas a retalho nos Estados Unidos da América (entre janeiro de 1978 e abril de 1995) e dividem-na em dois períodos, considerando que as condições econômicas diferem durante os mesmos. Os resultados indicam que, geralmente, o modelo de redes neuronais artificiais fornece melhores previsões do

que os métodos tradicionais, embora o modelo de Box-Jenkins continue a ser uma excelente alternativa. Quando as condições económicas são relativamente estáveis, o modelo de Holt-Winters também se mostra um método viável.

Noutro exemplo, Chu & Zhang (2003) comparam diversos métodos de previsão, nomeadamente, o modelo clássico de Box-Jenkins e modelos de regressão linear múltipla, que são abordagens lineares, e modelos de redes neuronais, que são metodologias não lineares. Tanto os modelos de regressão linear múltipla como os de redes neuronais incluem diferentes abordagens à modelação da sazonalidade, i.e., a componente sazonal é explicada ou através de variáveis indicatrizes (*dummy*) ou através de funções trigonométricas. No caso dos modelos de redes neuronais existem também as opções de construir o modelo com os dados originais, não existindo, assim, decomposição da série, ou com os dados sazonalmente ajustados (*seasonally adjusted*). Nesse estudo, os dados utilizados correspondem a vendas a retalho e foram recolhidos mensalmente entre janeiro de 1985 e dezembro de 1999. Os resultados sugerem que os métodos não lineares são os mais apropriados, sendo o modelo de redes neuronais com dados sazonalmente ajustados apontado como o mais preciso. Além disso, enquanto o uso de variáveis indicatrizes para a modelação da sazonalidade parece melhorar as previsões, o mesmo não acontece com as funções trigonométricas e, por isso, não é aconselhado o uso das mesmas para dados do segmento do retalho. Contrariamente ao que é feito noutros estudos, a série de confirmação (ou série de teste) não se resume apenas às últimas observações recolhidas (sejam elas os últimos dias, meses, anos, etc.). Usando uma metodologia de validação cruzada múltipla, são consideradas cinco séries de confirmação distintas, correspondendo cada uma delas a um dos 5 últimos anos de observações, i.e., são construídos modelos cuja série de modelação exclui, um a um, os anos de 1995 a 1999. Esta abordagem permite testar a robustez dos modelos, dando a perceber se os resultados obtidos podem ser replicados para outras situações ou se só se verificam para uma determinada amostra.

Também em Veiga et al. (2016) são comparados métodos lineares com métodos não lineares, mais especificamente o modelo ARIMA de Box-Jenkins e os modelos de Holt-Winters com o modelo de Takagi-Sugeno e o modelo de redes neuronais Wavelet. Para tal, usam dados mensais das vendas a retalho de produtos lácteos (recolhidos entre 2005 e 2013), que são agrupados em três grupos distintos consoante as suas características. De uma forma geral, concluem que os métodos de previsão utilizados acompanham o padrão sazonal existente nos dados, apesar de os não lineares o fazerem melhor. Em termos de precisão das previsões, também os métodos

não lineares se revelam como sendo os mais precisos, nomeadamente o modelo de redes neuronais Wavelet que mostra, nesse estudo, ser o mais apropriado para dados do segmento do retalho alimentar.

Já no segmento não alimentar, Ramos et al. (2015) comparam o modelo clássico de Box-Jenkins (ARIMA) com um modelo em espaço de estados, analisando, para tal, as vendas mensais (entre janeiro de 2007 e abril de 2012) de cinco categorias de calçado feminino da marca Foreva. Os resultados mostram que a precisão das previsões é bastante semelhante para os dois métodos aplicados.

De notar também que, nos últimos anos, os modelos de redes neuronais têm sido imensamente explorados enquanto método de previsão, inclusive no segmento do retalho. Encontram-se, na literatura, vários exemplos que comparam estes modelos com o modelo clássico de Box-Jenkins (e.g., Zhang & Qi, 2005; Kuvulmaz et al., 2005). Uma revisão mais pormenorizada sobre os modelos de redes neuronais pode ser vista em Zhang et al. (1998).

Em alguns casos, *“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”*, (Zhang, 2003). Pan (2013) e Aburto & Weber (2007) combinam o modelo de Box-Jenkins com o de redes neuronais, com aplicação às vendas a retalho e à gestão da cadeia de abastecimento, respetivamente. Alguma da bibliografia existente sobre modelos combinados encontra-se compilada em Clemen (1989).

Desta forma, e dada a existência de tantas abordagens distintas, existe uma questão que se coloca: “Com base em que critério se escolhe entre os diversos métodos de previsão?”. Yokum & Armstrong (1995) tentam responder a esta questão realizando dois diferentes estudos. No primeiro avaliam se os critérios diferem de acordo com a principal função do inquirido (e.g., investigador, professor, etc.), enquanto que, no segundo pretendem verificar se o critério de seleção varia ou não consoante a natureza da previsão (i.e., da quantidade de dados disponível, do número de previsões necessárias e/ou do método utilizado). Os resultados destes e de anteriores estudos revelam que o critério considerado mais importante é a precisão (*accuracy*), sendo essa a opinião dos especialistas, independentemente da sua função ou da natureza da previsão. Apesar disso, outros critérios, como a facilidade de utilização, a facilidade de interpretação, a facilidade de implementação, a flexibilidade, a poupança de custos, etc., devem também ser avaliados, pois é recomendada a utilização de vá-

rios critérios na seleção e avaliação dos métodos de previsão (Yokum & Armstrong, 1995). De facto, apesar de serem mencionados alguns outros, a preferência pelo critério da precisão é também notória nos estudos referidos anteriormente (Alon et al., 2001; Chu & Zhang, 2003; Ramos et al., 2015; Veiga et al., 2016), uma vez que esse é, quase que exclusivamente, o critério utilizado para comparar e escolher entre os diferentes métodos.

Também para avaliar a precisão das previsões existem vários critérios. No entanto, nem todos os critérios são adequados para todas as situações (Hyndman & Koehler, 2006; Hyndman, 2006). Por exemplo, medidas que dependam 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 úteis para comparar diferentes métodos aplicados ao mesmo conjunto de dados, porém não devem ser utilizadas para comparar dados com diferentes escalas. Para esses casos existem as medidas baseadas em erros percentuais (e.g., o erro percentual absoluto médio, EPAM), que são independentes da escala, mas têm a desvantagem de poderem ser infinitas ou indefinidas caso a série possua observações nulas. Outras alternativas para lidar com os problemas de escala podem ser as medidas de erros relativos ou as medidas de erros escalados.

Hyndman & Koehler (2006) mostram a inadequabilidade de várias medidas de precisão, utilizando três séries específicas que contêm zeros e/ou observações negativas e que, por isso, são representativas de muitas séries reais. Propõem a utilização de medidas de erros escalados para esses casos, nomeadamente o erro escalado absoluto médio (EEAM), que aparenta ser amplamente aplicável e sempre definido e finito, além de facilmente interpretável. No entanto, admitem a existência de situações em que as medidas tradicionais continuam a ser preferidas, dada a sua simplicidade. Por exemplo, se todas as séries estão na mesma escala, deve optar-se pelo EAM ou, se os dados são positivos e muito maiores do que zero, a utilização do EPAM é recomendada.

Uma vez que não existe uma medida que seja universalmente aceite como a mais adequada para ser aplicada a todas as situações de previsão, devem ser utilizadas várias para melhor avaliar e decidir sobre os diferentes métodos (Chu & Zhang, 2003; Ramos et al., 2015). Analisando os artigos que comparam os diferentes métodos para o segmento do retalho, a medida mais utilizada é o EPAM, seguida do EAM e do REQM. No entanto, em Alon et al. (2001) é aplicado um só critério (EPAM), sendo este complementado por dois testes de hipóteses (teste de Diebold-Mariano e teste dos sinais) que testam a significância da diferença entre as medidas do erro para métodos de previsão distintos.

Relativamente aos restantes parâmetros de comparação entre métodos, Veiga et al. (2016) mencionam que os modelos lineares têm a vantagem de serem mais fáceis de interpretar e de implementar que os não lineares, além da simplicidade e baixo custo que apresentam. Pan (2013) refere que, devido a essas características, os métodos lineares são muito populares na previsão de vendas no segmento do retalho e que, se estes forem capazes de capturar o comportamento dos dados, devem ser os escolhidos. No entanto, algumas características de dados reais (e.g., ciclos irregulares, *outliers*, etc.) não são possíveis de modelar usando esses métodos (Makridakis et al., 1998), pelo que se deve optar, nesses casos, por métodos não lineares.

Capítulo 3

Séries temporais

Neste Capítulo são introduzidas algumas noções sobre séries temporais e sobre processos estocásticos, fundamentais para a compreensão deste trabalho.

3.1 Conceito de série temporal

Uma *série temporal* é um conjunto de observações medidas sequencialmente no tempo. Essas medições podem ser feitas continuamente no tempo ou apenas em momentos específicos, geralmente igualmente espaçados (e.g., dias, meses, trimestres, anos, etc.). Por convenção, esses dois tipos de séries são designados de séries temporais *contínuas* e *discretas*, respetivamente, independentemente da natureza da variável medida, que tanto pode ser discreta como contínua, em ambos os casos (Chatfield, 2000, 2004). Além disso, as séries temporais podem também ser classificadas em *univariadas*, se são constituídas por observações de uma só variável, e em *multivariadas*, se se observarem mais variáveis em cada instante.

Usualmente, para analisar séries temporais contínuas, estas são transformadas em discretas, sendo, para isso, amostradas (ou discretizadas) em intervalos de tempo iguais (e.g., temperatura medida em intervalos horários). Desde que o intervalo de amostragem seja suficientemente pequeno, este procedimento não resulta, em regra, numa perda significativa de informação.

Quanto às séries temporais discretas, estas surgem de diversas formas: além de poderem resultar de uma série temporal contínua por discretização (como visto anteriormente), podem também ocorrer quando são agregados valores de uma variável ao longo de um período de tempo (e.g., total de vendas em meses sucessivos) ou até serem inerentemente discretas (e.g., o dividendo pago por uma empresa em anos sucessivos).

Independentemente do tipo de série, os dados de séries temporais têm, regra geral, uma propriedade muito particular: observações sucessivas são correlacionadas e, portanto, a análise desses dados deve ter em conta a ordem em que as observações são recolhidas. Efetivamente, no caso de séries temporais univariadas, cada observação pode ser vista como bivariada, considerando que a segunda variável corresponde ao tempo em que esta é observada.

Em termos de aplicações, a previsão de séries temporais é fundamental em diversas áreas: na Economia, e.g., preços de ações diários, totais de exportação em meses sucessivos, rendimentos médios mensais, lucros anuais de uma empresa, etc.; nas Ciências Físicas, nomeadamente na Meteorologia, Geofísica e Ciências Marinhas, e.g. precipitação diária, temperatura do ar medida a cada hora, dia ou mês sucessivo, etc.; no *Marketing*, e.g., planejar a produção com base em previsões de vendas mensais, etc.; nas Ciências Sociais, particularmente na Demografia, e.g., prever mudanças na população daí a 10 ou 20 anos, etc.; na Indústria, designadamente no controlo de processos, e.g., detetar mudanças no desempenho de um determinado processo de fabricação e efetuar as correções necessárias, etc..

3.1.1 Objetivos da análise de uma série temporal

O estudo de uma (ou várias) série temporal tem, dependendo de cada caso, objetivos específicos. No entanto, de uma forma geral, podem mencionar-se como principais motivações (Box et al., 2013; Chatfield, 2000):

- descrever os dados usando estatísticas descritivas e/ou métodos gráficos e compreender o mecanismo gerador da série, i.e., procurar encontrar razões que justifiquem o comportamento da série, monitorizar a sua trajetória, analisar periodicidades relevantes nos dados, etc.;
- encontrar um modelo estatístico adequado para descrever a evolução da série temporal. São vários os modelos propostos, entre os quais se destacam os modelos ARIMA, os modelos de alisamento exponencial, os modelos de regressão múltipla, etc.;
- prever o comportamento futuro da série, o que pode revelar-se extremamente útil na construção e execução de planos a curto, médio ou longo prazos e/ou no controlo de um determinado processo. Estas previsões podem ser a *1-passo*, se realizadas apenas para a observação seguinte, ou *multi-passos*, se englobam várias observações futuras.

Em Economia, de acordo com Tsay (2000), a análise de séries temporais é utilizada entre outros propósitos para, (a) estudar a estrutura dinâmica de um processo, (b) investigar as relações dinâmicas entre variáveis, (c) proceder ao ajustamento sazonal de dados económicos, (d) melhorar a análise de regressão quando os erros são correlacionados e (e) produzir previsões pontuais e/ou intervalos de previsão.

3.1.2 Componentes de uma série temporal

Segundo Persons (1919), a variação de uma série temporal pode ser decomposta em 4 componentes: a tendência (T), a componente sazonal (S), a componente cíclica (C) e a componente irregular/residual (E) (Kirchgässner & Wolters, 2008). Estas podem ser descritas como (Jebb et al., 2015; Alpuim, 1998):

Tendência (T) : é a inclinação que a série temporal apresenta ao longo do tempo, podendo esta ser linear ou não, crescente ou decrescente. A tendência pode ser consequência do facto dos valores observados dependerem de uma componente determinística que é função monótona do tempo, embora para muitos autores esta possa ser de natureza estocástica;

Sazonalidade (S) : corresponde a um padrão de aumento e diminuição que ocorre regularmente na série em períodos específicos, originando oscilações que se repetem (e.g., as vendas a retalho apresentam frequentemente um padrão sazonal mensal, onde cada mês, considerando períodos anuais, mostra constantemente a mesma posição relativamente aos restantes: picos de vendas ocorrem em meses de férias e, nos meses seguintes, verifica-se uma diminuição acentuada; Hyndman & Athanasopoulos, 2014). O movimento dentro de um período tem, então, duração fixa e é atribuído a fatores “sazonais”, i.e., relacionados com aspetos do calendário (e.g., os meses ou trimestres de um ano ou os dias de uma semana). A sazonalidade pode ser classificada como *aditiva*, se não depende do nível da série, ou *multiplicativa*, quando é proporcional ao mesmo (Chatfield, 2000). Esta fonte de variação é muitas vezes removida através de um procedimento designado de ajustamento sazonal (*seasonal adjustment*), uma vez que esta pode omitir outras componentes relevantes, tais como a tendência;

Componente cíclica (C) : é um padrão de flutuação que não apresenta qualquer periodicidade definida (i.e., a sua duração não é fixa e, portanto, o seu comprimento varia frequentemente de ciclo para ciclo) nem causa atribuída a fatores “sazonais”. Por outras palavras, os ciclos são qualquer componente não-sazonal

que apresenta um padrão reconhecível (e.g., ciclos económicos). Como estas flutuações são irregulares e se podem prolongar por vários anos, é frequente ignorar esta componente para séries “curtas”. Além disso, como se torna difícil dissociar a componente cíclica da tendência, estas são muitas vezes aglomeradas e referidas como componente de tendência-cíclica (*trend-cycle*);

Aleatoriedade (E) : é a componente que contém qualquer variação não explicada pelas componentes anteriores e representa o ruído aleatório. Quando esta componente é modelada por um processo estocástico de variáveis aleatórias não correlacionadas e identicamente distribuídas, é referida como um ruído branco.

Decomposição de uma série temporal

A maioria dos métodos tradicionais para a análise de séries temporais baseia-se na decomposição da variação da série nas componentes anteriores.

Sejam Y_t o valor da série temporal no tempo t , T_t a tendência no tempo t , S_t a componente sazonal no tempo t e E_t a componente irregular no tempo t . Considere-se que a componente cíclica (C), quando presente, se encontra incorporada na tendência (note-se que esta situação foi referida anteriormente como tendência-cíclica).

O *modelo de decomposição aditivo* descreve cada valor da série temporal como sendo a soma das suas componentes, i.e.,

$$Y_t = T_t + S_t + E_t,$$

enquanto que no *modelo de decomposição multiplicativo* cada observação é o produto dessas mesmas componentes, i.e.,

$$Y_t = T_t \times S_t \times E_t.$$

Um modelo aditivo é apropriado quando a magnitude das oscilações sazonais não varia com o nível da série. No entanto, se estas aumentam ou diminuem proporcionalmente com a tendência da série, então um modelo multiplicativo é o mais adequado (Makridakis et al., 1998).

Na Economia, a decomposição multiplicativa é a dominante, uma vez que a maioria das séries económicas sazonais apresenta uma variação sazonal que depende do nível da série. Muitas vezes, nesses casos, opta-se por aplicar uma transformação aos dados, nomeadamente a logarítmica, de forma a converter um modelo multipli-

cativo num modelo aditivo, i.e., aplicando a transformação logarítmica ao modelo multiplicativo e obtendo

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

Desta forma, pode, então, construir-se um modelo multiplicativo por ajustamento de um modelo aditivo ao logaritmo dos dados. Contudo, deve notar-se que um modelo de decomposição multiplicativo não deve ser implementado para séries temporais de valores negativos ou nulos (Caiado, 2016).

Evidentemente, os modelos de decomposição aditivo e multiplicativo não são as únicas formas de decompor uma série temporal e podem até ser misturados, dando origem a outros modelos que incluem relações tanto aditivas como multiplicativas, e.g., um *modelo multiplicativo com erros aditivos*, i.e.,

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

Uma ferramenta vantajosa no que toca à decomposição de séries temporais são os gráficos de decomposição (introduzidos por Cleveland & Terpenning, 1992), que permitem visualizar graficamente as várias componentes (Makridakis et al., 1998). No ambiente R, a função *decompose* divide a série temporal nas três componentes – sazonalidade (S), tendência (T) e aleatoriedade (E) –, usando médias móveis. Esta função mostra-se bastante útil, uma vez que permite avaliar as diversas componentes separadamente e, portanto, ajudar a identificar o comportamento individual das mesmas.

3.2 Processos estocásticos

Dado um processo estocástico $\{Y(t), t \in \mathcal{T}\}$, uma série temporal é um conjunto de observações do processo estocástico em instantes t_1, t_2, \dots, t_n . Geralmente, considera-se t inteiro (i.e., $t = 0, \pm 1, \pm 2, \dots$) e as observações são feitas em intervalos de tempo regulares, isto é, com a mesma amplitude (Alpuim, 1998).

Definição 1. *Um processo estocástico é qualquer família ou coleção de variáveis aleatórias $Y(t), t \in \mathcal{T}$, em que \mathcal{T} é um conjunto de índices representando o tempo.*

Ao conjunto de índices \mathcal{T} dá-se o nome de *espaço de parâmetros* e ao contradomínio das variáveis aleatórias $Y(t)$ dá-se o nome de *espaço de estados*, que se representa por S . Quanto à natureza de \mathcal{T} , se $\mathcal{T} = \mathbb{Z}$ ou $\mathcal{T} = \mathbb{N}$ diz-se que o processo é de

tempo discreto e se $\mathcal{T} = \mathbb{R}$ ou, mais comumente, $\mathcal{T} = \mathbb{R}^+$ diz-se que o processo é de tempo contínuo.

Uma maneira de caracterizar um processo estocástico consiste em especificar a distribuição de probabilidade conjunta de n variáveis aleatórias $(Y(t_1), \dots, Y(t_n))$ para todos os inteiros n e quaisquer pontos t_1, \dots, t_n . No entanto, esta forma de definir um processo estocástico é complicada e, na prática, não é viável. Assim, uma alternativa mais simples de descrever um processo estocástico é através dos momentos do processo, em particular os primeiro e segundo momentos, designados por valor médio $\mu(t) = E[Y(t)]$ e função de autocovariância $\gamma(t_1, t_2) = E[(Y(t_1) - \mu(t_1))(Y(t_2) - \mu(t_2))]$, respetivamente. A variância $\sigma^2(t) = Var[Y(t)]$ é um caso particular da função de autocovariância (quando $t_1 = t_2$) mas, por si só não é suficiente para definir os segundos momentos de uma sequência de variáveis aleatórias (Chatfield, 2004).

A série de valores observados, que compõem a série temporal, é considerada apenas uma única realização (ou trajetória) de um processo estocástico, de entre todas as possíveis. Ou seja, o processo estocástico é um só, mas as séries que ele pode gerar são infinitas. Assim, a relação entre uma série temporal e o processo estocástico que a gerou é semelhante à relação entre uma amostra e a população de onde foi extraída. Em termos gerais, a análise de séries temporais consiste em fazer inferências sobre um processo estocástico desconhecido tendo como informação disponível uma única realização observada (Cordeiro, 2011).

Os processos estocásticos dividem-se em estacionários e não estacionários. Nesta Secção introduzem-se os dois tipos de estacionariedade (forte e fraca), alguns procedimentos que permitem transformar processos não estacionários em estacionários e outras ferramentas essenciais para a posterior modelação das séries temporais (e.g., as funções de autocorrelação, FAC, e autocorrelação parcial, FACP, e o processo de ruído branco).

3.2.1 Processos estocásticos estacionários

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

$$F_{(Y(t_1), \dots, Y(t_n))}(y_1, \dots, y_n) = F_{(Y(t_1 + \delta), \dots, Y(t_n + \delta))}(y_1, \dots, y_n)$$

em todos os pontos (y_1, \dots, y_n) .

Desta forma, pode dizer-se que um processo fortemente estacionário usufrui da propriedade de que a distribuição de um qualquer conjunto de margens se mantém a mesma, quando estas são sujeitas a uma translação no tempo (Alpuim, 1998).

A estacionariedade no sentido estrito é uma propriedade demasiado exigente e, na maioria dos casos, de difícil verificação. Os processos estacionários de 2.^a ordem (ou fracamente estacionários) obedecem a uma propriedade mais fraca mas que, grosso modo, descreve o mesmo tipo de comportamento (Murteira et al., 1993).

Definição 3. *Um processo $\{Y(t), t \in \mathcal{T}\}$ diz-se estacionário de 2.^a ordem (ou fracamente estacionário) se e só se todos os momentos até à 2.^a ordem de $(Y(t_1), \dots, Y(t_n))$ existem e são iguais aos momentos correspondentes até à 2.^a ordem de $(Y(t_1 + \delta), \dots, Y(t_n + \delta))$. Logo, num processo fracamente estacionário:*

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

Está implícito, pelas condições de estacionariedade fraca, que se assume, nesse caso, que os momentos até à 2.^a ordem existem e são finitos. Assim, pelas Definições 2 e 3, se $Y(t)$ é estritamente estacionário e os seus momentos até à 2.^a ordem são finitos, então $Y(t)$ também é estacionário de 2.^a ordem. A recíproca pode não se verificar. No entanto, se a série temporal $Y(t)$ tem distribuição Normal, então a estacionariedade fraca é equivalente à estacionariedade forte (Tsay, 2010).

De agora em diante, considerem-se apenas processos estacionários de 2.^a ordem, designados simplesmente por processos estacionários, e Y_t a representar um processo estocástico, independentemente do tempo t ser discreto ou contínuo.

Funções de autocovariância, autocorrelação e autocorrelação parcial

Definição 4. *Para um processo estacionário, define-se a função de autocovariância*

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

que mede a intensidade com que covariam (se acompanham) pares de valores do processo separados por um intervalo (lag) de amplitude k .

A função de autocovariância γ_k é definida para $k \in \mathbb{R}$ se o processo é de tempo contínuo e para $k \in \mathbb{Z}$ se for de tempo discreto, i.e., $k = 0, \pm 1, \pm 2, \dots$. Além disso, esta função satisfaz as seguintes propriedades:

1. $\gamma_0 = Cov[Y_t, Y_t] = Var[Y_t] = \sigma^2$;
2. $\gamma_k = \gamma_{-k}$, i.e., a função é par e dispensa a representação gráfica para $k < 0$;
3. $|\gamma_k| \leq \gamma_0$, como consequência da desigualdade de Cauchy-Schwarz¹;
4. A função γ_k é semidefinida positiva, i.e., para qualquer conjunto de números reais $\alpha_1, \dots, \alpha_n$ e instantes de tempo t_1, \dots, t_n ,

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

Definição 5. Para um processo estacionário, define-se a função de autocorrelação

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

que mede a correlação entre pares de valores do processo separados por um intervalo (lag) de amplitude k .

A representação gráfica de ρ_k em função de k designa-se por correlograma. O comportamento do correlograma dá indicações cruciais sobre as características da série e constitui um auxiliar importante na identificação do modelo que lhe é mais adequado (como será visto no Capítulo 4). Na maioria dos casos, o aumento de k traduz-se no decréscimo de ρ_k (e, obviamente, de γ_k). De facto, à medida que a amplitude do intervalo (k) aumenta é de esperar que a capacidade de memória do processo seja limitada, e, portanto, que no momento $t + k$ se encontre pouco refletido o que se passou no momento t (Murteira et al., 1993). Desta forma, espera-se que a correlação temporal diminua ($\rho_k \rightarrow 0$), quando o desfazamento entre duas observações aumenta ($k \rightarrow +\infty$).

Intuitivamente, pode interpretar-se ρ_k como uma medida da semelhança entre cada realização e a mesma realização deslocada k unidades de tempo (Murteira et al., 1993).

Tal como no caso da função de autocovariância, a função de autocorrelação ρ_k pode estar definida para $k \in \mathbb{R}$ ou para $k \in \mathbb{Z}$, consoante o processo for de tempo

¹Desigualdade de Cauchy-Schwarz: $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$.

contínuo ou discreto, respetivamente. Além disso, esta função satisfaz as seguintes propriedades:

1. $\rho_0 = Corr[Y_t, Y_t] = 1$;
2. $\rho_k = \rho_{-k}$, i.e., a função é par e dispensa a representação gráfica para $k < 0$;
3. $|\rho_k| \leq 1$, como consequência da desigualdade de Cauchy-Schwarz;
4. A função ρ_k é semidefinida positiva, i.e., para qualquer conjunto de números reais $\alpha_1, \dots, \alpha_n$ e instantes de tempo t_1, \dots, t_n ,

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

Como os processos estacionários se caracterizam pelos parâmetros já referidos, a estimação dos mesmos tem muita importância. Se se considerar um conjunto de n observações de um processo estacionário Y_t durante um certo período de tempo, ou seja, Y_1, Y_2, \dots, Y_n , podem utilizar-se os estimadores clássicos dos parâmetros (Alpuim, 1998), i.e.,

- para a média μ usar $\bar{Y} = \frac{1}{n} \sum_{t=1}^n Y_t$;
- para a autocovariância γ_k usar $\hat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$;
- para a autocorrelação ρ_k usar $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$.

Por vezes, além de se estudar a correlação de uma forma global, interessa também investigar a correlação parcial que existe entre Y_t e Y_{t+k} quando se fixam as variáveis intermédias $Y_{t+1}, Y_{t+2}, \dots, Y_{t+k-1}$, i.e., a correlação simples entre Y_t e Y_{t+k} depois de eliminar o efeito que as variáveis intermédias exercem sobre elas (Caiado, 2016).

Sejam $E[Y_t] = 0$ e $Var[Y_t] = 1$ e considere-se a regressão linear múltipla de Y_{t+k} sobre Y_{t+k-1}, \dots, Y_t , i.e.,

$$Y_{t+k} = \phi_{k1} Y_{t+k-1} + \dots + \phi_{kk} Y_t + \epsilon_{t+k}, \quad (3.1)$$

onde ϕ_{kj} , $j = 1, \dots, k$, são os coeficientes do modelo de regressão linear que se considera ter erros gaussianos. O valor ϕ_{kk} é o coeficiente de correlação do modelo

de regressão linear (3.1) onde $\{\epsilon_t, t \in \mathbb{Z}\}$ tem distribuição Normal de parâmetros $(0, \sigma^2)$ e ϵ_{t+k} é independente de $\{Y_{t+k-j}, j \geq 1\}$. O coeficiente ϕ_{kk} exprime a variação em Y_{t+k} que acompanha em média uma variação unitária em Y_t quando $Y_{t+1}, \dots, Y_{t+k-1}$ são constantes; tal variação pode interpretar-se como a correlação parcial entre Y_t e Y_{t+k} .

Multiplicando, então, ambos os membros de (3.1) por Y_{t+k-j} , $j = 1, \dots, k$, calculando os valores esperados e dividindo por ρ_0 obtém-se um sistema constituído pelas equações

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \dots + \phi_{kk}\rho_{k-j}, \quad j = 1, 2, \dots, k; \quad (3.2)$$

resolve-se o sistema em ordem a ϕ_{kj} , $j = 1, 2, \dots, k$, utilizando a regra de Cramer, e obtém-se, assim, a função de autocorrelação parcial, ϕ_{kk} .

Uma definição alternativa para esta função é:

Definição 6. O conjunto de autocorrelações parciais de desfasamento (lag) k é dado por $\{\phi_{kk} : k = 1, 2, \dots\}$ onde

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

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

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

Resolvendo o sistema constituído pelas equações em (3.2) ou seguindo a Definição 6, obtém-se as seguintes propriedades:

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

Ruído branco

Um processo estocástico $\{\epsilon_t, t \in \mathbb{Z}\}$ diz-se um processo puramente aleatório ou processo de ruído branco quando é formado por uma sucessão de variáveis aleatórias não correlacionadas e identicamente distribuídas, de média e variância constantes, i.e., um processo estocástico diz-se um ruído branco se e só se satisfaz as seguintes condições:

1. $E[\epsilon_t] = \mu_\epsilon$ (usualmente $\mu_\epsilon = 0$);
2. $Var[\epsilon_t] = \sigma_\epsilon^2$;
3. $Cov(\epsilon_t, \epsilon_{t+k}) = \gamma_k = 0, \quad k = \pm 1, \pm 2, \dots$

Se, além disso, as variáveis aleatórias seguem uma distribuição Normal (i.e., $\epsilon_t \sim N(\mu_\epsilon, \sigma_\epsilon^2)$), então o processo é designado de ruído branco gaussiano.

Um ruído branco é, então, um processo estacionário cujas funções de autocorrelação (FAC) e autocorrelação parcial (FACP) são nulas para todo o $k \neq 0$. Na Figura 3.1 encontra-se representada uma trajetória de um processo de ruído branco e as respetivas FAC e FACP empíricas.

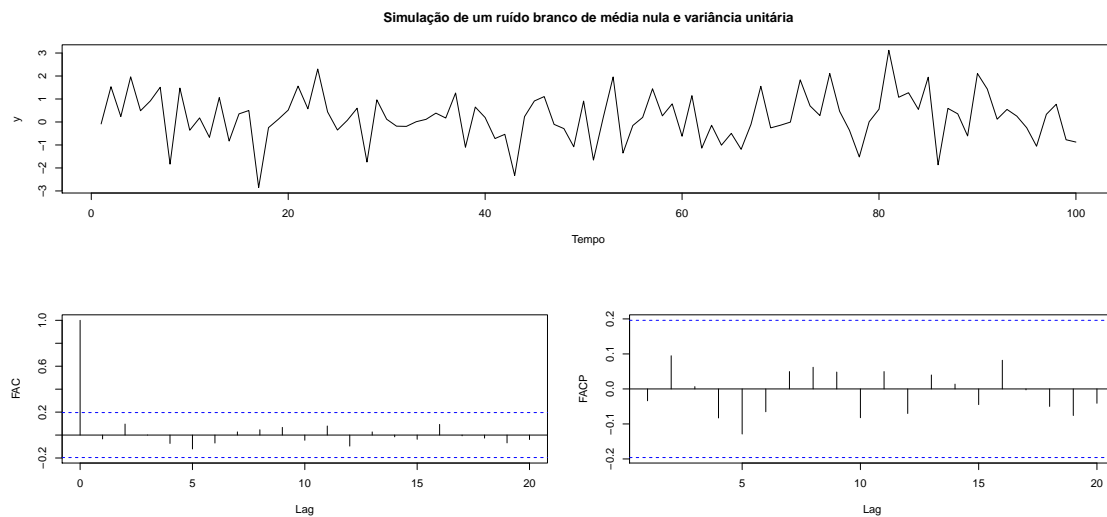


Figura 3.1: Simulação de um ruído branco e respetivas FAC e FACP empíricas.

Embora o ruído branco dificilmente se observe em séries reais, este desempenha um papel fundamental na construção de modelos probabilísticos ou estocásticos. Além disso, um bom modelo de previsão deve ser aquele que produz erros de previsão com comportamento análogo a um ruído branco, isto porque um ruído branco é imprevisível (Caiado, 2016).

3.2.2 Processos estocásticos não estacionários

Numa série estacionária, os valores futuros serão similares aos do passado e, por isso, a estacionariedade é um importante pressuposto quando se pretende obter previsões com base em observações passadas. Alguns modelos de previsão de séries temporais assumem que a série já é ou pode ser transformada numa série estacionária (Jebb et al., 2015).

Muitas séries temporais, nomeadamente as associadas a fenómenos ambientais ou económicos, são não estacionárias. Um processo pode ser não estacionário por a média e/ou a variância serem funções do tempo e não constantes. Uma série estacionária em média não é necessariamente estacionária em variância. Para ultrapassar este problema pode recorrer-se a transformações que estabilizam a média e/ou a variância convertendo, assim, uma série temporal não estacionária numa série estacionária. No caso de se estar perante uma série não estacionária em média nem em variância deve proceder-se, em primeiro lugar, à estabilização da variância e só depois da média (Murteira et al., 1993; Caiado, 2016). Na Figura 3.2 podem ver-se alguns tipos de não estacionariedade: do lado esquerdo, uma série não estacionária em variância, ainda que estacionária em média; e do lado direito, uma série com tendência linear (não estacionária em média) e variância crescente com a tendência (não estacionária em variância).

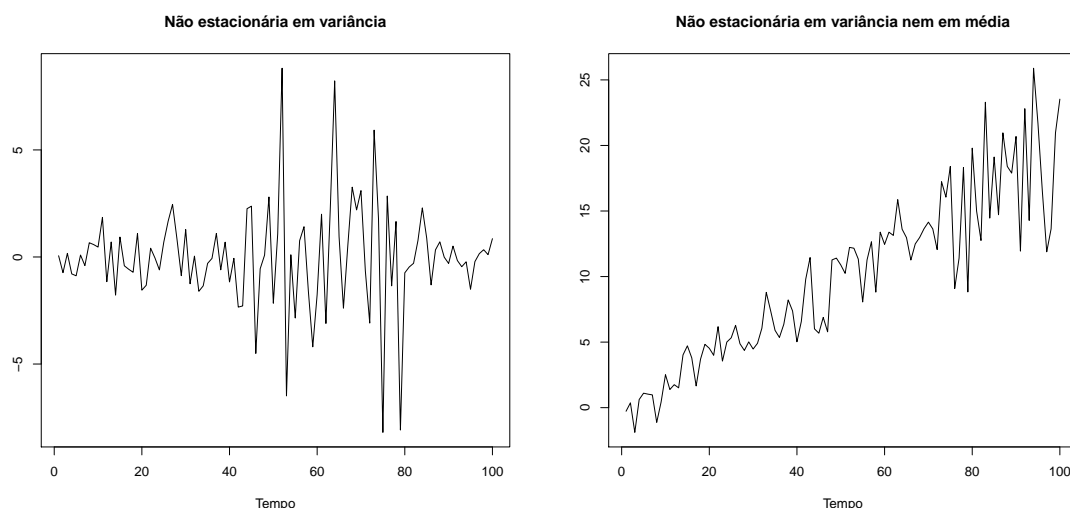


Figura 3.2: Duas realizações de séries temporais não estacionárias.

Em muitas situações, há procedimentos que procuram “retirar” a tendência e a sazonalidade a uma série temporal, permitindo, assim, que se atinja a estacionariedade. Estes métodos seguem os modelos de decomposição anteriormente descritos

(Secção 3.1.2) e consistem, naturalmente, na estimação das componentes tendência (T_t) e sazonalidade (S_t) através de funções determinísticas (ou outras abordagens), de modo que a série após “remoção” dessas componentes passe a ser bem modelada por um processo estacionário (Alpuim, 1998).

No entanto, em muitas séries temporais é possível realizar outro tipo de transformações, que permitem transformar séries não estacionárias em estacionárias.

Transformações para a estacionariedade

Uma forma de estabilizar a média consiste em usar processos de diferenciação (regular), que resultam da aplicação do operador diferença ∇ , definido como $\nabla Y_t = Y_t - Y_{t-1}$, à série temporal não estacionária. Assim, se uma série, Y_t , for não estacionária, pode pensar-se em transformá-la numa série estacionária, aplicando uma diferenciação de primeira ordem, i.e.,

$$\nabla Y_t = Y_t - Y_{t-1}, \quad t = 2, 3, \dots, n.$$

Se a diferenciação de 1.^a ordem não for suficiente para obter uma série estacionária, podem obter-se as diferenças de 2.^a ordem, que correspondem às diferenças das primeiras diferenças da série original, i.e.,

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = \nabla(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}, \quad t = 3, 4, \dots, n.$$

O operador de diferenciação de ordem d , para qualquer inteiro $d \geq 1$, consiste em diferenciar a série d vezes, ou seja,

$$\nabla^d Y_t = \nabla(\nabla^{d-1} Y_t), \quad t = d + 1, \dots, n.$$

Contudo, a diferenciação inapropriada de uma série já estacionária é indesejável e deve evitar-se. De facto, o objetivo é determinar a série estacionária obtida pela menor diferenciação, uma vez que a variância aumenta com a diferenciação. De uma forma geral, se a série transformada $\nabla^{d_0} Y_t$ é estacionária, então, para qualquer $d > d_0$, a série $\nabla^d Y_t$ é também estacionária, mas tem maior variância. Conclui-se, portanto, que se deve evitar a sobrediferenciação para não introduzir variação indesejada na série transformada. Efetivamente, caso este procedimento seja adequado, as diferenciações de 1.^a e 2.^a ordens devem ser suficientes para obter uma série estacionária.

Se uma série temporal é diferenciada uma vez e a série diferenciada é estacionária,

então diz-se que a série original é integrável de ordem 1 e representa-se por $I(1)$. Em geral, se a série for diferenciada d vezes, é integrável de ordem d ou $I(d)$. Por convenção, se $d = 0$, o processo $I(0)$ é um processo estacionário.

Para estabilizar a variância de uma série não estacionária, pode utilizar-se um método de transformação paramétrica, conhecido como transformação de Box-Cox, baseado na seguinte expressão

$$Z_t = T(Y_t) = \begin{cases} \frac{Y_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log Y_t, & \lambda = 0 \end{cases},$$

onde os valores de λ estão no intervalo $[-1; 1]$. As transformações mais recorrentes encontram-se sumariadas na Tabela 3.1.

Tabela 3.1: Transformações usuais de Box-Cox.

λ	Transformação
-1	$Z_t = \frac{1}{Y_t}$
-0,5	$Z_t = \frac{1}{\sqrt{Y_t}}$
0	$Z_t = \log Y_t$
0,5	$Z_t = \sqrt{Y_t}$
1	$Z_t = Y_t$

No entanto, note-se que algumas destas transformações (nomeadamente a do logaritmo) apenas estão definidas para séries de valores positivos. Para contornar esse problema, pode encontrar-se uma constante c tal que $Y_t + c > 0$ e só depois aplicar as transformações.

Como já foi referido, a transformação logarítmica permite converter o efeito sazonal multiplicativo em aditivo, sendo, também por isso, uma das transformações mais usadas dentro da família Box-Cox. Além disso, uma vez que esta transformação permite estabilizar as diferenças entre os valores da série temporal (i.e., a sua variância), ela também pode ser utilizada para atenuar o efeito de possíveis *outliers*.

É importante lembrar que, sempre que se aplica alguma transformação aos dados, todas as previsões geradas pelo modelo selecionado estarão nas unidades transformadas. Uma vez ajustado o modelo e estimados os parâmetros, devem, então, ser revertidas as transformações de modo a obter previsões nas unidades originais (Jebb et al., 2015).

Passeio aleatório

Considere-se o seguinte modelo de tendência estocástica,

$$Y_t = Y_{t-1} + \epsilon_t, \quad (3.3)$$

onde ϵ_t é um ruído branco. Este modelo é vulgarmente conhecido por modelo de passeio aleatório (*random walk*), uma vez que descreve a série no momento t com base no seu valor passado (no momento $t - 1$) mais um choque aleatório.

Em geral, os “passeios aleatórios” caracterizam-se por movimentos de tendência crescente ou decrescente em períodos longos, seguidos de mudanças bruscas de sentido imprevisível (Caiado, 2016). Contudo, uma vez que as aparentes tendências são apenas o resultado da existência de um elevado número de valores positivos em relação ao número de valores negativos (no caso de tendências crescentes) ou vice-versa (no caso de tendências decrescentes), o valor médio de um processo aleatório é constante. De facto, se Y_0 for conhecido, o modelo de passeio aleatório pode ser reescrito como

$$Y_t = Y_0 + \sum_{i=1}^t \epsilon_i, \quad (3.4)$$

e, calculando o valor esperado de Y_t , conclui-se que este é igual a Y_0 e, portanto, constante. Apesar disso, o modelo de passeio aleatório é um processo não estacionário, uma vez que a sua variância depende de t (Enders, 2015). No entanto, aplicando uma transformação, nomeadamente uma diferenciação de 1.^a ordem, $\nabla Y_t = \epsilon_t$, é possível obter um ruído branco, que é um processo estacionário.

Este modelo é muito utilizado na modelação e previsão de séries financeiras não estacionárias e na análise de importantes relações económicas. Por exemplo, nos mercados financeiros, as previsões dos preços das ações têm um comportamento do tipo “passeio aleatório”, no sentido em que as melhores previsões para o dia de amanhã correspondem aos preços observados hoje (Caiado, 2016).

Na Figura 3.3 pode observar-se uma série simulada com 100 observações de um modelo de passeio aleatório, com ϵ_t um ruído branco gaussiano de média nula e variância unitária, bem como o resultado da diferenciação de 1.^a ordem da mesma. O correlograma de um passeio aleatório apresenta correlações positivas que decaem muito lentamente para zero.

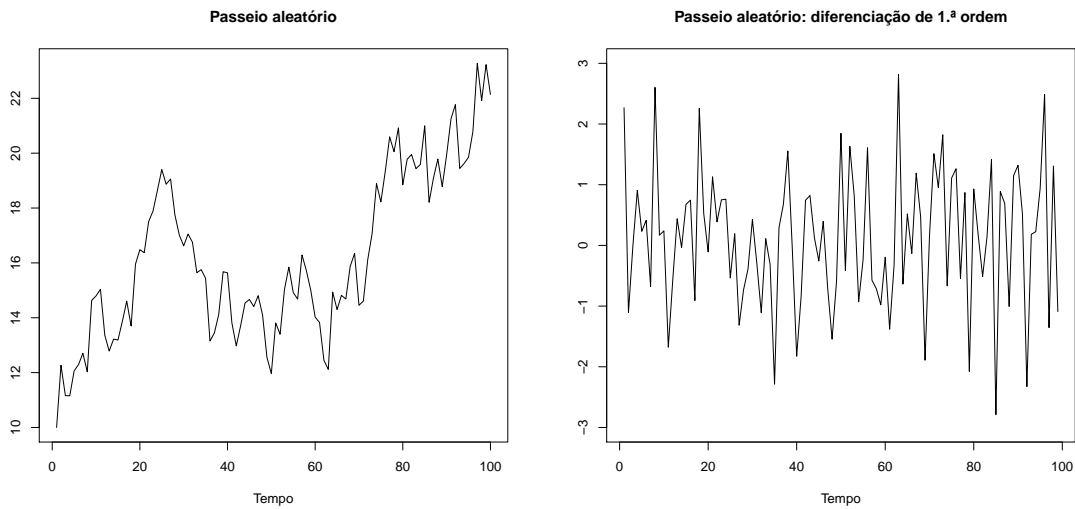


Figura 3.3: Simulação de um modelo de passeio aleatório e as suas diferenças de 1.^a ordem.

Passeio aleatório com *drift*

O modelo de passeio aleatório com *drift* é uma extensão do modelo anterior, quando se adiciona um termo constante a_0 , i.e.,

$$Y_t = a_0 + Y_{t-1} + \epsilon_t, \tag{3.5}$$

onde ϵ_t é um ruído branco. Neste caso, considera-se que a tendência é parcialmente determinística e parcialmente estocástica. Por esse motivo, o valor médio deste processo depende de t , não sendo constante. De facto, se Y_0 for conhecido, o modelo de passeio aleatório com *drift* pode ser reescrito como

$$Y_t = Y_0 + a_0 t + \sum_{i=1}^t \epsilon_i, \tag{3.6}$$

e, calculando o valor esperado de Y_t , conclui-se que este é igual a $Y_0 + a_0 t$ e, portanto, dependente de t . Desta forma, também o modelo de passeio aleatório com *drift* é não estacionário, podendo ser transformado em estacionário se se aplicar uma diferenciação de 1.^a ordem.

Na Figura 3.4 pode observar-se uma série simulada com 100 observações de um modelo de passeio aleatório com *drift*, onde ϵ_t é um ruído branco gaussiano de média nula e variância unitária e $a_0 = 0,5$, bem como o resultado da diferenciação de 1.^a ordem da mesma.

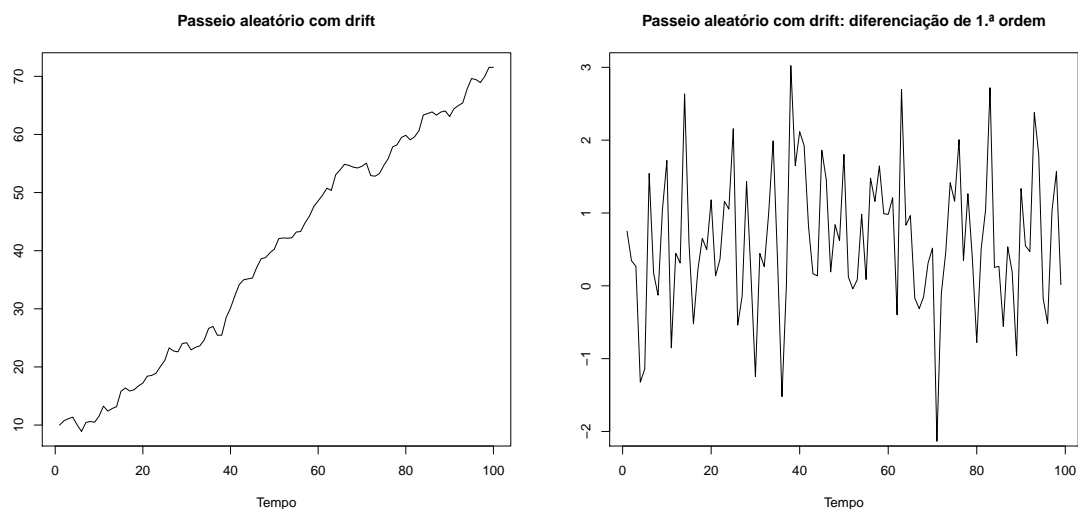


Figura 3.4: Simulação de um modelo de passeio aleatório com *drift* e as suas diferenças de 1.^a ordem.

Usualmente, neste tipo de processos, a tendência determinística é dominante. No entanto, o aumento da variância de ϵ_t ou a diminuição do valor absoluto de a_0 podem mascarar a existência de uma tendência determinística (Enders, 2015).

Análise de estacionariedade

Uma forma elementar de analisar a estacionariedade de uma série é representar graficamente os dados ao longo do tempo. No entanto, esta análise é subjetiva e, apesar de útil, deve ser confirmada através de testes estatísticos formais. Existem vários testes para realizar este estudo, baseados, na sua maioria, em encontrar uma raiz unitária. Segundo a literatura, deve utilizar-se mais do que um teste de raiz unitária de modo a avaliar a estacionariedade da série. Alguns dos testes mais utilizados são os testes de Dickey-Fuller (DF), Dickey-Fuller Aumentado (*Augmented Dickey Fuller*, ADF), Phillips-Perron (PP) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS). Dentro destes testes salientam-se dois tipos: aqueles cuja hipótese a testar é a presença de uma raiz unitária (ou seja, a não estacionariedade) e, no caso de esta não ser rejeitada, são fornecidas informações sobre o número de diferenciações necessárias para atingir a estacionariedade (DF, ADF e PP); e aqueles cuja hipótese nula é a estacionariedade (KPSS). Mais detalhes sobre estes testes podem ver-se em Dickey & Fuller (1979); Said & Dickey (1984); Phillips & Perron (1988); Kwiatkowski et al. (1992).

Teste de Dickey-Fuller

Considere-se um processo definido por

$$Y_t = \phi Y_{t-1} + \epsilon_t, \quad -1 \leq \phi \leq 1, \quad (3.7)$$

onde ϵ_t é um ruído branco. Este processo é estacionário se $|\phi| < 1$ (trata-se de um processo autorregressivo de ordem 1, como se verá na Secção 4.1.1). No entanto, quando $\phi = 1$ este processo é um passeio aleatório, o que, como foi visto anteriormente, equivale a fazer uma diferenciação de 1.^a ordem. Assim, se $\phi = 1$, pode dizer-se que a série é não estacionária.

O processo descrito pela equação (3.7) pode ser escrito na forma das diferenças, ou seja,

$$Y_t = \phi Y_{t-1} + \epsilon_t \Leftrightarrow Y_t - Y_{t-1} = \phi Y_{t-1} - Y_{t-1} + \epsilon_t \quad (3.8)$$

$$\Leftrightarrow \nabla Y_t = (\phi - 1)Y_{t-1} + \epsilon_t \quad (3.9)$$

$$\Leftrightarrow \nabla Y_t = \delta Y_{t-1} + \epsilon_t, \quad (3.10)$$

onde $\delta = \phi - 1$ e ϵ_t é um processo estacionário. Assim, as hipóteses para o teste DF em termos de δ são dadas por

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

De notar que, se $\delta = 0$ então $Y_t \sim I(1)$. Isto é, a série das primeiras diferenças é estacionária e a série original é um passeio aleatório. Se $-2 < \delta < 0$, então o processo Y_t é estacionário. Com base nesta modificação, Dickey & Fuller (1979) propõem o teste

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

Se a hipótese H_0 não é rejeitada então o processo Y_t é não estacionário, i.e., $Y_t \sim I(1)$. No caso de $\delta < 0$ então $\phi - 1 < 0 \Leftrightarrow \phi < 1$ e, tendo em conta as condições de estacionariedade para este processo, a série Y_t é estacionária.

Além da equação (3.10), Dickey e Fuller consideram também duas equações alternativas:

$$\nabla Y_t = a_0 + \delta Y_{t-1} + \epsilon_t \quad (3.11)$$

$$\nabla Y_t = a_0 + a_1 t + \delta Y_{t-1} + \epsilon_t \quad (3.12)$$

A diferença entre as três equações reside, essencialmente, na presença/ausência de

uma constante a_0 e/ou de um termo determinístico $a_1 t$. Assim, a rejeição da hipótese nula, $H_0 : \delta = 0$, considerando a equação (3.11), implica que o processo Y_t é estacionário de média não nula. No caso da equação (3.12), a indicação é de que $Y_t \sim I(0)$ com uma tendência determinística.

Ao aplicar o teste DF deve verificar-se se a sequência ϵ_t é não correlacionada, uma vez que, quando este pressuposto não se verifica, estes testes tendem a exibir um fraco comportamento. Caso isso aconteça, o teste DF não pode ser aplicado e deve optar-se por uma versão do teste para ordens superiores, ou seja, o teste ADF (Cordeiro, 2011).

Teste de Dickey-Fuller Aumentado

Muitas séries apresentam uma estrutura mais complexa do que a captada pelo modelo (3.7). Para lidar com esses casos, surge uma variante do teste DF capaz de integrar modelos mais complexos. Esta inovação deve-se a Said & Dickey (1984) e, desde então, este tem sido um dos testes mais usados no estudo da estacionariedade de séries temporais.

Considere-se, neste caso, um processo mais complexo que o apresentado em (3.7) definido por

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t. \quad (3.13)$$

Assim, repetindo o processo da (3.8) a (3.10), obtém-se

$$\nabla Y_t = \delta Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \nabla Y_{t-j} + \epsilon_t, \quad (3.14)$$

onde $\delta = \sum_{i=1}^p \phi_i - 1$, $\gamma_j = -\sum_{i=j+1}^p \phi_i$, $\nabla Y_{t-j} = Y_{t-j} - Y_{t-j-1}$ e ϵ_t é um ruído branco. Esta decomposição separa o modelo (3.13) em dois termos: Y_{t-1} e as $p-1$ primeiras diferenças. No caso em que Y_t é $I(1)$ (passeio aleatório), esta separação envolve uma componente $I(1)$ e $p-1$ componentes $I(0)$ (estacionárias). Na terminologia comum, diz-se que o modelo original foi aumentado por $p-1$ componentes de primeiras diferenças, dando origem à designação $ADF(p-1)$.

Para o modelo (3.14), a existência de uma raiz unitária é garantida se $\sum_{i=1}^p \phi_i = 1$, i.e., se $\delta = 0$. Assim, as hipóteses a testar são, também neste caso,

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

e a não rejeição da hipótese nula implica a não estacionariedade de Y_t . Como este teste é unilateral à esquerda, a hipótese nula é rejeitada a um nível de significância

α se a estatística de teste for inferior ou igual ao quantil $(1 - \alpha)100\%$ da distribuição correspondente (valor crítico).

Além da equação (3.14) e à semelhança do que acontece no teste DF, são propostas duas outras equações

$$\nabla Y_t = a_0 + \delta Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \nabla Y_{t-j} + \epsilon_t, \quad (3.15)$$

$$\nabla Y_t = a_0 + a_1 t + \delta Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \nabla Y_{t-j} + \epsilon_t, \quad (3.16)$$

cuja diferença reside, novamente, na presença/ausência de uma constante a_0 e/ou de um termo determinístico $a_1 t$.

Um dos principais problemas do teste ADF é decidir qual o número de termos a incluir na equação a ser testada, ou seja, o valor de p . Para essa escolha existem várias estratégias, entre as quais:

1. Escolher o valor de p empiricamente. A ideia é incluir um número suficiente de termos de forma que a estrutura dos resíduos não apresente correlação;
2. Minimizar algum critério de informação como, por exemplo, o critério de informação de Akaike (*Akaike information criterion*, AIC);
3. Usar a regra proposta por Ng & Perron (1995): primeiro, define-se um limite máximo para p , p_{max} ; de seguida, calcula-se o teste ADF considerando $p = p_{max}$; se o valor absoluto da estatística t para testar a significância da diferença de ordem p for maior do que 1,6, define-se $p = p_{max}$ e prossegue-se com o teste; caso contrário, reduz-se p em uma unidade e repete-se o processo. Para determinar p_{max} pode utilizar-se (proposta de Schwert (1989))

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

onde $[x]$ representa a parte inteira de x e T o número de observações. Esta será a estratégia adotada para definir o valor de p .

Independentemente da estratégia escolhida, o importante é averiguar se os resíduos do modelo atuam como um ruído branco antes da aplicação do teste.

Teste de Phillips-Perron

O teste de Phillips-Perron (PP) difere do teste ADF essencialmente na estrutura permitida para os erros, ϵ_t . De facto, enquanto que o teste ADF exige que estes sejam independentes, o teste PP permite que os erros sejam correlacionados e, possivelmente, heterocedásticos. Assim, as equações a testar são iguais às apresentadas para o teste ADF, com a única diferença que ϵ_t deixa de ser um ruído branco. As hipóteses a testar são também idênticas e, portanto, a rejeição da hipótese nula implica a estacionariedade do processo.

Teste de KPSS

O teste de KPSS, proposto por Kwiatkowski, Phillips, Schmidt e Shin, serve, tal como os anteriores, para avaliar a estacionariedade de um processo. No entanto, neste caso, as hipóteses a testar estão trocadas, ou seja,

H_0 : O processo é estacionário *vs* H_1 : O processo é não estacionário.

Se se considerar o processo Y_t , a equação deste teste decompõe Y_t numa soma de três componentes: uma tendência determinística (T_t), um passeio aleatório (μ_t) e um erro estacionário (u_t), i.e.,

$$Y_t = T_t + \mu_t + u_t, \quad (3.17)$$

$$\mu_t = \mu_{t-1} + \epsilon_t, \quad (3.18)$$

onde ϵ_t é um ruído branco. Este teste é unilateral à direita e, portanto, a hipótese nula é rejeitada a um nível de significância α se a estatística de teste for superior ou igual ao quantil $(1 - \alpha)100\%$ da distribuição correspondente (valor crítico).

Capítulo 4

Modelos e métodos de previsão

A distinção entre modelo, representação matemática da estrutura estocástica de uma série temporal através de uma equação ou sistema de equações, e método, procedimento para calcular previsões, nem sempre foi clara (Cordeiro, 2011).

Um modelo estatístico determina um processo gerador dos dados, que pode ser usado para obter toda a distribuição de probabilidade para um instante futuro. Os modelos utilizados para as séries temporais são processos estocásticos e, portanto, controlados por leis de probabilidade. Além das previsões pontuais para o futuro, um modelo também permite calcular os intervalos de previsão, com um nível de confiança associado.

Um método de previsão é um procedimento para calcular previsões a partir de valores presentes e passados. Como tal, pode ser simplesmente um algoritmo e não depender de um modelo de probabilidade subjacente ou, alternativamente, surgir da identificação de um modelo específico para os dados fornecidos e da localização de previsões condicionadas a esse modelo (Chatfield, 2000). De acordo com Chatfield (2004), os métodos de previsão podem ser classificados em três grupos:

Métodos subjetivos, que, tal como o nome indica, envolvem alguma subjetividade, baseada em julgamento, intuição, conhecimento comercial ou qualquer outra informação relevante;

Métodos univariados, onde as previsões dependem apenas dos valores presentes e passados de uma série temporal única;

Métodos multivariados, em que as previsões para uma dada variável dependem, pelo menos em parte, dos valores de uma ou mais variáveis de séries temporais adicionais, designadas de preditores ou variáveis explicativas.

Na prática, um método de previsão pode resultar da combinação de várias abordagens como, por exemplo, quando métodos univariados ou multivariados envolvem alguma subjetividade, de forma a ter em conta informações externas que dificilmente se expressam através de um modelo matemático.

A escolha do método depende de uma variedade de considerações, incluindo:

- o objetivo do cálculo das previsões;
- o tipo de série temporal e a presença/ausência de componentes como a tendência e/ou a sazonalidade;
- a dimensão da série temporal (n);
- o horizonte de previsão (h);
- o conhecimento e a experiência do analista;
- e a disponibilidade dos programas informáticos.

De facto, existe uma enorme variedade de métodos de previsão que se podem aplicar a uma série temporal, cada um com as suas capacidades e limitações. Em cada caso, deve escolher-se o método que pareça mais adequado, nunca excluindo a hipótese de experimentar outros métodos. Nesta dissertação são estudadas três abordagens distintas: a clássica de Box-Jenkins, a não paramétrica de Holt-Winters e a de regressão linear múltipla com erros correlacionados. Estas metodologias permitem modelar tendências e flutuações sazonais presentes em séries temporais e, dentro dos métodos de previsão mais tradicionais, são frequentemente utilizadas em estudos comparativos na literatura (Alon et al., 2001; Chu & Zhang, 2003; Veiga et al., 2016).

4.1 Modelos SARIMA e a metodologia Box-Jenkins

Os modelos SARIMA (*seasonal autoregressive integrated moving average*), introduzidos por Box & Jenkins (1970), permitem modelar e prever séries temporais estacionárias e não estacionárias, descrevendo a série Y_t como função dos seus valores passados e como combinação linear de uma sucessão de choques aleatórios.

Dentro dos modelos SARIMA, podem considerar-se vários modelos mais simples: o modelo autorregressivo (*autoregressive*, AR), que considera que o comportamento

da série pode ser explicado através do seu passado; o modelo de médias móveis (*moving average*, MA), que explica a série temporal através de uma sucessão de choques aleatórios; e ainda o modelo autorregressivo e de médias móveis (*autoregressive moving average*, ARMA), que, tal como o nome indica, se trata de uma combinação dos dois modelos anteriores. Estes modelos são úteis para séries estacionárias, revelando-se, no entanto, insuficientes para modelar casos de não estacionariedade. Nesses casos, deve optar-se pelos modelos integrados (ARIMA) ou, para séries que apresentam sazonalidade, pelos equivalentes sazonais (SARIMA).

A metodologia Box-Jenkins trata-se de um processo iterativo de três fases (identificação, estimação e diagnóstico), que facilita a escolha do modelo SARIMA que melhor descreve a série temporal em questão. Na primeira fase, tal como o nome indica, pretende-se identificar, através da análise da representação gráfica dos dados e das respetivas FAC e FACP empíricas, o modelo SARIMA mais apropriado. Os parâmetros do modelo escolhido são estimados na segunda fase (estimação) e avaliados quanto à sua significância na fase de diagnóstico. Também na última fase se avalia o comportamento dos resíduos, que se deve assemelhar a um ruído branco. Estas três fases são aplicadas iterativamente até que o modelo final não possa ser melhorado.

No entanto, a fase de identificação do modelo é bastante subjetiva e complexa, resultando, muitas vezes, em modelos distintos para a mesma análise. Para contornar esta situação, muitos *softwares* já possuem algoritmos automatizados de seleção, que permitem não só que a metodologia possa ser utilizada por leigos, mas também que os resultados obtidos por diferentes analistas sejam os mesmos.

A maior vantagem destes modelos é que, ao contrário do que acontece no alisamento exponencial e na regressão linear, estes têm em consideração uma das principais características dos dados de séries temporais: a dependência temporal, ou seja, a autocorrelação. Desta forma, os modelos SARIMA são apropriados quando se pode assumir que existe algum tipo de relação entre o passado e o futuro, sendo, no entanto, isso que os torna pouco recomendados para previsões a longo prazo e/ou previsões de séries com mudanças bruscas de comportamento.

Mais detalhes sobre estes modelos e a metodologia Box-Jenkins podem ser consultados em Box et al. (2016).

4.1.1 Modelos de processos estacionários

Uma série estacionária fica completamente definida pelas suas funções média, variância e de autocorrelação. Fazendo uso desta característica, pretende-se, com

a metodologia Box-Jenkins, identificar um modelo com base no comportamento da função de autocorrelação empírica.

Os processos ARMA são considerados como um grupo bastante diversificado e de grande fiabilidade na modelação de inúmeras séries temporais estacionárias. No entanto, para a modelação de séries que apresentem oscilações bruscas ao longo do tempo, este tipo de processos revela-se insuficiente. Ainda assim, os processos ARMA têm especial importância na modelação de séries não estacionárias, uma vez que estas são facilmente convertidas em estacionárias através de transformações adequadas.

Processo autorregressivo de ordem p , $AR(p)$

Os processos autorregressivos baseiam-se no pressuposto de que a observação da variável no instante t se relaciona, de forma linear, com as observações nos instantes anteriores. Assim, o processo Y_t diz-se um processo autorregressivo de ordem p , $AR(p)$, quando satisfaz a equação

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_{t-p} Y_{t-p} + \epsilon_t, \quad (4.1)$$

onde ϵ_t é um ruído branco de média nula, independente de Y_{t-k} para todo o $k \geq 1$. De facto, Y_t pode ser vista como uma variável dependente que é explicada através de uma regressão linear múltipla, em que as observações em p instantes anteriores funcionam como variáveis explicativas e ϕ_i são os coeficientes de cada Y_{t-i} .

Alternativamente, a representação de um processo $AR(p)$ pode ser feita através do operador atraso B^k , que se define como sendo $B^k Y_t = Y_{t-k}$. Com efeito, a equação (4.1) pode ser reescrita como

$$\Phi_p(B)Y_t = \epsilon_t, \quad (4.2)$$

onde $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ é o polinómio autorregressivo de ordem p . Tendo em consideração as p raízes (reais ou complexas), $G_1^{-1}, G_2^{-1}, \dots, G_p^{-1}$, da equação característica $\Phi_p(B) = 0$, torna-se possível fatorizar o polinómio autorregressivo do seguinte modo

$$\Phi_p(B) = \prod_{i=1}^p (1 - G_i B). \quad (4.3)$$

Para que o processo seja estacionário é condição necessária e suficiente que as raízes da equação característica sejam todas de módulo maior do que a unidade, ou, de forma equivalente, que $|G_i| < 1$, para $i = 1, 2, \dots, p$. Qualquer processo autorregressivo que seja estacionário é também invertível, o que, em termos práticos,

significa que a dependência do passado se vai atenuando à medida que o passado se torna mais remoto.

Se o processo Y_t é um processo $AR(p)$, então a sua função de autocorrelação parcial, ϕ_{kk} , é igual a zero para todo o $k > p$. Assim, a FACP de um processo $AR(p)$ apresenta, graficamente, uma queda brusca para zero a partir do *lag* $p + 1$, enquanto que a respetiva FAC tem um decaimento exponencial ou sinusoidal amortecido para zero. Na Figura 4.1 encontra-se representado um processo autorregressivo de ordem 2, $AR(2)$, e as respetivas FAC e FACP empíricas. Este processo é estacionário, uma vez que as raízes da equação característica $1 - 0,6B + 0,3B^2 = 0$ são as duas, em módulo, aproximadamente iguais a 1,83.

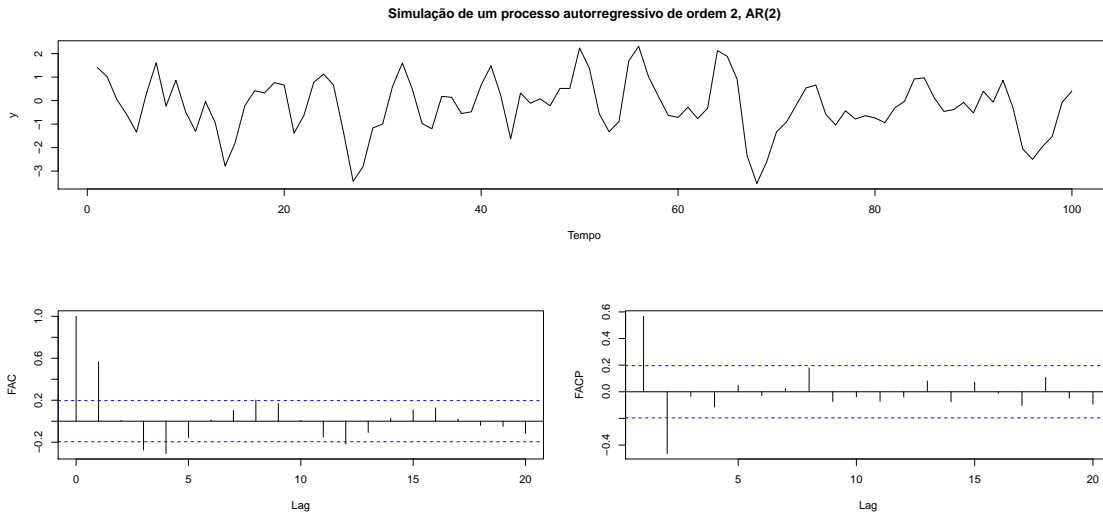


Figura 4.1: Simulação de um processo autorregressivo de equação $(1 - 0,6B + 0,3B^2)Y_t = \epsilon_t$, e respetivas FAC e FACP empíricas.

Processo de médias móveis de ordem q , $MA(q)$

O processo Y_t diz-se um processo de médias móveis de ordem q , $MA(q)$, quando assume a expressão

$$Y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \dots + \theta_q\epsilon_{t-q} \quad (4.4)$$

ou

$$Y_t = \Theta_q(B)\epsilon_t, \quad (4.5)$$

onde ϵ_t é um ruído branco de média nula e $\Theta_q(B) = 1 + \theta_1B + \theta_2B^2 + \dots + \theta_qB^q$ é o polinómio de médias móveis de ordem q . Pretende-se, através destes processos, exprimir Y_t em termos de um processo mais simples, como é o ruído branco. Assim,

um processo de médias móveis de ordem q define-se, em cada instante t , como a média ponderada das $q + 1$ observações de um processo de ruído branco. Desta forma, graças à estacionariedade inerente ao ruído branco, os processos de médias móveis são sempre estacionários. Adicionalmente, um processo de médias móveis é invertível se puder ser escrito como um processo autorregressivo estacionário de ordem infinita. Para garantir a invertibilidade do processo, basta que, à semelhança do que acontece no caso da estacionariedade de processos autorregressivos, as raízes da equação característica $\Theta_q(B) = 0$ se encontrem todas fora do círculo unitário, isto é, sejam, em módulo, todas superiores a 1 (Cowpertwait & Metcalfe, 2009).

Se o processo Y_t é um processo $MA(q)$, então a sua função de autocorrelação, ρ_k , é igual a zero para todo o $k > q$, e, por isso, a FAC de um processo $MA(q)$ apresenta, graficamente, uma queda brusca para zero a partir do *lag* $q + 1$. No que respeita à FACP, esta exhibe um decaimento exponencial ou sinusoidal amortecido para zero tendo, portanto, a mesma estrutura que a FAC de um processo $AR(q)$. Na Figura 4.2 encontra-se representado um processo de médias móveis de ordem 2, $MA(2)$, e as respetivas FAC e FACP empíricas. Este processo é invertível, uma vez que as raízes da equação característica $1 - 0,7B + 0,1B^2 = 0$ são iguais a 2 e a 5.

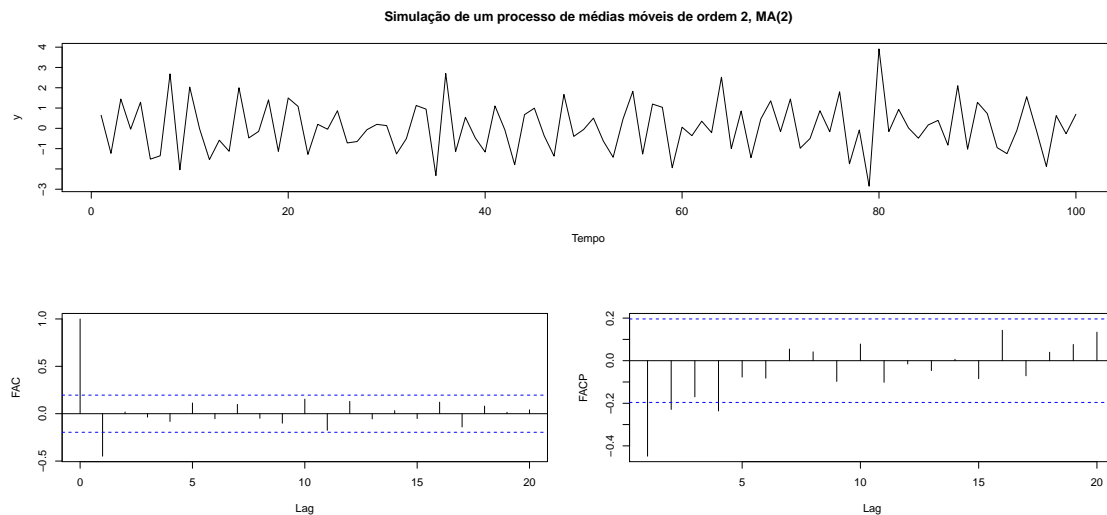


Figura 4.2: Simulação de um processo de médias móveis de equação $Y_t = (1 - 0,7B + 0,1B^2)\epsilon_t$, e respetivas FAC e FACP empíricas.

Processo autorregressivo e de médias móveis, $ARMA(p, q)$

Os processos estacionários e invertíveis podem ser representados quer na forma autorregressiva quer na forma de médias móveis. Contudo, pode acontecer que qualquer um destes processos tenha uma representação com um número excessivo

de parâmetros, o que pode conduzir a uma perda de eficiência na sua estimação (Caiado, 2016). Se assim for, pode construir-se um modelo mais parcimonioso que inclua tanto termos autorregressivos como de médias móveis. Este modelo designa-se de processo misto autorregressivo e de médias móveis de ordens p e q e representa-se por $\text{ARMA}(p, q)$.

Assim, o processo Y_t diz-se um processo autorregressivo e de médias móveis de ordens p e q , $\text{ARMA}(p, q)$, se satisfaz a equação

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} \quad (4.6)$$

ou a equação

$$\Phi_p(B)Y_t = \Theta_q(B)\epsilon_t, \quad (4.7)$$

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

A FAC e a FACP de um processo $\text{ARMA}(p, q)$ resultam da combinação das respetivas funções dos processos $\text{AR}(p)$ e $\text{MA}(q)$. Recorde-se que a FAC de um processo $\text{MA}(q)$ é insignificante a partir do lag $q + 1$, o mesmo acontecendo para a FACP de um processo $\text{AR}(p)$ depois do lag p . Na Figura 4.3 encontra-se representado um processo autorregressivo e de médias móveis, $\text{ARMA}(2, 2)$, e as respetivas FAC e FACP empíricas. Este processo é estacionário e invertível.

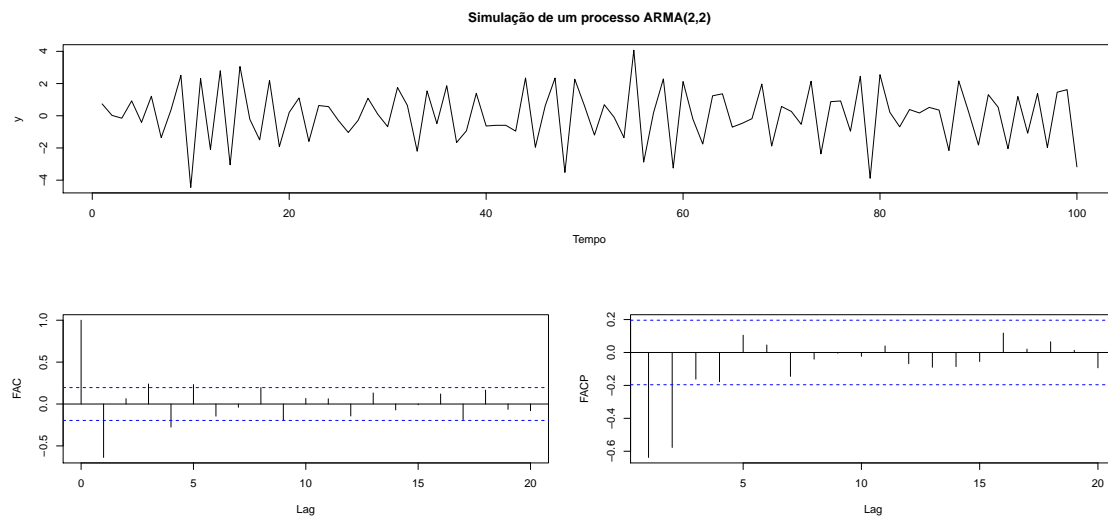


Figura 4.3: Simulação de um processo autorregressivo e de médias móveis, $\text{ARMA}(2, 2)$, de equação $(1 + 0,5B + 0,3B^2)Y_t = (1 - 0,7B + 0,1B^2)\epsilon_t$, e respetivas FAC e FACP empíricas.

De facto, uma vez que o processo $\text{ARMA}(p, q)$ é uma combinação dos processos $\text{AR}(p)$ e $\text{MA}(q)$, a estacionariedade e a invertibilidade do processo ficam garantidas se as raízes das equações características $\Phi_p(B) = 0$ e $\Theta_q(B) = 0$ são, em módulo, maiores do que a unidade. Como é evidente, estes processos generalizam os processos anteriormente mencionados e, por exemplo, um processo $\text{ARMA}(p, 0)$ é equivalente a um processo $\text{AR}(p)$, o mesmo acontecendo com um $\text{ARMA}(0, q)$ relativamente a um $\text{MA}(q)$.

Processo autorregressivo e de médias móveis sazonal, $\text{SARMA}(P, Q)_s$

Algumas séries temporais exibem, por vezes, uma componente sazonal ou uma componente periódica de período s , caracterizada pela existência de correlação significativa entre observações desfasadas em s unidades temporais, $Y_t, Y_{t-s}, Y_{t-2s}, \dots$. De forma a contemplar essa correlação, existem os processos sazonais, que permitem modelar séries que apresentem um comportamento periódico.

O processo autorregressivo e de médias móveis sazonal (*seasonal autoregressive moving average*, SARMA) de ordens P e Q , $\text{SARMA}(P, Q)_s$, define-se através da equação

$$Y_t = \nu_1 Y_{t-s} + \dots + \nu_P Y_{t-Ps} + \epsilon_t + \eta_1 \epsilon_{t-s} + \dots + \eta_Q \epsilon_{t-Qs} \quad (4.8)$$

ou

$$N_P(B^s)Y_t = H_Q(B^s)\epsilon_t, \quad (4.9)$$

onde ϵ_t é um ruído branco de média nula, independente de Y_{t-ks} para todo o $k \geq 1$, $N_P(B^s) = 1 - \nu_1 B^s - \dots - \nu_P B^{Ps}$ e $H_Q(B^s) = 1 + \eta_1 B^s + \dots + \eta_Q B^{Qs}$ são os polinómios autorregressivo e de médias móveis sazonais de ordens P e Q , respetivamente. Tanto a FAC como a FACP do modelo $\text{SARMA}(P, Q)_s$ apresentam um comportamento do tipo exponencial ou sinusoidal amortecido sobre os *lags* múltiplos de s , mantendo-se nulas nos restantes *lags*. Na Figura 4.4 encontra-se representado um processo autorregressivo e de médias móveis sazonal de período 12, $\text{SARMA}(1, 1)_{12}$, e as respetivas FAC e FACP empíricas.

Os processos autorregressivos sazonais (*seasonal autoregressive*, SAR) de ordem P e os processos de médias móveis sazonais (*seasonal moving average*, SMA) de ordem Q são casos particulares dos processos $\text{SARMA}(P, Q)_s$, que se designam por $\text{SAR}(P)_s$ e $\text{SMA}(Q)_s$, respetivamente. A FAC do processo $\text{SAR}(P)_s$ tem um decaimento exponencial ou sinusoidal amortecido sobre os *lags* múltiplos de s , anulando-se nos restantes, e a FACP tem uma queda brusca para zero a partir do P -ésimo *lag* múltiplo de s . Já a FAC do $\text{SMA}(Q)_s$ apresenta uma queda brusca para zero a

partir do Q -ésimo *lag* múltiplo de s , enquanto que a FACP tem uma queda lenta para zero sobre os *lags* múltiplos de s (Caiado, 2016).

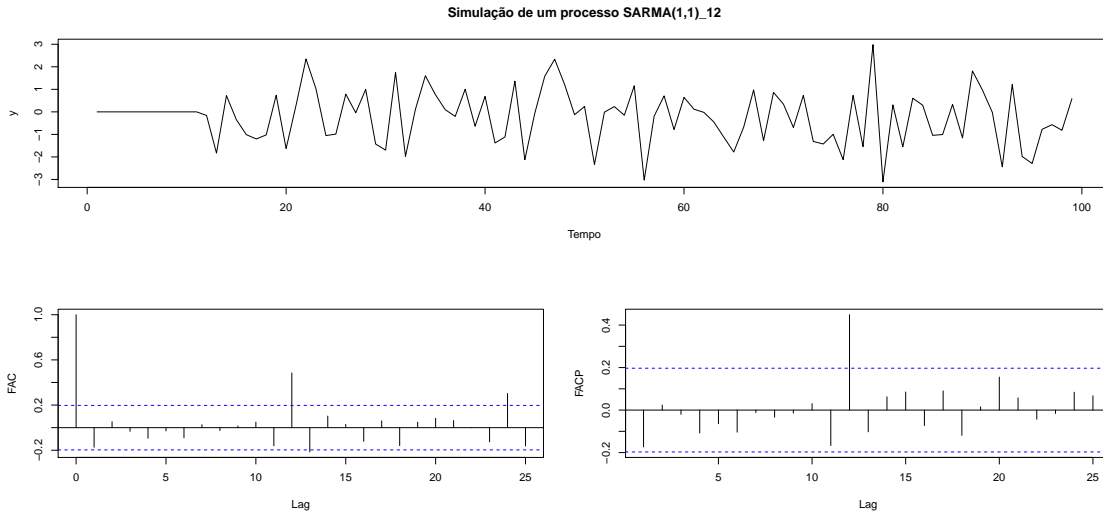


Figura 4.4: Simulação de um processo autorregressivo e de médias móveis sazonal, $SARMA(1, 1)_{12}$, de equação $(1 - 0,65B^{12})Y_t = (1 - 0,25B^{12})\epsilon_t$, e respetivas FAC e FACP empíricas.

Além disso, a combinação dos processos $SARMA(P, Q)_s$ com os processos $ARMA(p, q)$ permite obter o processo multiplicativo, com componente sazonal e não sazonal, $SARMA(p, q)(P, Q)_s$, com a equação

$$\Phi_p(B)N_P(B^s)Y_t = \Theta_q(B)H_Q(B^s)\epsilon_t, \quad (4.10)$$

onde $\Phi_p(B)$ e $N_P(B^s)$ representam os polinómios autorregressivos regular (ou não sazonal) e sazonal, $\Theta_q(B)$ e $H_Q(B^s)$ representam os polinómios de médias móveis regular e sazonal, todos iguais aos anteriormente descritos. Quando as raízes destes quatro polinómios se encontram fora do círculo unitário ficam garantidas a estacionariedade e a invertibilidade.

Os comportamentos das FAC e FACP destes processos resultam também da combinação das características das respetivas funções para os processos $ARMA(p, q)$ e $SARMA(P, Q)_s$. Por exemplo, num processo $SARMA(1, 0)(1, 0)_{12}$, a FAC pode apresentar um decaimento sinusoidal amortecido para zero sobre os *lags* não sazonais e sazonais e a FACP apresentar autocorrelações significativas nos *lags* 1 e 12 e quedas bruscas para zero nos *lags* não sazonais e sazonais seguintes, como é possível observar na Figura 4.5 (Caiado, 2016).

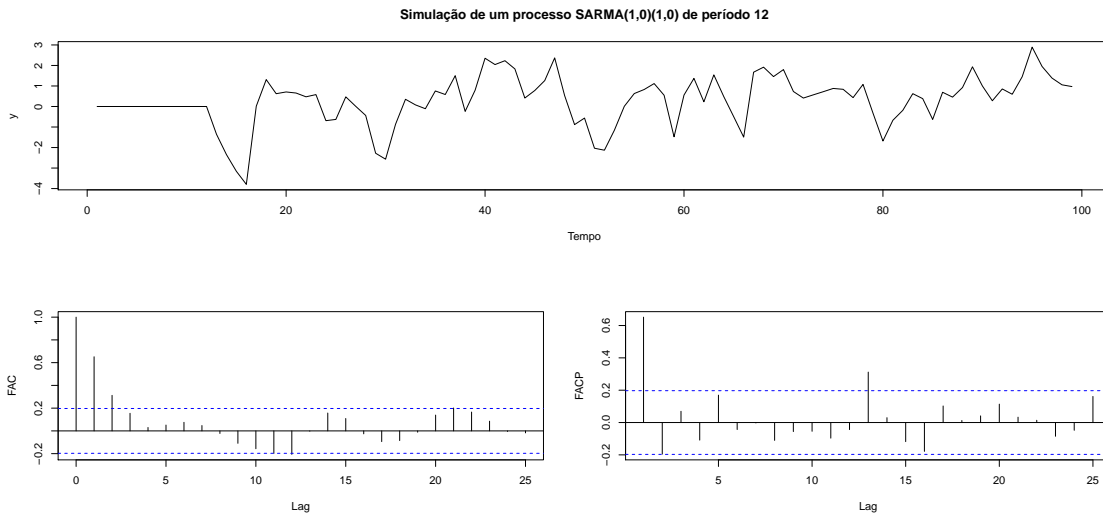


Figura 4.5: Simulação de um processo multiplicativo, $SARMA(1,0)(1,0)_{12}$, de equação $(1 - 0,7B)(1 + 0,25B^{12})Y_t = \epsilon_t$, e respectivas FAC e FACP empíricas.

4.1.2 Modelos de processos não estacionários

Na prática, a maioria das séries temporais é não estacionária. Quando isso acontece, é necessário remover dos dados as fontes de variação não estacionárias (e.g., tendência, sazonalidade), de forma a possibilitar o ajustamento de um modelo estacionário. Como visto na Secção 3.2.2, se a série temporal observada for não estacionária na média, pode aplicar-se uma (ou várias) diferenciação (regular) à mesma. Assim, se se substituir Y_t por $\nabla^d Y_t$ na equação (4.7), obtém-se um modelo capaz de descrever séries não estacionárias (modelo ARIMA). Este tipo de modelo é designado de modelo “integrado”, uma vez que o modelo estacionário que é ajustado aos dados diferenciados deve ser somado ou “integrado” para fornecer um modelo para os dados não estacionários. Além disso, estes modelos podem, à semelhança dos modelos ARMA, ser generalizados para incluir termos sazonais, dando origem aos modelos SARIMA.

Processo autorregressivo e de médias móveis integrado, $ARIMA(p, d, q)$

O processo Y_t diz-se um processo autorregressivo e de médias móveis integrado, $ARIMA(p, d, q)$, quando assume a expressão

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - B)^d Y_t = (1 + \theta_1 B + \dots + \theta_q B^q) \epsilon_t \quad (4.11)$$

ou

$$\Phi_p(B)\nabla^d Y_t = \Theta_q(B)\epsilon_t, \quad (4.12)$$

onde $\nabla^d Y_t = (1 - B)^d Y_t$, com $d \geq 1$, é a série estacionária depois de diferenciada d vezes, $\phi_1, \phi_2, \dots, \phi_p$ são os parâmetros autorregressivos, $\theta_1, \theta_2, \dots, \theta_q$ são os parâmetros de médias móveis e $\Phi_p(B)$ e $\Theta_q(B)$ são os polinômios autorregressivo e de médias móveis regulares.

Por se tratar de um processo não estacionário, um processo deste tipo apresenta uma FAC com coeficientes positivos e decaimento muito lento para zero, pelo que a necessidade de uma diferenciação (regular) é facilmente identificável. Na Figura 4.6 encontra-se representado um processo autorregressivo e de médias móveis integrado, ARIMA(2, 1, 1), e as respectivas FAC e FACP empíricas.

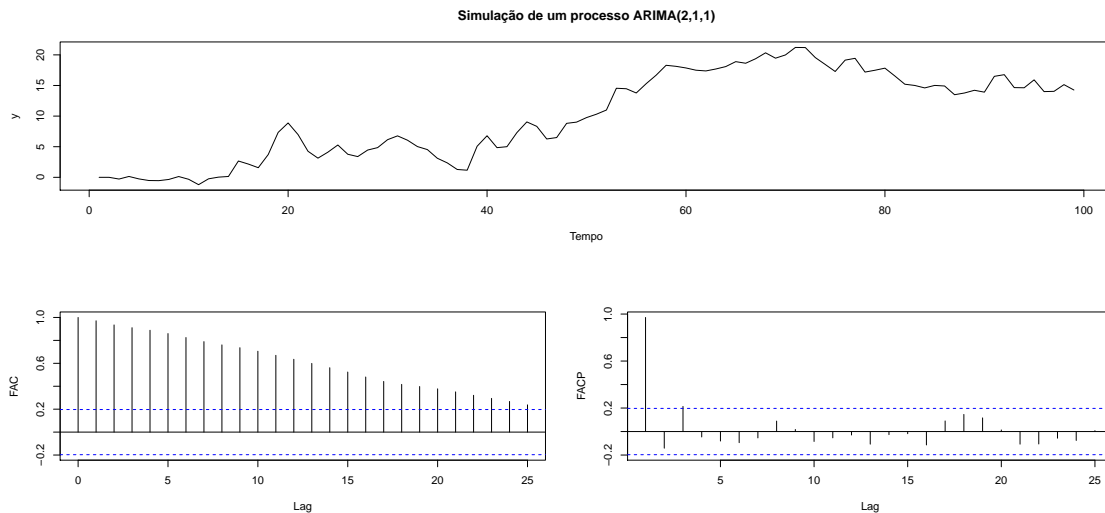


Figura 4.6: Simulação de um processo autorregressivo e de médias móveis integrado, ARIMA(2, 1, 1), de equação $(1 + 0,2B + 0,3B^2)(1 - B)Y_t = (1 + 0,7B)\epsilon_t$, e respectivas FAC e FACP empíricas.

Processo autorregressivo e de médias móveis integrado sazonal, SARIMA(p, d, q)(P, D, Q)_s

Em muitas séries temporais sazonais é de esperar que a componente sazonal esteja de alguma forma relacionada com as componentes não sazonais. Isto é, se as observações vizinhas de uma série, $Y_t, Y_{t-1}, Y_{t-2}, \dots$, estão relacionadas, é muito provável que as observações vizinhas espaçadas em s unidades temporais, $Y_t, Y_{t-s}, Y_{t-2s}, \dots$, também estejam relacionadas. Assim sendo, pode estender-se o processo ARIMA(p, d, q) a um processo multiplicativo integrado sazonal, que generaliza todos os processos apresentados anteriormente e se representa por

SARIMA(p, d, q)(P, D, Q) $_s$.

Antes de se introduzir a expressão do modelo completo, é conveniente apresentar-se a noção de diferenciação sazonal. Uma diferenciação sazonal não é mais do que a diferença entre a observação no instante t e a observação que ocorre s momentos antes, ou seja, a observação no instante $t - s$. Desta forma, quando uma série apresenta um comportamento periódico repetitivo, pode-se aplicar uma diferenciação sazonal, isto é,

$$\nabla_s Y_t = Y_t - Y_{t-s} = (1 - B^s)Y_t.$$

A série resultante desta diferenciação corresponde, então, à mudança entre observações separadas por períodos de tempo s . Por exemplo, para uma série mensal, com $s = 12$, a série resultante de uma diferenciação sazonal representa a mudança que ocorre de ano para ano.

À semelhança do que acontece para a diferenciação (regular) introduzida na Secção 3.2.2, a diferenciação sazonal pode ser aplicada a uma série D vezes, dando origem ao operador de diferenciação sazonal de ordem D , para qualquer inteiro $D \geq 1$, que se define por

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

Assim, um processo Y_t diz-se um processo autorregressivo e de médias móveis integrado sazonal, SARIMA(p, d, q)(P, D, Q) $_s$, quando satisfaz a equação

$$\Phi_p(B)N_P(B^s)\nabla^d\nabla_s^D Y_t = \Theta_q(B)H_Q(B^s)\epsilon_t, \quad (4.13)$$

em que $\Phi_p(B)$, $N_P(B^s)$, $\Theta_q(B)$ e $H_Q(B^s)$ são os polinómios já referidos, d e D são as ordens de diferenciação das partes regular e sazonal, respetivamente.

Em geral, a necessidade de uma diferenciação sazonal pode ser indicada quando a FAC de um processo decai lentamente nos lags múltiplos de s e é insignificante nos restantes (Shumway & Stoffer, 2006). Na Figura 4.7 encontra-se representado um processo SARIMA(2, 1, 1)(1, 1, 1) $_{12}$ e as respetivas FAC e FACP empíricas.

Na maioria das aplicações práticas, os valores de p, q, P e Q são quase sempre inferiores ou iguais a 2, enquanto que os valores de d e D usualmente assumem os valores inteiros 0 ou 1 (Caiado, 2016).

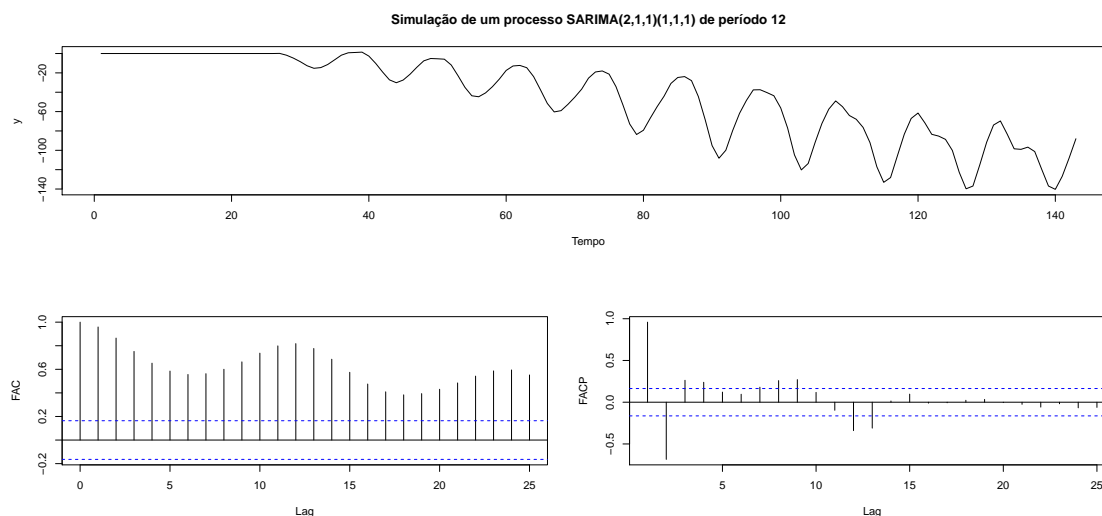


Figura 4.7: Simulação de um processo autorregressivo e de médias móveis integrado sazonal, SARIMA(2, 1, 1)(1, 1, 1)₁₂, de equação $(1 - 1,2B + 0,8B^2)(1 - 0,3B^{12})(1 - B)(1 - B^{12})Y_t = (1 + 0,4B)(1 + 0,7B^{12})\epsilon_t$, e respectivas FAC e FACP empíricas.

4.1.3 Metodologia Box-Jenkins

Baseados nos trabalhos de Yule (1926) e Wold (1938), Box & Jenkins (1970) introduziram uma abordagem prática e sistemática para a construção de modelos SARIMA, a conhecida metodologia Box-Jenkins. Esta metodologia não é mais do que um processo de modelação iterativo em três fases: identificação do modelo, estimação dos parâmetros e análise de diagnóstico (ou validação do modelo). Este processo é tipicamente repetido várias vezes até que um modelo satisfatório seja selecionado.

A ideia base por detrás da identificação do modelo é que, se uma série temporal é gerada a partir de um processo SARIMA, então deve ter algumas propriedades teóricas de autocorrelação. Assim, ao comparar os padrões empíricos de autocorrelação com os teóricos, é frequentemente possível identificar um ou vários potenciais modelos para a série temporal a estudar. Box & Jenkins (1970) propuseram, então, usar a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) como ferramentas básicas para identificar as ordens do modelo SARIMA (Zhang, 2003).

Descrevem-se, de seguida, as três fases da metodologia Box-Jenkins, esquematizadas na Figura 4.8: identificação, estimação e diagnóstico.

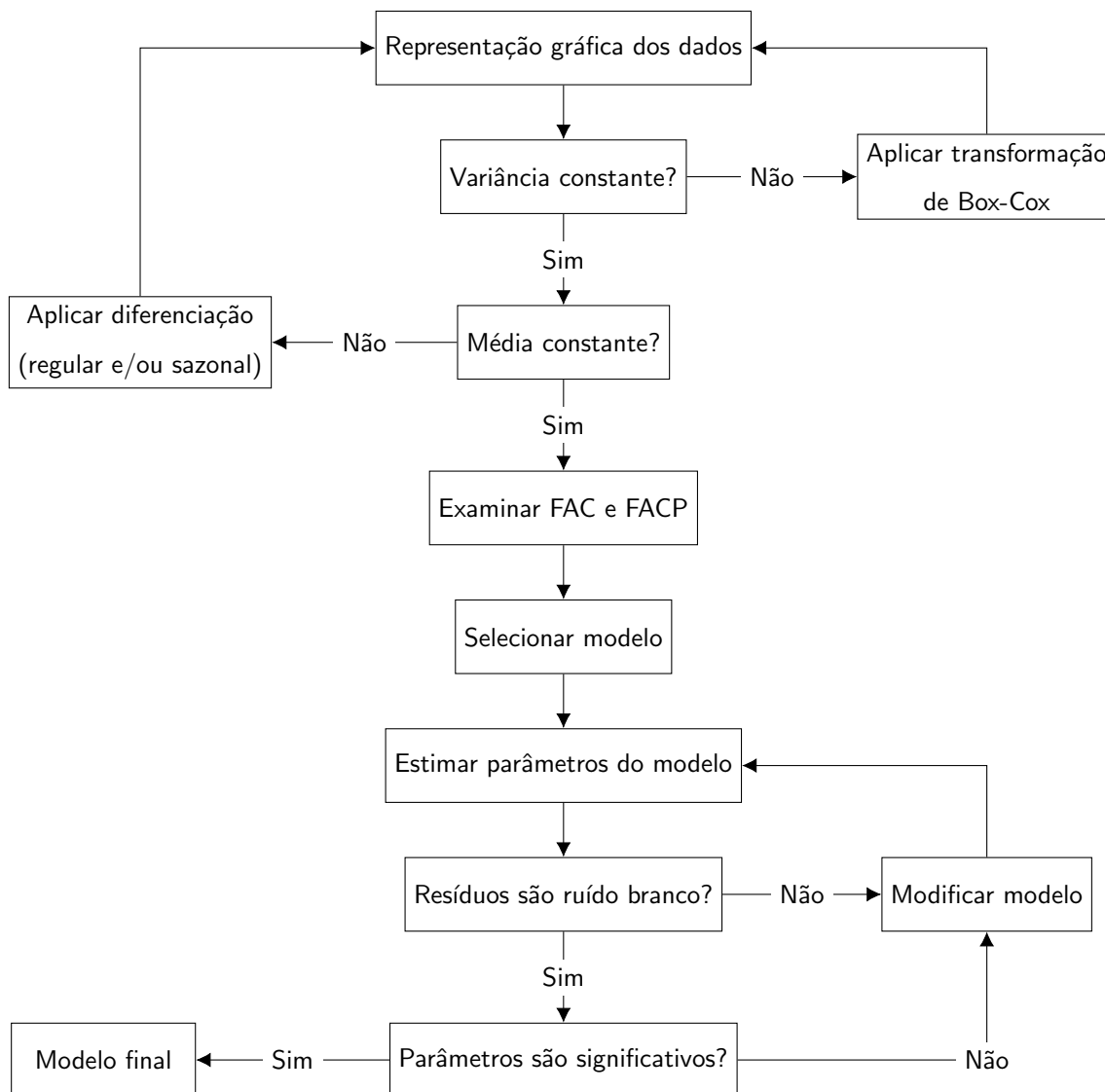


Figura 4.8: Esquema da metodologia Box-Jenkins.

Identificação

A primeira etapa de modelação de uma série temporal consiste na identificação de um modelo $SARIMA(p, d, q)(P, D, Q)_s$ que descreva a relação existente entre as suas observações. Esta etapa compreende três importantes passos na tentativa de identificação dos valores dos inteiros d, s, D, p, q, P e Q :

Passo 1: Representação gráfica da série e sua estacionarização

O estudo de uma qualquer série temporal deve começar pela análise detalhada da sua representação gráfica, a fim de se poder identificar a existência ou não de fontes de não estacionariedade. No caso da série ser não estacionária,

deve, então, proceder-se à sua estacionarização através das transformações adequadas: para a estabilização da variância recorrer a transformações Box-Cox; para a eliminação da tendência usar a diferenciação regular; e para a eliminação de movimentos periódicos, a diferenciação sazonal. Recorde-se que, se a estabilização da variância for necessária, esta deve ser realizada antes de qualquer outra transformação.

Passo 2: Estimação das FAC e FACP da série original

Analisar o comportamento das FAC e FACP da série original pode ser útil para complementar a informação obtida através da representação gráfica (passo 1), nomeadamente no que diz respeito à utilização de diferenciações. Por exemplo, um decaimento lento para zero na FAC de uma série temporal pode indicar a necessidade da aplicação de uma diferenciação não sazonal (regular).

Passo 3: Estimação das FAC e FACP da série estacionária e identificação dos inteiros p, q, P e Q

Depois de se identificarem as ordens de diferenciação, d e D , e o período, s , a escolha do modelo que descreve a série temporal só fica concluída quando são determinados os inteiros p, q, P e Q . Uma das formas de identificar esses inteiros é através da comparação do comportamento das FAC e FACP empíricas com o das FAC e FACP teóricas dos modelos $ARMA(p, q)$ e $SARMA(p, q)(P, Q)_s$. A Tabela 4.1 resume os padrões teóricos apresentados pelas FAC e FACP dos modelos AR, MA, ARMA, SAR, SMA, SARMA e SARMA multiplicativo.

Estimação

Uma vez identificados os modelos candidatos a descrever a série em estudo, segue-se a etapa de estimação dos seus parâmetros. Nesta fase, é imprescindível o auxílio de um *software* estatístico adequado, dado que a estimação dos parâmetros requer a aplicação de um conjunto de métodos numéricos e de cálculos computacionais com alguma complexidade. Os dois principais métodos de estimação dos parâmetros do modelo SARIMA são o método da máxima verosimilhança e o método dos mínimos quadrados.

A ideia base do método da máxima verosimilhança consiste em determinar os valores dos parâmetros que tornam mais verosímil a ocorrência de um conjunto de observações idênticas aquelas de que efetivamente se dispõe. Este método obtém estimativas dos parâmetros através de um processo iterativo em que se maximiza a função de verosimilhança dos estimadores (mais detalhes em Box et al. (2016)).

O método dos mínimos quadrados é, provavelmente, o método estatístico mais utilizado na estimação de modelos. Contudo, em modelos $ARMA(p, q)$ este método não permite obter estimadores consistentes com os verdadeiros parâmetros. A única exceção ocorre nos casos em que os modelos têm apenas componente autorregressiva (modelos $AR(p)$), em que os estimadores se podem obter por minimização da soma de quadrados dos resíduos (Caiado, 2016).

Tabela 4.1: Padrões teóricos das FAC e FACP dos modelos $SARMA(p, q)(P, Q)_s$.

Modelo	FAC	FACP
$AR(p)$	Decaimento exponencial ou sinusoidal amortecido para zero	Queda brusca para zero a partir do $lag\ p + 1$
$MA(q)$	Queda brusca para zero a partir do $lag\ q + 1$	Decaimento exponencial ou sinusoidal amortecido para zero
$ARMA(p, q)$	Decaimento exponencial ou sinusoidal amortecido para zero	Decaimento exponencial ou sinusoidal amortecido para zero
$SAR(P)_s$	Decaimento exponencial ou sinusoidal amortecido para zero sobre os $lags\ s, 2s, \dots$	Queda brusca para zero a partir do $lag\ (P + 1)s$
$SMA(Q)_s$	Queda brusca para zero a partir do $lag\ (Q + 1)s$	Decaimento exponencial ou sinusoidal amortecido para zero sobre os $lags\ s, 2s, \dots$
$SARMA(P, Q)_s$	Decaimento exponencial ou sinusoidal amortecido para zero sobre os $lags\ s, 2s, \dots$	Decaimento exponencial ou sinusoidal amortecido para zero sobre os $lags\ s, 2s, \dots$
$SARMA(p, q)(P, Q)_s$	Decaimento exponencial ou sinusoidal amortecido para zero	Decaimento exponencial ou sinusoidal amortecido para zero

Diagnóstico

Após se identificar o modelo SARIMA e estimar os respetivos parâmetros, é necessário verificar se o modelo é realmente adequado. A fase de diagnóstico engloba duas importantes etapas: a avaliação da qualidade das estimativas obtidas e a avaliação da qualidade do ajustamento do modelo às observações da série em estudo.

Na avaliação da qualidade das estimativas obtidas interessa, sobretudo, analisar a significância estatística dos parâmetros estimados. Para tal, a cada parâmetro,

diga-se β_i (onde $i = 1, \dots, m$ e m representa o número de parâmetros estimados), deve aplicar-se um teste de hipóteses que avalie a necessidade (ou não) de incluir esse parâmetro no modelo. Interessa, assim, testar a hipótese de que β_i é estatisticamente nulo, isto é, $H_0 : \beta_i = 0$. A rejeição desta hipótese acontece, a um nível de significância α , quando a estatística T associada ao coeficiente estimado for, em valor absoluto, superior ou igual ao quantil $1 - \alpha/2$ de uma distribuição t -Student com $n - m$ graus de liberdade, (onde n representa o número de observações), isto é,

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

De forma equivalente, a hipótese nula é rejeitada, a um nível de significância α , quando o valor de prova obtido é inferior ou igual a α .

Tendo sempre presente o princípio da parcimónia, devem incluir-se no modelo apenas os parâmetros que possam considerar-se significativamente diferentes de zero (Caiado, 2016).

No que diz respeito à avaliação da qualidade do ajustamento do modelo SARIMA, deve proceder-se à análise do comportamento dos respetivos resíduos (mais detalhes na Secção 4.1.4). Se os resíduos tiverem comportamento semelhante a um ruído branco, pode dizer-se que o modelo estimado descreve bem a série em estudo.

Um modelo que não satisfaça os critérios aplicados em alguma das duas etapas, deve ser rejeitado.

4.1.4 Análise dos resíduos

Após a modelação de uma série temporal, os resíduos, que correspondem à informação não captada pelo modelo, podem ser calculados através da diferença entre os valores observados e os valores estimados correspondentes, ou seja,

$$e_t = Y_t - \hat{Y}_t.$$

Como referido anteriormente, um bom modelo deve gerar resíduos com o comportamento idêntico ao de um ruído branco, e, portanto, estes devem apresentar média nula e satisfazer o pressuposto da não correlação. Adicionalmente, para a construção de intervalos de previsão, torna-se pertinente verificar se os resíduos têm variância constante e apresentam uma distribuição aproximadamente Normal (Hyndman & Athanasopoulos, 2014).

Distribuição Normal

A condição de normalidade pode ser avaliada quer por análise gráfica, quer por testes estatísticos, ou, idealmente, por ambos.

No caso das representações gráficas, as mais usuais são o histograma e o *QQ-plot*. Para indicar a normalidade, o histograma deve aproximar-se do comportamento da função densidade de uma distribuição Normal. No que respeita ao *QQ-plot*, uma vez que se trata de uma representação gráfica dos quantis reais e dos teóricos, este deve apresentar um conjunto de pontos que se posicione mais ou menos sobre uma reta.

Para uma verificação rigorosa, os testes estatísticos mais comuns são o teste de Shapiro-Wilk (para amostras de pequenas dimensões, com menos de 50 observações) e o teste de Kolmogorov-Smirnov. Em ambos é testada a hipótese nula “os erros seguem uma distribuição Normal”.

Não correlação

Relativamente ao pressuposto de não correlação, este deve ser verificado tanto individualmente como de forma conjunta.

Individualmente, as autocorrelações dos resíduos podem ser avaliadas através da observação da FAC que, se o modelo for apropriado, deve apresentar um comportamento semelhante ao da FAC de um ruído branco, ou seja, com autocorrelações não significativamente diferentes de zero.

Para testar várias autocorrelações como um grupo pode recorrer-se a um teste de Portmanteau. Um dos mais utilizados dentro desta classe de testes é o de Ljung-Box, cuja estatística de teste Q é definida por

$$Q = n(n + 2) \sum_{j=1}^k \frac{\hat{\rho}_j^2}{n - j}$$

e segue aproximadamente uma distribuição do Qui-Quadrado com $k - m$ graus de liberdade (com k a corresponder ao número de autocorrelações a serem testadas e m ao número de parâmetros estimados). Em caso de rejeição da hipótese nula, $H_0 : \rho_1 = \rho_2 = \dots = \rho_k = 0$, conclui-se que o modelo escolhido não é apropriado. Quanto ao valor de k , não existe um critério específico para a sua escolha, pelo que se compreende que a melhor abordagem passa por realizar o teste para vários valores diferentes.

Média nula e variância constante

A condição exigida à média dos resíduos pode ser facilmente verificada através de um teste estatístico muito comum: o teste t para o valor médio. Neste teste, rejeita-se a hipótese da média ser nula se

$$\left| \frac{\bar{e}\sqrt{n}}{s_e} \right| \geq t_{1-\alpha/2;n-1},$$

onde \bar{e} corresponde à média dos resíduos, n à dimensão da amostra, s_e ao desvio padrão dos resíduos e $t_{1-\alpha/2;n-1}$ ao valor do quantil $1 - \alpha/2$ de uma distribuição t -Student com $n - 1$ graus de liberdade. De forma equivalente, a hipótese de média nula é rejeitada, a um nível de significância α , quando o valor de prova obtido é inferior ou igual a α . No entanto, este teste só deve ser aplicado quando os pressupostos de normalidade e não correlação (independência, no caso de não se rejeitar a normalidade) se verificam.

A estabilidade da variância ou homocedasticidade pode ser avaliada, visualmente, através da análise do gráfico dos resíduos ao longo do tempo.

4.1.5 Seleção de modelos

Em qualquer análise estatística de dados, podem encontrar-se vários modelos adequados para descrever o fenómeno em estudo. Mais concretamente, na modelação de uma série temporal pode existir mais do que um modelo que verifique os diferentes critérios de avaliação do diagnóstico, o que torna difícil a tarefa de escolher o melhor modelo. Assim sendo, devem procurar-se critérios de seleção de modelos que tomem em consideração as estatísticas baseadas nos resíduos do modelo ajustado.

Um critério plausível para escolher o melhor modelo SARIMA poderia ser escolher o modelo que fornece a menor soma dos quadrados dos erros (ou erro quadrático médio) ou o maior valor para a função de verosimilhança. No entanto, esta abordagem nem sempre funciona porque, muitas vezes, o erro quadrático médio pode ser reduzido e a função de verosimilhança aumentada simplesmente pelo aumento do número de parâmetros no modelo.

Para solucionar esta questão, a função de verosimilhança deve ser penalizada por cada parâmetro adicional no modelo, ou seja, se o parâmetro extra não melhorar o valor da função de verosimilhança mais do que o valor da penalização, esse parâmetro não deve ser acrescentado ao modelo (Makridakis et al., 1998).

São vários os critérios, baseados na função de verosimilhança, existentes na lite-

ratura, sendo os mais utilizados o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (*Bayesian information criterion*, BIC).

Critério de informação de Akaike

Considere-se que um modelo com $m = p + q + P + Q$ parâmetros foi ajustado a uma série com n observações. Para avaliar a qualidade do ajustamento, Akaike (1974) introduziu um critério baseado na quantidade de informação, definido por

$$\text{AIC} = -2 \log L + 2m, \quad (4.14)$$

onde L é a função de verosimilhança.

Nem todos os *softwares* estatísticos possuem a capacidade de determinar o AIC ou a função de verosimilhança L e, por isso, nem sempre é possível encontrar o AIC exato para um determinado modelo. No entanto, uma aproximação útil para o AIC é obtida através da aproximação

$$-2 \log L \approx n(1 + \log 2\pi) + n \log \sigma^2,$$

onde σ^2 representa a variância dos resíduos. Esta variância é facilmente estimada por qualquer *software* estatístico, permitindo, assim, que o AIC possa ser encontrado aproximadamente através da fórmula

$$\text{AIC} \approx n(1 + \log 2\pi) + n \log \hat{\sigma}^2 + 2m. \quad (4.15)$$

Às vezes, o primeiro termo em (4.15) é omitido por ser igual para todos os modelos.

Note-se que o AIC não tem muito significado por si só e, por isso, só é útil em comparação com o AIC de outro modelo ajustado ao mesmo conjunto de dados. Assim, deve escolher-se o modelo que tenha o menor AIC, considerando que uma diferença de valores de duas unidades (2) ou menos não é substancial. Nesses casos, deve optar-se pelo modelo mais simples, seja pela parcimónia, ou para obter um melhor ajustamento do modelo.

Critério de informação Bayesiano

O critério de informação Bayesiano, proposto por Schwarz (1978), define-se como

$$\text{BIC} = -2 \log L + m \log(n), \quad (4.16)$$

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

Ao contrário do AIC, o BIC depende da dimensão da amostra (n) pelo que, para $\log(n) > 2$, isto é, para uma amostra de dimensão superior a 7, a penalização do BIC é superior à penalização do AIC. Consequentemente, a minimização do BIC leva, em geral, à seleção de modelos com um menor número de parâmetros do que os obtidos pela minimização do critério AIC, evitando, de certa forma, a sobrestimação do número de componentes.

4.1.6 Previsão pontual e intervalar

Depois de se escolher o modelo que melhor descreve a série temporal, pode prosseguir-se para o cálculo de previsões, sejam estas pontuais ou intervalares.

As previsões pontuais podem ser facilmente determinadas fazendo uso da própria expressão do modelo escolhido. De facto, para obter previsões a h -passos, isto é, para um instante $t + h$, basta calcular a esperança condicionada aos valores observados, ou seja, $E[Y_{t+h}|Y_1, Y_2, \dots, Y_t]$. Hyndman & Athanasopoulos (2014) descrevem este processo em apenas três passos:

1. Expandir a equação do modelo de forma a obter Y_t separado dos restantes termos (ou seja, Y_t num dos membros da equação e os restantes termos no outro);
2. Reescrever a equação substituindo t por $t + h$ (considera-se que o modelo é estimado usando as t observações);
3. Substituir as observações futuras pelas respetivas previsões, os erros futuros por zero e os erros passados pelos resíduos correspondentes.

Iniciando o processo para uma previsão a 1-passo, isto é, para $h = 1$, e repetindo-o para $h = 2, 3, \dots$, é, então, possível obter todas as previsões pretendidas.

Os intervalos de previsão usuais são construídos com base em estimativas do desvio padrão das próprias previsões. Partindo, assim, do pressuposto que os erros são independentes e seguem uma distribuição Normal, a previsão intervalar para o instante $t + h$ é dada por

$$\left(\hat{y}_{t+h|t} - z_{1-\alpha/2} \hat{\sigma}_h, \hat{y}_{t+h|t} + z_{1-\alpha/2} \hat{\sigma}_h \right),$$

onde z é o quantil da distribuição Normal padrão, $1 - \alpha$ corresponde ao nível de confiança do intervalo e $\hat{\sigma}_h$ é a estimativa do desvio padrão da previsão para o passo h .

Os níveis de confiança mais utilizados em intervalos de previsão são 90%, 95% e 99%, aos quais correspondem valores de z iguais a 1,645, 1,960 e 2,576, respectivamente.

Em geral, os intervalos de previsão aumentam conforme o horizonte de previsão, h , aumenta. No entanto, em modelos estacionários (isto é, com $d = 0$) as sucessões dos limites inferiores e superiores são convergentes e, portanto, para horizontes distantes, os intervalos de previsão terão amplitudes idênticas (Hyndman & Athanasopoulos, 2014).

Mais detalhes sobre estes e outros processos para a obtenção de previsões, tanto pontuais como intervalares, podem ser consultados em Box et al. (2016).

4.2 Métodos de alisamento exponencial

Os métodos de alisamento exponencial integram um conjunto de métodos de previsão adaptativos, i.e., capazes de acompanhar mudanças no nível, declive e/ou padrão sazonal e, por isso, são considerados dos mais eficientes quando aplicados a séries cujas componentes, tendência e sazonalidade, variam com o tempo. Nestes métodos, as previsões não são mais do que uma combinação ponderada das observações anteriores, onde às observações mais recentes é atribuído um “peso” maior do que às mais antigas. Por outras palavras, quanto mais recente for a observação, maior é a sua contribuição na previsão e, portanto, pode dizer-se que a ponderação decresce exponencialmente com a antiguidade das observações (daí a designação de *métodos de alisamento exponencial*).

Os primeiros desenvolvimentos destes métodos surgem no final da década de 50 do século XX. A ideia parece ser introduzida por Robert Brown, em 1944, enquanto trabalhava para a Marinha dos Estados Unidos da América como analista de Investição Operacional. Brown é contratado para desenvolver um modelo matemático de controlo de incêndios em submarinos, que se mostra, essencialmente, um alisamento exponencial simples de dados contínuos. Durante essa década, ele estende o método para séries discretas e desenvolve métodos que permitem lidar com tendência e sazonalidade (descritos em Brown (1959, 1963)).

Também durante a mesma década, Charles Holt desenvolve um método de alisamento exponencial, diferente do de Brown na forma de lidar com as componentes de tendência e de sazonalidade. Os métodos sazonais propostos por Holt (1957) são, posteriormente, testados por Winters (1960) para dados empíricos, daí a atual designação de métodos de Holt-Winters.

De acordo com as diferentes formas que a tendência e a sazonalidade podem

assumir (nula, aditiva ou multiplicativa), Pegels (1969) constrói, então, uma primeira classificação dos métodos de alisamento exponencial (Tabela 4.2).

Tabela 4.2: Classificação dos métodos de alisamento exponencial proposta por Pegels (1969).

Tendência	Sazonalidade		
	N (Nula)	A (Aditiva)	M (Multiplicativa)
N (Nula)	N,N	N,A	N,M
A (Aditiva)	A,N	A,A	A,M
M (Multiplicativa)	M,N	M,A	M,M

Mais tarde, e como consequência dos trabalhos de Gardner & McKenzie (1985), Hyndman et al. (2002) e Taylor (2003), surgem novos métodos de alisamento exponencial, que levam ao alargamento da classificação inicial proposta por Pegels. O aparecimento de uma componente “amortecida” resulta na introdução de dois novos tipos de tendência, que culminam numa classificação com 15 métodos (ignorando a componente erro, Tabela 4.3).

Tabela 4.3: Classificação atual dos métodos de alisamento exponencial.

Tendência	Sazonalidade		
	N (Nula)	A (Aditiva)	M (Multiplicativa)
N (Nula)	N,N	N,A	N,M
A (Aditiva)	A,N	A,A	A,M
Ad (Aditiva amortecida)	Ad,N	Ad,A	Ad,M
M (Multiplicativa)	M,N	M,A	M,M
Md (Multiplicativa amortecida)	Md,N	Md,A	Md,M

Alguns dos métodos referidos na Tabela 4.3 apresentam recorrentemente outras designações. Por exemplo, o método que corresponde à célula (N,N) é usualmente denominado de método de alisamento exponencial simples, enquanto que o método linear de Holt corresponde à entrada (A,N). Já as células (A,A) e (A,M) representam o método de Holt-Winters nas suas duas versões, aditiva e multiplicativa, respetivamente. As restantes entradas da Tabela 4.3 correspondem a outros métodos menos utilizados em previsão.

Para cada um dos 15 métodos apresentados, existem dois modelos em espaço de estados possíveis, um que considera erros aditivos e outro erros multiplicativos,

perfazendo um total de 30 modelos. Não se deve, no entanto, confundir métodos de alisamento exponencial com modelos em espaço de estados. Um método de alisamento exponencial trata-se de um algoritmo que produz apenas previsões pontuais. O respetivo modelo em espaço de estados fornece as mesmas previsões pontuais, mas também disponibiliza uma estrutura para determinar intervalos de confiança, entre outras propriedades. Uma vez que a distinção entre erros aditivos e multiplicativos não altera as estimativas pontuais, antes do aparecimento dos modelos em espaço de estados, a componente erro foi muitas vezes ignorada (Hyndman et al., 2008).

Os métodos de alisamento exponencial têm vindo a destacar-se devido à sua versatilidade, rapidez, fiabilidade e facilidade de automatização (Armstrong, 2001). Ainda assim, Gardner (2006) refere que a propriedade mais importante destes métodos é a robustez.

As vantagens dos métodos de alisamento exponencial em relação a outros métodos de previsão são a simplicidade de utilização, o baixo custo de computação e a facilidade de implementação em situações reais em que é necessário estabelecer previsões para muitas séries simultaneamente. No entanto, apesar de simples e versátil, a implementação destes métodos exige alguns cuidados, nomeadamente com os processos de inicialização das suas componentes e constantes de alisamento. A principal desvantagem dos métodos de alisamento exponencial prende-se com o facto de estes serem métodos não paramétricos e, por isso, não permitirem realizar inferências estatísticas de uma forma direta. Assim, a construção e interpretação de intervalos de previsão deve ser feita com algum cuidado sempre que os erros de previsão não tiverem um comportamento próximo de uma distribuição Normal de média zero. Em particular, os intervalos de previsão deixam de ser válidos sempre que os erros de previsão forem correlacionados entre si (Caiado, 2016).

Pode consultar-se uma revisão mais abrangente destes métodos em Gardner (2006).

Na Tabela 4.4 encontra-se descrita a notação utilizada no alisamento exponencial.

Tabela 4.4: Notação para o alisamento exponencial.

Símb.	Definição	Símb.	Definição
\hat{Y}_t	Previsão a 1-passo para o instante t	h	Horizonte de previsão
α	Constante de alisamento associada ao nível	l_t	Nível no instante t
β	Constante de alisamento associada ao declive	b_t	Declive no instante t
γ	Constante de alisamento associada à sazonalidade	s_t	Fator sazonal no instante t
ϕ	Parâmetro de amortecimento da tendência	s	Período sazonal

4.2.1 Método de alisamento exponencial simples

O método de alisamento exponencial simples, desenvolvido por Brown (1959), considera que a previsão para o período seguinte, \hat{Y}_{t+1} , não é mais do que a previsão para o período corrente, \hat{Y}_t , ajustada pelo respetivo erro de previsão, $e_t = Y_t - \hat{Y}_t$, i.e.,

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t), \quad (4.17)$$

onde α é uma constante de alisamento que varia entre 0 e 1. Quando α toma um valor próximo de 1, a previsão para o período seguinte sofre um ajustamento substancial através do erro de previsão, enquanto que, quando α se encontra próximo de 0, esse ajustamento é muito reduzido (Hyndman et al., 2008).

De forma equivalente, a equação (4.17) escreve-se como

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\hat{Y}_t. \quad (4.18)$$

A previsão \hat{Y}_{t+1} pode, então, ser interpretada como sendo uma média pesada da mais recente previsão e da mais recente observação, uma vez que se trata de uma ponderação entre ambas, com pesos $1 - \alpha$ e α , respetivamente. Além disso, desenvolvendo esta última equação, é fácil mostrar que \hat{Y}_{t+1} representa uma média móvel pesada de todas as observações anteriores, com pesos exponencialmente decrescentes (Hyndman et al., 2008). Efetivamente, se se substituir na equação (4.18) o valor de \hat{Y}_t por $\alpha Y_{t-1} + (1 - \alpha)\hat{Y}_{t-1}$ obtém-se

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1 - \alpha)Y_{t-1} + (1 - \alpha)^2\hat{Y}_{t-1},$$

e, repetindo esse processo sucessivamente, tem-se

$$\hat{Y}_{t+1} = (1 - \alpha)^t\hat{Y}_1 + \alpha \sum_{i=0}^{t-1} (1 - \alpha)^i Y_{t-i}.$$

Assim, pode dizer-se que, para cada observação, o peso decresce exponencialmente dependendo do valor do parâmetro α : quando α tem um valor grande, é atribuído um peso mais elevado às observações mais recentes; e quando α tem um valor pequeno, são os dados históricos que têm mais peso. Portanto, o valor do parâmetro α pode ser utilizado para ajustar a sensibilidade da previsão a mudanças no nível da série: quanto maior o valor de α , maior é o peso atribuído a Y_t e mais sensível é a previsão às mudanças no nível (Wang, 2006).

Uma das principais limitações de métodos recursivos prende-se com a inicializa-

ção do processo. De facto, se uma nova previsão depende sempre de uma previsão anterior, será necessário iniciar o processo com um valor pré-definido para a primeira previsão. Desta forma, Makridakis et al. (1998) recomendam considerar como previsão inicial a primeira observação, ou seja, $\hat{Y}_1 = Y_1$. Alternativamente, propõem usar a média dos primeiros quatro ou cinco valores da série temporal. No ambiente R, a inicialização do processo é feita usando a primeira opção, ou seja, considerando que o valor da previsão inicial corresponde à primeira observação.

Quanto à previsão para períodos futuros, a mais de um passo, pode dizer-se que, uma vez que este método é adequado apenas para séries sem tendência, sazonalidade ou qualquer outro tipo de padrão subjacente, a função de previsão é “constante”, i.e., $\hat{Y}_{t+h} = \hat{Y}_{t+1}$, com $h = 2, 3, \dots$.

Existe, ainda, outra forma de escrever a equação (4.18) que, embora mais “pesada”, se revela útil para a extensão deste método (Secções 4.2.2 e 4.2.3). Assim, para se obter a equação do nível, basta considerar que o nível da série é igual à previsão, i.e., $\hat{Y}_{t+1} = l_t$ e, portanto, $l_t = \alpha Y_t + (1 - \alpha)l_{t-1}$ e $\hat{Y}_{t+h} = l_t$.

4.2.2 Método linear de Holt

Holt (1957) estende o método de alisamento exponencial simples, desenvolvendo um método que permite lidar com séries temporais que apresentem tendência: o método linear de Holt (também conhecido apenas por método de Holt). Este método funciona de forma idêntica ao anterior, sendo, no entanto, necessária a atualização de duas componentes em cada período t , o nível e o declive, dando, por isso, origem a duas equações de atualização:

$$\text{Nível: } l_t = \alpha Y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (4.19)$$

$$\text{Declive: } b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}, \quad (4.20)$$

para α e β com valores compreendidos entre 0 e 1.

Analisando a expressão (4.21), pode dizer-se que a existência de tendência impossibilita que se façam previsões só com base na estimativa do nível, sendo necessário um segundo termo, hb_t , para projetar essa tendência para o futuro:

$$\text{Previsão: } \hat{Y}_{t+h} = l_t + hb_t, \quad h = 1, 2, \dots \quad (4.21)$$

No que respeita à inicialização do processo, o método de Holt exige duas estimativas, uma para l_1 e outra para b_1 . Assim, Makridakis et al. (1998) recomendam que

se defina $\hat{l}_1 = Y_1$ e, para a estimativa de b_1 , sugerem duas alternativas: $\hat{b}_1 = Y_2 - Y_1$ ou $\hat{b}_1 = (Y_4 - Y_1)/3$. Outra opção, quando os dados históricos estão disponíveis, passa por utilizar a regressão e, pelo método dos mínimos quadrados, determinar \hat{l}_1 e \hat{b}_1 a partir de alguns valores iniciais da série temporal. Já no ambiente R, os valores iniciais para l_1 e b_1 são Y_2 e $Y_2 - Y_1$, respetivamente.

Brown (1963) também propõe um método que permite lidar com tendências: o método de alisamento exponencial duplo. Este método consiste na aplicação do método de alisamento exponencial simples duas vezes, utilizando a mesma constante de alisamento. Para proceder ao alisamento duplo de Brown utilizam-se as equações de atualização

$$m_t = \alpha Y_t + (1 - \alpha)m_{t-1}, \quad (4.22)$$

$$d_t = \alpha m_t + (1 - \alpha)d_{t-1}, \quad (4.23)$$

onde m_t é a série de alisamento exponencial simples, d_t a série de alisamento exponencial duplo e α a constante de alisamento.

As previsões a h -passos são obtidas através da expressão (4.21), considerando que as estimativas do nível e do declive são dadas por:

$$\text{Nível: } l_t = 2m_t - d_t \quad (4.24)$$

$$\text{Declive: } b_t = \frac{\alpha}{1 - \alpha}(m_t - d_t). \quad (4.25)$$

Comparando as equações (4.19) e (4.20) com (4.24) e (4.25), pode ver-se que quando $\alpha = \beta$, o método de Holt e o método de alisamento exponencial duplo são equivalentes (Caiado, 2016). Desta forma, pode concluir-se que o método de alisamento exponencial duplo é um caso particular do método de Holt.

4.2.3 Método de Holt-Winters

O método de Holt-Winters trata-se de uma extensão do método de Holt proposta para situações em que, além da tendência, a série também apresenta sazonalidade. De forma a incorporar a componente sazonal, surge a necessidade de se considerar um terceiro parâmetro ($0 < \gamma < 1$) e, conseqüentemente, é introduzida uma terceira equação de atualização.

Na verdade, existem dois métodos de Holt-Winters distintos, dependendo se a sazonalidade se encontra na sua forma aditiva ou multiplicativa. Na sua forma aditiva, o método adapta-se a séries com amplitude sazonal independente ao nível da

série, enquanto que, na sua forma multiplicativa, o método é utilizado em séries cuja amplitude sazonal é proporcional ao nível da série. Os modelos de decomposição para os métodos aditivo e multiplicativo são, neste caso, o modelo de decomposição aditivo ($Y_t = T_t + S_t + \epsilon_t$) e o modelo de decomposição multiplicativo com erros aditivos ($Y_t = T_t \times S_t + \epsilon_t$), respetivamente (Cordeiro, 2011). Podem ver-se na Tabela 4.5 as equações que constituem cada um dos métodos, com $h_s^+ = [(h - 1) \bmod s] + 1$.

Tabela 4.5: Equações do método de Holt-Winters.

Método aditivo	Método multiplicativo
$l_t = \alpha(Y_t - s_{t-s}) + (1 - \alpha)(l_{t-1} + b_{t-1})$	$l_t = \alpha \frac{Y_t}{s_{t-s}} + (1 - \alpha)(l_{t-1} + b_{t-1})$
$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$	$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$
$s_t = \gamma(Y_t - l_t) + (1 - \gamma)s_{t-s}$	$s_t = \gamma \frac{Y_t}{l_t} + (1 - \gamma)s_{t-s}$
$\hat{Y}_{t+h} = l_t + hb_t + s_{t-s+h_s^+} \quad h = 1, 2, \dots$	$\hat{Y}_{t+h} = (l_t + hb_t)s_{t-s+h_s^+}, \quad h = 1, 2, \dots$

À semelhança do que acontece para os métodos de alisamento exponencial já apresentados, para inicializar o método de Holt-Winters são precisos valores iniciais para o nível, para o declive e para os índices de sazonalidade. Note-se que, segundo Makridakis et al. (1998), existem agora exigências sobre o número mínimo de observações: para determinar as estimativas iniciais para o nível e os índices de sazonalidade são necessárias, no mínimo, s observações (s corresponde ao período sazonal); e para as estimativas iniciais do declive são precisas, pelo menos, $2s$ observações. As expressões para a inicialização dos dois métodos de Holt-Winters (aditivo e multiplicativo) encontram-se na Tabela 4.6.

Tabela 4.6: Inicialização do método de Holt-Winters.

Modelo aditivo	Modelo multiplicativo
$\hat{l}_s = \frac{1}{s} \sum_{i=1}^s Y_i$	$\hat{l}_s = \frac{1}{s} \sum_{i=1}^s Y_i$
$\hat{b}_s = \frac{1}{s^2} (\sum_{i=s+1}^{2s} Y_i - \sum_{i=1}^s Y_i)$	$\hat{b}_s = \frac{1}{s^2} (\sum_{i=s+1}^{2s} Y_i - \sum_{i=1}^s Y_i)$
$\hat{s}_i = Y_i - \hat{l}_s$, para $i = 1, \dots, s$	$\hat{s}_i = \frac{Y_i}{\hat{l}_s}$, para $i = 1, \dots, s$

Para inicializar estes métodos, o ambiente R utiliza a função *decompose* para realizar uma decomposição da série nas suas componentes – tendência e sazonalidade –, através de um processo de médias móveis, e assim determinar os valores iniciais para l_s , b_s e s_i , $i = 1, \dots, s$ (\hat{l}_s e \hat{b}_s são determinadas através de uma regressão linear simples da tendência das primeiras observações).

Os métodos de Holt-Winters, além de permitirem a modelação de séries que apresentem tendência e sazonalidade, generalizam os métodos de alisamento exponencial já referidos. De facto, tanto o método de alisamento exponencial simples como o método de Holt podem ser derivados através das equações apresentadas, considerando os parâmetros β e γ (e os valores iniciais para o declive e para os fatores sazonais, respetivamente) como sendo nulos, sempre que se revele oportuno, i.e., sempre que a série não apresente tendência ($\beta = 0$) nem sazonalidade ($\gamma = 0$).

4.2.4 Outros métodos de alisamento exponencial

Como referido anteriormente, com os trabalhos de Gardner & McKenzie (1985), Hyndman et al. (2002) e Taylor (2003), são introduzidos os métodos de alisamento exponencial com tendência amortecida, particularmente úteis na previsão de séries temporais com taxas de crescimento não constantes ao longo do tempo e de efeitos amortecidos (Caiado, 2016).

Para séries não sazonais, pode usar-se o *método de Holt-Winters com tendência aditiva amortecida* (Ad,N), cujas equações de atualização são:

$$\text{Nível: } l_t = \alpha Y_t + (1 - \alpha)(l_{t-1} + \phi b_{t-1}) \quad (4.26)$$

$$\text{Declive: } b_t = \beta(l_t - l_{t-1}) + (1 - \beta)\phi b_{t-1}, \quad (4.27)$$

onde α , β e ϕ variam entre 0 e 1. A previsão a h -passos é, neste caso, dada por

$$\text{Previsão: } \hat{Y}_{t+h} = l_t + \sum_{i=1}^h \phi^i b_t, \quad h = 1, 2, \dots \quad (4.28)$$

Assim, o crescimento para a previsão a 1-passo é ϕb_t e o crescimento é amortecido por um fator de ϕ para cada período de tempo futuro adicional (Hyndman et al., 2008). Quando $\phi = 1$ este método é equivalente ao método linear de Holt.

Para séries sazonais, podem destacar-se dois métodos de tendência amortecida: o *método de Holt-Winters multiplicativo com tendência aditiva amortecida* (Ad,M) e o *método de Holt-Winters aditivo com tendência aditiva amortecida* (Ad,A). Note-se que, quando $\phi = 1$, estes métodos são equivalentes aos métodos de Holt-Winters multiplicativo e aditivo, respetivamente. Podem ver-se na Tabela 4.7 as equações que constituem cada um dos métodos (aditivo e multiplicativo), com $h_s^+ = [(h - 1) \bmod s] + 1$ e $h = 1, 2, \dots$

Tabela 4.7: Equações do método de Holt-Winters com tendência aditiva amortecida.

Método aditivo (Ad,A)	Método multiplicativo (Ad,M)
$l_t = \alpha(Y_t - s_{t-s}) + (1 - \alpha)(l_{t-1} + \phi b_{t-1})$	$l_t = \alpha \frac{Y_t}{s_{t-s}} + (1 - \alpha)(l_{t-1} + \phi b_{t-1})$
$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)\phi b_{t-1}$	$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)\phi b_{t-1}$
$s_t = \gamma(Y_t - l_t) + (1 - \gamma)s_{t-s}$	$s_t = \gamma \frac{Y_t}{l_t} + (1 - \gamma)s_{t-s}$
$\hat{Y}_{t+h} = l_t + \sum_{i=1}^h \phi^i b_t + s_{t-s+h_s^+}$	$\hat{Y}_{t+h} = (l_t + \sum_{i=1}^h \phi^i b_t) s_{t-s+h_s^+}$

Um método de tendência amortecida é, então, apropriado quando é detetada uma tendência que aparenta não se prolongar no futuro. As equações deste tipo de método fazem o que o nome indica: atenuam a tendência à medida que o horizonte de previsão aumenta, o que, geralmente, melhora a precisão das previsões, principalmente para horizontes de previsão longínquos (Hyndman et al., 2008).

4.2.5 Intervalos de previsão

Depois de se obter previsões pontuais, pode ser útil construir um intervalo de previsão para cada instante temporal. Efetivamente, como o futuro é desconhecido e as previsões não passam disso mesmo, a indicação de uma estimativa intervalar permite quantificar a incerteza associada às previsões pontuais.

Os intervalos de previsão são, usualmente, baseados na medida do erro quadrático médio (EQM),

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

que constitui uma estimativa da variância do erro de previsão a 1-passo (i.e., com horizonte temporal de uma unidade temporal). Se se admitir a hipótese de que os erros de previsão têm distribuição aproximadamente Normal de média zero, pode, então, construir-se um intervalo de previsão aproximado para cada instante do tempo.

Assim, supondo a normalidade dos erros de previsão, o intervalo de previsão a 1-passo para o instante t pode ser obtido através da expressão (Makridakis et al., 1998)

$$\left(\hat{Y}_t - z_{1-\alpha/2} \sqrt{\text{EQM}}, \hat{Y}_t + z_{1-\alpha/2} \sqrt{\text{EQM}} \right),$$

onde z é o quantil da distribuição Normal padrão e $1 - \alpha$ corresponde ao nível de confiança do intervalo.

Se o objetivo da previsão, no que se refere ao horizonte temporal, for de médio ou longo prazo, devem calcular-se previsões a h -passos utilizando exclusivamente a

informação disponível da série até ao momento t . Neste caso, a medida do EQM passa a ser determinada com base nos erros de previsão a h -passos, dando lugar à expressão

$$\text{EQM}_{(h)} = \frac{1}{n-h} \sum_{t=h+1}^n (Y_t - \hat{Y}_{t|t-h})^2, \quad (4.29)$$

onde h é o horizonte temporal. Nos casos em que a série temporal apresenta sazonalidade, a expressão (4.29) deve ser substituída por

$$\text{EQM}_{(h)} = \frac{1}{n-h-s+1} \sum_{t=h+s}^n (Y_t - \hat{Y}_{t|t-h})^2,$$

em que s é o período sazonal. Desta forma, para construir intervalos de previsão empíricos (ou aproximados) multi-passos (h unidades temporais) deve utilizar-se a expressão

$$\left(\hat{Y}_{t+h} - z_{1-\alpha/2} \sqrt{\text{EQM}_{(h)}}, \hat{Y}_{t+h} + z_{1-\alpha/2} \sqrt{\text{EQM}_{(h)}} \right). \quad (4.30)$$

Como referido anteriormente, estes intervalos constituem uma medida da incerteza das previsões obtidas através dos diferentes métodos de previsão. De facto, a previsão intervalar é uma das partes mais importantes da fundamentação estatística, uma vez que as previsões não são exatas e os intervalos destacam isso mesmo.

4.3 Modelos de regressão linear

A análise de regressão linear é uma técnica estatística que permite estudar a relação entre uma ou mais variáveis explicativas (independentes) e a variável resposta (dependente), para que, com base nas primeiras, seja possível prever a última ou simplesmente identificar quais os fatores (variáveis independentes) que explicam o fenómeno em estudo (variável resposta).

A origem da análise de regressão linear remonta ao século XIX (1877) e a sua primeira documentação foi feita por Galton, em Inglaterra, num artigo científico que mostrou a existência de uma relação linear entre o diâmetro dos grãos de ervilhas e o diâmetro dos grãos descendentes (Stanton, 2001).

As componentes de tendência e sazonalidade, presentes nas séries temporais, podem também ser explicadas através de modelos de regressão linear. A componente tendência é, na análise de regressão, vista como determinística e, portanto, pode ser explicada através de funções polinomiais do tempo. Já para modelar a componente sazonal, pode optar-se tanto por uma descrição qualitativa do padrão sazonal, atra-

vés de variáveis indicatrizes, como pela utilização de funções harmônicas, com ondas sinusoidais. Desta forma, quando uma série temporal revela tendência e sazonalidade, deve construir-se uma combinação das várias abordagens de forma a explicar num só modelo todo o comportamento da série.

No entanto, os dados de séries temporais apresentam, usualmente, uma forte autocorrelação, o que resulta na violação de um dos principais pressupostos da análise de regressão linear: o pressuposto da não correlação dos erros. Assim, é necessário modelar a componente aleatória do modelo de regressão linear, o que, usualmente, é conseguido através de um modelo autorregressivo.

Comparativamente com os restantes modelos de séries temporais, os modelos de regressão linear têm a vantagem de permitir, além da inclusão de informações de observações passadas (através de variáveis explicativas que são funções do tempo), a inclusão de outras informações que possam ser relevantes na explicação do comportamento da série (Hyndman & Athanasopoulos, 2014).

4.3.1 Conceitos fundamentais

Num modelo de regressão linear (univariado) pretende-se determinar uma relação funcional entre uma variável dependente (ou resposta), Y , e uma ou mais variáveis independentes (ou explicativas), X_j , com $j = 1, \dots, p$,

$$Y = f(X_1, X_2, \dots, X_p).$$

No entanto, é comum que a variável Y seja afetada por outros fatores (desconhecidos), além dos considerados no modelo, e, portanto, ao analisar Y em função de X_1, X_2, \dots, X_p deve incluir-se, além da parte determinística, uma componente aleatória, isto é,

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon.$$

Modelo de regressão linear múltipla

O *modelo de regressão linear múltipla* descreve a relação entre uma ou mais variáveis independentes e uma variável dependente através da forma funcional

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \dots + \beta_p X_i^p + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (4.31)$$

onde Y é a variável dependente, X_1, X_2, \dots, X_p são as p variáveis independentes, $\beta_0, \beta_1, \dots, \beta_p$ são os parâmetros ou coeficientes de regressão, e ϵ_i é o erro aleatório

(não observável) associado ao valor observado Y_i , que representa os outros fatores (desconhecidos) que influenciam a variável em estudo. Os coeficientes de regressão, β_0 e β_j ($j = 1, \dots, p$), representam, respetivamente, o valor esperado da variável Y quando as variáveis explicativas são simultaneamente nulas e a variação do valor esperado da variável Y por cada incremento unitário na variável X_j , quando se mantêm constantes as restantes variáveis explicativas.

Este modelo pode também ser apresentado através da sua forma matricial

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad (4.32)$$

onde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1^1 & \cdots & X_1^p \\ 1 & X_2^1 & \cdots & X_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n^1 & \cdots & X_n^p \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

ou seja, em que \mathbf{Y} é o vetor coluna ($n \times 1$) das observações da variável dependente, \mathbf{X} é a matriz ($n \times (p + 1)$) cujas linhas são constituídas pelos valores das variáveis independentes, β é o vetor coluna ($(p + 1) \times 1$) dos parâmetros ou coeficientes da regressão, e ϵ é o vetor coluna ($n \times 1$) dos erros aleatórios.

O *modelo de regressão linear simples* é o caso particular quando $p = 1$, isto é, quando o modelo contém apenas uma variável explicativa, X . A relação entre Y e X pode ser descrita através da simplificação da forma funcional (4.31), ou seja,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (4.33)$$

onde Y_i é a i -ésima observação da variável dependente Y , X_i é a i -ésima observação da variável independente X , β_0 e β_1 são os parâmetros ou coeficientes de regressão, e ϵ_i é o erro aleatório associado ao valor observado Y_i . Estes coeficientes de regressão, β_0 e β_1 , representam, respetivamente, a ordenada na origem (ou nível) e o declive, e têm a mesma interpretação que os parâmetros do modelo de regressão linear múltipla.

Pressupostos

Geralmente, a aplicação do modelo de regressão linear múltipla assenta num conjunto de pressupostos, que não devem ser violados:

- P1:** o valor esperado das perturbações aleatórias é nulo, $E[\epsilon_i] = 0, \forall i$;
- P2:** os erros são homocedásticos, isto é, são constantes ao longo das observações, $Var[\epsilon_i] = \sigma^2, \forall i$ (variância constante desconhecida);
- P3:** não existe autocorrelação nos erros aleatórios, isto é, a covariância entre dois quaisquer erros aleatórios ϵ_i e ϵ_j ($i \neq j$) é nula, $Cov(\epsilon_i, \epsilon_j) = 0, \forall i, j, i \neq j$;
- P4:** os erros aleatórios têm uma distribuição aproximadamente Normal multivariada, $\epsilon \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, onde \mathbf{I}_n é a matriz identidade;
- P5:** os valores das variáveis independentes X_1, X_2, \dots, X_p são conhecidos na amostra, em que a matriz \mathbf{X} é não aleatória;
- P6:** as variáveis explicativas não estão correlacionadas entre si.

No caso do modelo de regressão linear simples, os pressupostos de média nula (P1), variância constante (P2) e independência dos erros (P3) mantêm-se. No entanto, por se estar perante um modelo com apenas uma variável explicativa, os erros aleatórios seguem uma distribuição Normal univariada, $\epsilon_i \sim N(0, \sigma^2)$, em vez de uma Normal multivariada. Quanto aos pressupostos estabelecidos sobre as variáveis explicativas (P5 e P6), pode apenas afirmar-se que a variável independente, X , assume valores conhecidos na amostra, não tendo carácter de variável aleatória (note-se que o P6 deixa de fazer sentido quando $p = 1$).

Estimação do modelo

A partir de uma amostra de n observações das variáveis Y e X_1, X_2, \dots, X_p , é possível estimar, através do método dos mínimos quadrados, os coeficientes de regressão, $\beta_0, \beta_1, \dots, \beta_p$. O princípio do método dos mínimos quadrados consiste em determinar o conjunto de valores $\beta_0, \beta_1, \dots, \beta_p$ que minimizam a soma dos quadrados dos desvios entre os valores observados (Y_i) e os valores estimados (\hat{Y}_i) da função de regressão da amostra, ou seja, que minimizam a soma dos quadrados dos resíduos (e_i) de estimação,

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i^1 + \dots + \hat{\beta}_p X_i^p))^2,$$

ou, em notação matricial,

$$\mathbf{e}^T \mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}^T \mathbf{Y} - 2\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{Y} + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}},$$

onde $\mathbf{e}^T \mathbf{e}$ é a soma de quadrados residual (SQE), $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ representa o vetor de estimadores dos coeficientes de regressão e $\mathbf{X}\hat{\beta} = (\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)$ o vetor de valores estimados. Derivando a expressão $\mathbf{e}^T \mathbf{e}$ em ordem ao vetor de estimadores $\hat{\beta}$ e igualando a zero, é possível obter as estimativas dos parâmetros, na sua forma matricial

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

A matriz de variâncias e covariâncias dos estimadores dos mínimos quadrados assume a expressão

$$\text{Var}[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1},$$

onde σ^2 é a variância (constante) dos erros aleatórios ϵ_i , e $(\mathbf{X}^T \mathbf{X})^{-1}$ é a matriz inversa dos produtos cruzados de \mathbf{X} . Uma vez que σ^2 é, geralmente, desconhecida, deve utilizar-se o seu estimador, definido pela expressão

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^n e_i^2 = \frac{\text{SQE}}{n - p - 1},$$

onde n é o número de observações e $p + 1$ é o número de parâmetros do modelo.

Inferência estatística

Assumindo a hipótese de normalidade dos erros aleatórios, pode testar-se a significância estatística de qualquer variável explicativa associada a um parâmetro $\beta_j, j = 1, \dots, p$, através de um teste de hipóteses de $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$. Sob a hipótese nula, a estatística do teste, definida por

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}},$$

segue uma distribuição t -Student com $n - p - 1$ graus de liberdade, em que $\hat{\sigma}_{\hat{\beta}_j}^2$ corresponde ao j -ésimo elemento da diagonal da matriz (estimada) de variâncias e covariâncias dos estimadores dos parâmetros, $\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$. Para um nível de significância α , a região de rejeição da hipótese nula corresponde ao conjunto $]-\infty, -t_{1-\alpha/2; n-p-1}] \cup [t_{1-\alpha/2; n-p-1}, +\infty[$. Analogamente, se o valor de prova for inferior ou igual a α , rejeita-se H_0 e pode concluir-se que o coeficiente β_j é estatisticamente significativo ($\beta_j \neq 0$), ou seja, que a variável independente X_j contribui significativamente para explicar a variável dependente, considerando as restantes variáveis explicativas constantes. Note-se que, a significância do parâmetro β_0 pode também ser testada por este método, não tendo, no entanto, interpretação relacio-

nada com a significância de nenhuma variável independente.

Para determinar o intervalo de confiança para o parâmetro β_j , com $j = 0, 1, \dots, p$, a um nível de confiança $(1 - \alpha)100\%$, basta calcular

$$\left(\hat{\beta}_j - t_{1-\alpha/2;n-p-1} \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{1-\alpha/2;n-p-1} \hat{\sigma}_{\hat{\beta}_j} \right).$$

Se se pretender testar a significância global da regressão, deve testar-se a hipótese conjunta $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ contra a hipótese alternativa de que nem todos os coeficientes são simultaneamente iguais a zero, isto é, contra a hipótese $H_1 : \exists_{j=1,\dots,p} : \beta_j \neq 0$ (existe pelo menos um coeficiente de regressão significativamente diferente de zero). A estatística de teste é dada por

$$F = \frac{(n - p - 1) \text{SQR}}{p \text{SQE}} \sim F_{p,n-p-1},$$

onde SQE é a soma de quadrados residual, SQR é a soma de quadrados da regressão e $F_{p,n-p-1}$ é a distribuição F de Fisher com p e $n - p - 1$ graus de liberdade. Se o valor de prova for inferior ou igual ao nível de significância α , rejeita-se H_0 e conclui-se que existe pelo menos um coeficiente β_j , com $j = 1, 2, \dots, p$, estatisticamente significativo. Para implementar o teste de significância global é usual construir-se uma tabela de análise de variância (*analysis of variance*, ANOVA), idêntica à que se encontra na Tabela 4.8.

Tabela 4.8: Tabela de análise de variância (ANOVA).

Fonte de variação	Soma de quadrados	Graus de liberdade	Média de quadrados	Valor F
Regressão	$\text{SQR} = \sum (\hat{Y}_i - \bar{Y})^2$	p	$\text{MQR} = \text{SQR}/p$	
Residual	$\text{SQE} = \sum (Y_i - \hat{Y}_i)^2$	$n - p - 1$	$\text{MQE} = \text{SQE}/(n - p - 1)$	$\frac{\text{MQR}}{\text{MQE}}$
Total	$\text{SQT} = \sum (Y_i - \bar{Y})^2$	$n - 1$	$\text{MQT} = \text{SQT}/(n - 1)$	

Qualidade do modelo

Para avaliar a qualidade de ajustamento do modelo de regressão linear é usual recorrer-se ao coeficiente de determinação

$$R^2 = 1 - \frac{\text{SQE}}{\text{SQT}} = \frac{\text{SQR}}{\text{SQT}}, \quad (4.34)$$

que representa a percentagem da variabilidade da variável resposta Y , que é explicada pelo conjunto das variáveis independentes. Note-se que, o coeficiente de determinação varia entre 0 e 1 (i.e., $0 \leq R^2 \leq 1$) e que, quanto mais próximo de 1 for o valor de R^2 , melhor é o ajustamento do modelo. No entanto, este coeficiente pode induzir em erro: um aumento do valor de R^2 não implica necessariamente que a qualidade de ajustamento do modelo tenha melhorado, uma vez que a adição de uma variável explicativa aumenta sempre o valor deste coeficiente. Por esse motivo, o coeficiente de determinação, R^2 , deve apenas ser utilizado para comparar modelos que tenham o mesmo número de variáveis independentes a explicar a mesma variável resposta.

Para os restantes casos, deve optar-se pelo coeficiente de determinação ajustado (R_a^2), que tem em consideração os graus de liberdade

$$R_a^2 = 1 - \frac{\text{MQE}}{\text{MQT}} = 1 - \frac{n-1}{n-p-1}(1-R^2). \quad (4.35)$$

De facto, enquanto que o R^2 aumenta sempre que uma nova variável explicativa é introduzida no modelo, o R_a^2 aumenta apenas se a nova variável proporciona um melhor ajustamento no modelo. Ao comparar dois modelos, deve escolher-se aquele que apresenta um maior coeficiente de determinação ajustado, porque será aquele que melhor explica a variabilidade de Y .

Análise dos resíduos

Para averiguar se um modelo de regressão linear é adequado, é usual proceder-se à análise dos resíduos resultantes, que devem cumprir os pressupostos associados.

Os pressupostos de média nula, homocedasticidade da variância e normalidade dos erros, podem ser verificados seguindo os procedimentos indicados na Secção 4.1.4.

Quanto ao pressuposto da independência, este pode ser avaliado através da observação da FAC dos resíduos, à semelhança do que acontece no caso dos modelos SARIMA. Um método alternativo para avaliar a validade do pressuposto de independência envolve o uso da estatística de Durbin-Watson, que testa a hipótese nula de independência (sem autocorrelação). A estatística de Durbin-Watson é dada por

$$\text{DW} = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}, \quad (4.36)$$

onde e_i corresponde ao resíduo da observação i e $0 \leq \text{DW} \leq 4$. Quando $\text{DW} \approx 2$,

não se rejeita a hipótese de independência. Além disso, valores inferiores a 2 indicam a existência de autocorrelação positiva, enquanto que valores superiores a 2 revelam autocorrelação negativa (Makridakis et al., 1998).

Para detetar a existência de multicolinearidade, isto é, de correlação entre variáveis explicativas, pode recorrer-se tanto à matriz de correlação das variáveis explicativas (se existirem pares de variáveis altamente correlacionadas, então deve-se eliminar uma delas, deixando aquela que fizer mais sentido do ponto de vista prático), como ao fator de inflação da variância (*variance inflation factor*, VIF), que é dado por

$$\text{VIF} = \frac{1}{1 - R_j^2},$$

onde R_j^2 é o coeficiente de determinação da regressão da variável X_j em função das restantes variáveis explicativas. Os VIFs medem o quanto a variância de um coeficiente de regressão estimado aumenta se as variáveis explicativas estão correlacionadas. Quando $\text{VIF} \approx 1$ as variáveis explicativas são independentes. Já se $\text{VIF} > 10$ as variáveis explicativas estão correlacionadas.

Previsão pontual e intervalar

Um dos objetivos da análise de regressão linear consiste em prever os valores da variável dependente \mathbf{Y} para valores hipotéticos ou observados de \mathbf{X} .

Uma previsão pontual de Y_k obtém-se substituindo no modelo de regressão linear múltipla as variáveis explicativas por valores observados ou hipotéticos $\mathbf{x}_k = (x_k^1, x_k^2, \dots, x_k^p)$, isto é,

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k^1 + \hat{\beta}_2 x_k^2 + \dots + \hat{\beta}_p x_k^p.$$

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

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

e limite superior dado por

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

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

4.3.2 Modelos de regressão para a tendência

Modelo de regressão linear

Ao examinar visualmente uma série temporal, é possível descrever as mudanças que a tendência apresenta ao longo do tempo. No entanto, o comportamento de uma tendência pode ser formalmente explicado por uma regressão sobre uma variável que representa o tempo, t (considera-se que a variável t toma valores de 1 a n para uma série com n observações igualmente espaçadas).

No caso mais simples, a tendência pode ser modelada como uma função linear do tempo, o que corresponde ao caso de um modelo de regressão linear simples em que a variável explicativa é a variável tempo, i.e.,

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t, \quad (4.37)$$

onde β_0 representa o nível da série no tempo $t = 0$, β_1 representa a variação da série temporal associada ao aumento de uma unidade na variável tempo, t é a variável tempo, e ϵ_t é a componente aleatória.

Quando o modelo linear não é suficiente para explicar a tendência da série temporal, pode resultar em previsões imprecisas e/ou apresentar tendência na série dos resíduos.

Modelo de regressão polinomial

Em algumas séries temporais, a tendência apresenta um comportamento não linear, que não é totalmente explicado pelo modelo mais simples. Nesses casos, é possível adicionar um (ou mais) termo polinomial de ordem superior a 1 ao modelo linear, o que corresponde a modelar a tendência como quadrática, cúbica, etc., isto é,

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_p t^p + \epsilon_t, \quad (4.38)$$

onde t representa a variável tempo definida anteriormente, e p representa o grau do polinômio que descreve a tendência. Nos casos particulares em que a tendência é quadrática e cúbica tem-se $p = 2$ e $p = 3$, respetivamente.

4.3.3 Séries temporais interrompidas

Uma série temporal interrompida não é mais do que uma série temporal dividida em duas ou mais partes, em pontos específicos no tempo (pontos de corte ou

change-points), por revelar alterações no seu comportamento resultantes de um dado acontecimento (diga-se, evento).

Caso o tempo de ocorrência do evento (ponto de corte) não seja especificado *a priori*, deve recorrer-se à análise exploratória ou inferencial da série de forma a identificar o momento da mudança. No entanto, esse momento nem sempre corresponde à ocorrência exata do evento, uma vez que, em alguns casos, é necessário algum tempo para o evento revelar algum efeito (chamado o período de equilíbrio). Assim, e dada a subjetividade da análise exploratória, pode proceder-se à comparação de modelos com diferentes tempos iniciais para o segmento pós-evento e, dessa forma, definir um ponto de corte (Jebb et al., 2015).

Modelo de regressão para séries temporais interrompidas

A análise de regressão segmentada é um método para modelar estatisticamente os dados de séries temporais interrompidas, com o objetivo de extrair conclusões mais formais sobre o impacto de um evento na medida de interesse.

Os modelos mais simples para séries interrompidas ajustam uma regressão linear a cada um dos segmentos da variável tempo (pré e pós-evento), fazendo, para tal, uso de dois parâmetros distintos para cada segmento: o nível e o declive. Sempre que ocorre uma mudança significativa nos parâmetros, esta representa um efeito do evento: uma mudança no nível da série indica uma alteração imediata, enquanto que, uma mudança no declive reflete um efeito mais gradual (Wagner et al., 2002).

Formalmente, este modelo de regressão define-se como

$$Y_t = \beta_0 + \beta_1 t + \beta_2 \times \text{evento}_t + \beta_3 t \times \text{evento}_t + \epsilon_t, \quad (4.39)$$

onde t é a variável tempo, evento_t é uma variável indicatriz que codifica a ocorrência ou não de um evento (i.e., assume o valor 0 nos tempos pré-evento e 1 a partir da ocorrência do mesmo), β_0 e β_1 representam, respetivamente, o nível para $t = 0$ (ou ordenada na origem) e o declive da tendência pré-evento, e β_2 e β_3 representam as mudanças no nível e no declive que ocorrem após o evento. Desta forma, a soma de β_1 e β_3 corresponde ao declive pós-evento, o mesmo acontecendo com a soma de β_0 e β_2 relativamente ao nível pós-evento.

Modelos mais complexos podem incluir outros tipos de tendência (nomeadamente quadrática, cúbica, etc.) ou até mais do que um ponto de corte (ou seja, mais do que um evento), tendo sempre por base a análise do comportamento da série temporal.

4.3.4 Modelos de regressão para a sazonalidade

Modelo com indicadores sazonais

Para modelar séries com sazonalidade pode optar-se por uma descrição qualitativa do padrão sazonal, através da especificação de um modelo de regressão que inclua uma variável indicatriz para representar cada um dos s períodos sazonais. Este modelo com indicadores sazonais pode ser escrito como

$$Y_t = T_t + \beta_1 D_1 + \dots + \beta_s D_s + \epsilon_t, \quad (4.40)$$

onde T_t representa a tendência (em função de t sem o termo constante, β_0), β_1, \dots, β_s são os coeficientes que refletem os s efeitos sazonais e D_i ($i = 1, \dots, s$) são as s variáveis indicatrizes que representam os diferentes períodos sazonais: tomam o valor 1 quando o tempo t pertence ao período i e 0 nos restantes casos. Por exemplo, para dados mensais, se D_1 corresponder às ocorrências no mês de janeiro (i.e., 1, se t ocorre em janeiro e 0, caso contrário), então β_1 só é tido em consideração para observações registadas nesse mês.

Intuitivamente, este modelo pode ser visto como um modelo linear onde existe um nível distinto para cada período, que representa o seu efeito.

Modelo sazonal harmónico

Nos modelos com indicadores sazonais, é atribuído um parâmetro (e uma variável) a cada período sazonal. No entanto, os efeitos sazonais geralmente variam de forma suave e contínua, pelo que pode ser mais eficiente e realista usar uma função de suavização em vez de indicadores sazonais. Assim, uma outra abordagem para modelar efeitos sazonais passa por ajustar um modelo sazonal harmónico, que recorre a funções seno e cosseno para descrever as oscilações observadas ao longo do tempo (Cowpertwait & Metcalfe, 2009).

Uma onda sinusoidal de frequência f (ciclos por intervalo amostral), amplitude A e constante de fase ϕ , pode ser escrita como

$$A \sin(2\pi ft + \phi) = \alpha_c \cos(2\pi ft) + \alpha_s \sin(2\pi ft),$$

onde $\alpha_s = A \cos(\phi)$ e $\alpha_c = A \sin(\phi)$.

Para uma série temporal de período s existem $s/2$ ciclos possíveis (no caso em que s não é divisível por 2, deve considerar-se a parte inteira da divisão) e, portanto,

pode definir-se o modelo sazonal harmónico

$$Y_t = T_t + \sum_{i=1}^{s/2} (\alpha_i \cos(2\pi it/s) + \beta_i \sin(2\pi it/s)) + \epsilon_t, \quad (4.41)$$

onde T_t representa a tendência (em função de t), α_i e β_i ($i = 1, \dots, s/2$) são os parâmetros desconhecidos de interesse, s é o período sazonal (e.g., $s = 12$ para dados mensais), i é um índice que varia entre 1 e $s/2$, e t é uma variável codificada que representa o tempo (e.g., $t = 1, \dots, 90$ para 90 observações igualmente espaçadas).

Apesar de parecer mais complexo, este modelo pode envolver um menor número de parâmetros relativamente ao modelo de indicadores sazonais, uma vez que, como os coeficientes harmónicos são independentes, todos os coeficientes que não são estatisticamente significativos para o modelo completo podem ser eliminados (Cowpertwait & Metcalfe, 2009).

Quando as tendências para os vários períodos apresentam não só diferenças em relação ao nível mas também relativamente ao declive, pode ser necessário incluir um novo termo no modelo harmónico, que descreva a variação dos declives ao longo do período sazonal. Nesses casos, Alpuim & El-Shaarawi (2009) sugerem que o modelo seja escrito como

$$Y_t = T_t + \sum_{i=1}^{s/2} \left[\alpha_i \cos\left(\frac{2\pi it}{s}\right) + \beta_i \sin\left(\frac{2\pi it}{s}\right) \right] + \sum_{i=1}^{s/2} \left[\gamma_i t \cos\left(\frac{2\pi it}{s}\right) + \delta_i t \sin\left(\frac{2\pi it}{s}\right) \right] + \epsilon_t, \quad (4.42)$$

onde o primeiro somatório descreve a variação sazonal dos dados e o segundo especifica a variação da tendência ao longo de cada período (quando $s = 12$, de cada ano).

4.3.5 Modelos de regressão com erros correlacionados

Como referido na introdução da Secção 4.3, no que respeita às séries temporais, observações sucessivas podem estar correlacionadas e, por isso, o ajustamento de um modelo de regressão a dados de séries temporais resulta, muitas vezes, no incumprimento do pressuposto da independência dos erros, o que acarreta algumas consequências.

De facto, a existência de autocorrelação faz com que os estimadores dos mínimos quadrados deixem de ser eficientes, ainda que se mantenham não enviesados e, em

geral, consistentes. Em casos de autocorrelação positiva, a variância dos estimadores é normalmente subestimada o que provoca a sobrestimação da estatística T e, conseqüentemente, deixa de ser válida toda a inferência estatística sobre os parâmetros do modelo (Caiado, 2016). Por esse motivo, é muito importante fazer uma análise dos resíduos quando se utiliza um modelo de regressão para a modelação de séries temporais.

Para estudar a estrutura de autocorrelação dos erros, podem, por exemplo, utilizar-se as FAC e FACP dos resíduos que, além de permitirem detetar a existência de autocorrelação, fornecem informação sobre qual o modelo mais adequado para a explicar.

A proposta apresentada por Alpuim & El-Shaarawi (2009) para o problema da existência de autocorrelação nos erros, consiste em considerar um modelo de regressão da forma

$$Y_t = \beta_0 + \beta_1 X_t^1 + \beta_2 X_t^2 + \dots + \beta_p X_t^p + \epsilon_t, \quad (4.43)$$

onde a componente aleatória, ϵ_t , é um processo gaussiano autorregressivo de ordem k , AR(k), ou seja,

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \dots + \phi_k \epsilon_{t-k} + a_t,$$

ou, alternativamente,

$$\Phi(B)\epsilon_t = a_t,$$

em que a_t é um processo de ruído branco gaussiano e $\Phi(B)$ é o polinómio autorregressivo de ordem k .

Quando a componente aleatória é um processo autorregressivo, prova-se (Alpuim & El-Shaarawi, 2008) que, sob certas condições da matriz de planeamento \mathbf{X} , os estimadores dos mínimos quadrados e os estimadores de máxima verosimilhança são assintoticamente equivalentes e totalmente eficientes. Para tal, o conjunto das p variáveis explicativas no tempo t , $\mathbf{X}_t^T = (X_t^1, X_t^2, \dots, X_t^p)$, deve verificar uma relação linear recursiva do tipo

$$\mathbf{X}_t = \Psi \mathbf{X}_{t-1}, \quad (4.44)$$

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

Além disso, para um modelo de equação (4.43) que verifique a condição (4.44), mostra-se que (Alpuim & El-Shaarawi, 2008) o vetor dos estimadores dos mínimos quadrados é assintoticamente Normal com matriz de variâncias e covariâncias dada

por

$$Var[\hat{\beta}] = \sigma_a^2 [\Phi(B)\mathbf{X}^T\Phi(B)\mathbf{X}]^{-1}, \quad (4.45)$$

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

$$X_t^{j*} = \Phi(B)X_t^j = X_t^j - \phi_1 X_{t-1}^j - \dots - \phi_k X_{t-k}^j, \quad (4.46)$$

para $j = 1, \dots, p$ e $t = k + 1, \dots, n$. Na prática, os valores dos coeficientes autor-regressivos, ϕ_1, \dots, ϕ_k , e da variância do ruído branco, σ_a^2 , são desconhecidos. No entanto, para amostras grandes (n elevado), estes podem ser substituídos por estimadores consistentes, o que permite realizar um teste assintótico à significância de cada variável explicativa, baseado na distribuição Normal (Alpuim & El-Shaarawi, 2009).

Em modelos harmónicos, deve ter-se em atenção que uma curva cosseno com um certo período deve ser sempre incluída ou eliminada juntamente com o seno correspondente com o mesmo período e vice-versa. Esta exigência é feita de forma a que o modelo verifique a equação (4.44), garantindo, por um lado, que os estimadores dos mínimos quadrados sejam ótimos e, por outro, que a fórmula para a variância dos estimadores (4.45) possa ser aplicada (Alpuim & El-Shaarawi, 2009).

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

4.4 Medidas de avaliação

Como referido neste Capítulo, são vários os métodos existentes para se fazer previsão de séries temporais, desde os mais complexos, que envolvem diferentes parâmetros, aos métodos mais simples e de fácil compreensão. No entanto, o facto de se usar procedimentos mais complexos nem sempre significa uma melhoria na qualidade das previsões. Então, como escolher o melhor método de previsão?

Yokum & Armstrong (1995) enumeram uma série de fatores que devem ser considerados aquando da escolha do método de previsão, incluindo a precisão das pre-

visões, a facilidade de utilização, de interpretação e de implementação do método, a poupança de custos, a flexibilidade, etc.. Chatfield (2000) acrescenta, ainda, a experiência do analista, as características da série em estudo, entre outros.

As medidas de avaliação são, em muitos dos casos de análise de previsões, o principal critério de seleção de um método de previsão. Através destas medidas pretende-se avaliar a eficácia de um modelo ou método de previsão, mostrando o quanto ele é capaz de reproduzir a série temporal em estudo e de obter previsões que venham a revelar-se o mais “corretas” possível. No caso das previsões, é a precisão da previsão futura que mais importa (Cordeiro, 2011).

Hyndman & Koehler (2006) classificam as medidas de avaliação em quatro tipos distintos:

Medidas dependentes da escala (*scale-dependent*): tal como o nome indica, estas medidas de precisão possuem uma escala que depende da escala dos dados. São bastante utilizadas e muito úteis para comparar diferentes métodos aplicados ao mesmo conjunto de dados, não devendo, no entanto, ser escolhidas, por exemplo, para comparar dados com diferentes escalas. As medidas dependentes da escala mais usadas baseiam-se em erros absolutos (e.g., erro absoluto médio, EAM) ou erros quadráticos (e.g., erro quadrático médio, EQM), o que evita que erros negativos e positivos se “compensem” uns aos outros;

Medidas de erros percentuais: o erro percentual é dado por $p_t = 100e_t/Y_t$. Os erros percentuais têm a vantagem de ser independentes da escala e, portanto, são usados com frequência para comparar o desempenho de previsão entre diferentes séries. Contudo, as medidas baseadas neste tipo de erros têm a desvantagem de ser infinitas ou indefinidas se $Y_t = 0$, para qualquer t no período de interesse, e de ter uma distribuição extremamente assimétrica quando Y_t está próximo de zero. Outro problema com estas medidas, frequentemente ignorado, é o facto de ser assumido um zero significativo (e.g., não faz sentido aplicar este tipo de erros para medir a precisão de previsões de temperatura nas escalas Fahrenheit ou Celsius);

Medidas de erros relativos: uma alternativa aos erros percentuais no cálculo de medidas independentes da escala envolve a divisão de cada erro pelo erro correspondente obtido através de um método de previsão de referência (usualmente, o método de referência é o método *naïve* em que \hat{Y}_t é igual à última observação). Por não dependerem da escala dos dados, as medidas de erros relativos também são recomendadas para avaliar a precisão das previsões entre várias

séries. No entanto, quando os erros são pequenos, o uso do método *naïve* como referência não é adequado, uma vez que envolve divisão por zero;

Medidas de erros escalados (*scale-free* ou *scaled*): estas medidas, propostas por Hyndman & Koehler (2006), são também independentes da escala e, por isso, podem ser utilizadas tanto para comparar diferentes métodos de previsão aplicados a uma única série como para comparar a precisão das previsões entre diversas séries. As medidas de erros escalados envolvem a divisão de cada erro pelo EAM *in-sample* (EAM calculado para a amostra de treino) do método *naïve* (passeio aleatório). A única circunstância em que estas medidas são infinitas ou indefinidas é quando todas as observações históricas são iguais.

Um dos aspetos fundamentais do exercício de previsão diz respeito à escolha da amostra de treino (*training set*) para ajustar o modelo e da amostra de teste (*test set*) para testar a sua qualidade preditiva. Apesar de um bom modelo de previsão não ser necessariamente o modelo que melhor se ajusta aos dados da amostra de treino, a avaliação da sua qualidade preditiva (isto é, a avaliação da precisão das suas previsões) só pode ser feita com uma amostra de observações que não tenha sido utilizada na sua estimação, ou seja, com uma amostra de teste (Caiado, 2016).

Desta forma, uma série temporal é geralmente dividida em série de treino, usada para estimar o modelo, e série de teste, para avaliar as previsões. Assim, se se considerar uma série temporal de dimensão n , $\{Y_1, Y_2, \dots, Y_n\}$, e um horizonte temporal de previsão, h , tem-se a seguinte partição da série:

$$\underbrace{Y_1, Y_2, \dots, Y_{n-h}}_{\text{série de treino}}, \underbrace{Y_{n-h+1}, \dots, Y_n}_{\text{série de teste}}.$$

Usualmente, a série de teste corresponde a 20% da série original, apesar deste valor depender da dimensão da amostra e do horizonte temporal pretendido para a previsão futura. É aconselhado escolher uma série de teste de dimensão igual ou superior ao horizonte temporal pretendido.

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

Os erros de previsão indicam se a metodologia de previsão é apropriada, sendo, por isso, importante medir a sua magnitude (Cordeiro, 2011). Um erro de previsão é a diferença entre o valor observado e a sua previsão. De forma geral, o erro de previsão (a h -passos) pode, então, ser escrito como $e_{t+h} = Y_{t+h} - \hat{Y}_{t+h|t}$. Se se considerar $h = 1$, ou seja, a previsão a 1-passo, o erro de previsão a 1-passo é dado simplesmente por $e_t = Y_t - \hat{Y}_t$. No entanto, os erros de previsão não devem ser confundidos com os resíduos do modelo. Isto porque os resíduos são calculados na série de treino e baseiam-se apenas em previsões a 1-passo, enquanto que os erros de previsão utilizam a série de teste e podem envolver previsões multi-passos (Hyndman & Athanasopoulos, 2014).

4.4.1 Medidas de avaliação utilizadas

Não existe nenhuma medida que seja consensual para ser aplicada a todas as situações de previsão e, por isso, é aconselhada a utilização de vários critérios para avaliar a qualidade preditiva de um modelo ou método de previsão. Desta forma, foram escolhidas cinco medidas para proceder à avaliação da qualidade preditiva dos métodos adotados: uma medida dependente da escala, EQM, e a sua correspondente na mesma escala dos dados, REQM; uma medida percentual, EPAM; uma medida escalada, EEAM; e uma medida alternativa, estatística U de Theil.

Erro quadrático médio

O erro quadrático médio (EQM), já definido na Secção 4.2.5, traduz o valor médio dos desvios ao quadrado entre os valores observados e as previsões para os instantes $1, 2, \dots, n$, ou seja,

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

Trata-se de uma medida dependente da escala dos dados, uma vez que é calculada através do erro de previsão a 1-passo, $e_t = Y_t - \hat{Y}_t$. Muitas vezes, a raiz do erro quadrático médio, $\text{REQM} = \sqrt{\text{EQM}}$, é preferida ao EQM porque permite reduzir a grandeza dos valores para a mesma escala dos dados.

Na comparação entre métodos de previsão, e uma vez que se $Y_t = \hat{Y}_t$ se obtém $\text{EQM} = 0$, considera-se que o método mais preciso é o que apresenta menor EQM (e, conseqüentemente, menor REQM).

Historicamente, a REQM e o EQM são medidas bastante utilizadas. No entanto, estas medidas são mais sensíveis a *outliers* do que outras do mesmo tipo, como, por

exemplo, o erro absoluto médio

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

Erro percentual absoluto médio

O erro percentual absoluto médio, EPAM, representa a percentagem média do erro de previsão em relação à grandeza das observações, sendo, portanto, uma medida de erros percentuais, que se define como

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

Tal como acontece no EQM, se $Y_t = \hat{Y}_t$ tem-se $\text{EPAM} = 0\%$ e, portanto, quanto menor for o EPAM mais preciso é o método de previsão.

Apesar de possibilitar a comparação da precisão entre diversas séries, esta medida não pode ser calculada quando existem zeros na série. Além disso, quando as observações se aproximam de zero, o EPAM apresenta valores extremos. De forma a contornar estes problemas foi introduzida uma outra medida, o erro percentual absoluto médio simétrico (EPAMS), definido por (Caiado, 2016)

$$\text{EPAMS} = \frac{1}{n} \sum_{t=1}^n \left| \frac{Y_t - \hat{Y}_t}{Y_t + \hat{Y}_t} \right| \times 200 \quad (\%).$$

No entanto, mesmo solucionando os inconvenientes da utilização do EPAM, o EPAMS tem sido criticado por alguns autores, nomeadamente Hyndman & Koehler (2006), por produzir erros extremos na presença de zeros na amostra.

Erro escalado absoluto médio

O erro escalado absoluto médio, EEAM, proposto por Hyndman & Koehler (2006), permite a comparação relativa de qualquer método de previsão com o método *naïve* (referência), seguindo, de certa forma, a ideia por detrás das medidas de erros relativos. O EEAM define-se como

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

Nesta medida, o erro escalado, q_t , é o quociente entre o erro de previsão, $e_t = Y_t - \hat{Y}_t$, e o erro absoluto médio da previsão *naïve*, $\text{EAMN} = \sum_{t=2}^n |Y_t - Y_{t-1}| / (n-1)$.

Na presença de séries de dados com sazonalidade, o denominador do erro escalado deverá ser substituído pelo erro absoluto médio da previsão *naïve* sazonal, $EAMNs = \sum_{t=s+1}^n |Y_t - Y_{t-s}| / (n - s)$, onde $s = 12$ para dados mensais (Caiado, 2016).

O EEAM pode ser utilizado para comparar a qualidade preditiva de séries com diferentes escalas, uma vez que a grandeza q_t é independente da escala dos dados. Além disso, esta medida é sempre definida e finita, exceto no caso em que todas as observações são iguais.

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

Estatística U de Theil

A estatística U de Theil combina duas características presentes em algumas medidas: por um lado, possibilita a comparação relativa de qualquer método de previsão com o método *naïve* (referência), o que remete para o que é feito nas medidas de erros relativos; e por outro, à semelhança do que acontece no EQM, os erros são elevados ao quadrado e, portanto, é dado maior peso aos grandes erros em detrimento dos mais pequenos. Matematicamente, esta estatística é definida como (Makridakis et al., 1998)

$$U = \sqrt{\frac{\sum_{t=1}^{n-1} (f_{t+1} - a_{t+1})^2}{\sum_{t=1}^{n-1} (a_{t+1})^2}}, \quad (4.47)$$

onde $f_{t+1} = (\hat{Y}_{t+1} - Y_t) / Y_t$ e $a_{t+1} = (Y_{t+1} - Y_t) / Y_t$, que correspondem às mudanças relativas prevista e real, respetivamente. Substituindo na expressão (4.47) os valores de f_{t+1} e a_{t+1} obtém-se, numa versão simplificada,

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

Analisando a expressão da estatística U, podem retirar-se as seguintes ilações:

- $U = 0$ se $f_{t+1} = a_{t+1}$, para $t = 1, 2, \dots, n - 1$, i.e., quando as previsões são exatas;
- $U = 1$ se $f_{t+1} = 0$, i.e., quando os erros de previsão do método adotado são

iguais aos erros obtidos pelo método *naïve*. Neste caso, pode dizer-se que o método adotado e o método *naïve* são igualmente precisos;

- $U > 1$ se f_{t+1} se encontra na direção oposta (sinais diferentes) de a_{t+1} , uma vez que o numerador será maior do que o denominador. Neste caso, o método de previsão adotado é menos preciso do que o método *naïve* e, portanto, não deve ser o escolhido;
- $U < 1$ se f_{t+1} se encontra na mesma direção (sinais iguais) de a_{t+1} . Assim, o método de previsão adotado é mais preciso do que o método *naïve* e, portanto, deve ser o escolhido. Quanto menor o valor de U , melhor é o método de previsão relativamente ao método *naïve*.

Tal como acontece no EEAM, também no caso da estatística U se considera que, na comparação entre diferentes métodos de previsão, o mais preciso é o que apresenta o valor de U mais baixo, desde que o método de referência utilizado seja sempre o mesmo (neste caso, o método *naïve*, cuja previsão corresponde à última observação disponível).

Capítulo 5

Aplicação dos métodos de previsão a dados do segmento do retalho

Os primeiros indícios da crise económica e financeira global na União Europeia (EU-28) surgem em 2008, pela redução considerável na taxa de crescimento real do produto interno bruto (PIB), mas foi em 2009 que se verificou uma grave recessão, caracterizada por uma queda de 4,3% do PIB real da União Europeia (ver Tabela 5.1). A recuperação foi vislumbrada em 2010, com uma taxa de crescimento do PIB de 2,1%, seguida de um aumento adicional de 1,8% em 2011. No entanto, houve ainda uma contração do PIB em 2012 (-0,4%), que rapidamente foi ultrapassada, como se comprova pelas taxas de variação progressivamente crescentes em 2013 (+0,3%), 2014 (+1,8%) e 2015 (+2,3%). Em 2016, o crescimento continuou, apesar de a um ritmo mais lento (+2,0%), e em 2017 a série anterior de crescimento regressou, refletindo-se num aumento do PIB de 2,4%, a maior taxa de variação anual desde o início da crise.

Na União Europeia, a taxa de crescimento real do PIB variou consideravelmente, tanto ao longo do tempo como entre os seus 28 Estados-Membros. Após uma queda do PIB em todos os Estados-Membros, com exceção da Polónia, em 2009, o crescimento económico recomeçou em 23 dos países em 2010 (manteve-se inalterado em Espanha). Em 2012, o desenvolvimento económico sofreu algumas alterações, uma vez que apenas metade (14) dos Estados-Membros conseguiu uma taxa de crescimento do PIB positiva, registando-se uma diminuição do PIB nos restantes (com exceção da Bulgária, que não sofreu alterações). A partir de 2013, a maioria dos Estados-Membros voltou a registar uma taxa de variação positiva: 17 países em 2013, 25 em 2014 e 27 em 2015 e 2016 (sendo a Grécia a única exceção). Em 2017, pela primeira vez desde 2007, ou seja, desde o início da crise, todos os Estados-Membros

registaram uma taxa de variação positiva e, portanto, crescimento económico.

Tabela 5.1: Taxa de crescimento real do PIB (%) na União Europeia e em alguns Estados-Membros, nos anos de 2007 a 2017.

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
EU-28	+3,1	+0,5	-4,3	+2,1	+1,8	-0,4	+0,3	+1,8	+2,3	+2,0	+2,4
DE	+3,3	+1,1	-5,6	+4,1	+3,7	+0,5	+0,5	+2,2	+1,7	+2,2	+2,2
ES	+3,8	+1,1	-3,6	+0,0	-1,0	-2,9	-1,7	+1,4	+3,6	+3,2	+3,0
FR	+2,4	+0,3	-2,9	+1,9	+2,2	+0,3	+0,6	+1,0	+1,1	+1,2	+2,2
IT	+1,5	-1,1	-5,5	+1,7	+0,6	-2,8	-1,7	+0,1	+0,9	+1,1	+1,6
NL	+3,8	+2,2	-3,7	+1,3	+1,6	-1,0	-0,1	+1,4	+2,0	+2,2	+2,9
PT	+2,5	+0,2	-3,0	+1,9	-1,8	-4,0	-1,1	+0,9	+1,8	+1,9	+2,8
UK	+2,5	-0,3	-4,2	+1,7	+1,6	+1,4	+2,0	+2,9	+2,3	+1,8	+1,7

Fonte: Eurostat.

Comércio na União Europeia

O comércio constitui uma fase intermediária entre a produção de bens ou serviços e a do seu consumo ou utilização, tendo por função levar as mercadorias até aos consumidores. Distinguem-se, frequentemente, dois tipos de comércio, que correspondem a duas etapas do processo de distribuição dos produtos: o comércio por grosso e o comércio a retalho. O primeiro grupo corresponde à atividade de compra e venda (por grosso), de bens novos ou usados, a comerciantes (retalhistas ou grossistas), a industriais, a utilizadores institucionais e profissionais ou a intermediários, ou seja, a entidades cujo objetivo é a revenda e não o consumo próprio. Por oposição, o comércio a retalho constitui a atividade de revenda (a retalho) de bens destinados ao público em geral, empresas e outras instituições para consumo ou uso pessoal ou doméstico, ou seja, a compradores que se comportam como consumidores finais. Desta forma, o retalhista funciona, muitas vezes, como último elo de ligação de uma cadeia de distribuição, permitindo que o produto (ou bem) chegue ao consumidor final. Além destes tipos de comércio, pode, ainda, destacar-se um terceiro tipo, que envolve volumes de vendas mais elevados e esporádicos: o comércio automóvel. Estes três tipos de comércio são mutuamente exclusivos, ou, por outras palavras, o volume de vendas relacionado com o comércio automóvel não é contabilizado em nenhum dos outros dois tipos de comércio.

O setor do comércio é um setor fundamental para a Economia Europeia. De acordo com o Eurostat, em 2015, este setor empregou cerca de 33 milhões de pessoas e representou 9,9% do valor acrescentado bruto total da União Europeia. Em termos

de volume de negócios, o setor produziu, em 2016, cerca de 9,9 triliões de euros, donde 57,8% corresponderam ao comércio por grosso (−1% face a 2015), 28,9% ao comércio a retalho (tal como em 2015) e os restantes 13,3% ao comércio automóvel (+1%).

O comércio a retalho é um setor dinâmico e complexo, que engloba diversos tipos de empresas, cuja estrutura reflete as características culturais da sociedade em que se insere, sofrendo o impacto de toda a evolução sociológica, económica e tecnológica. Assim, torna-se fundamental, do ponto de vista económico, o estudo do setor do comércio a retalho de forma a avaliar o desenvolvimento de uma economia nacional. De facto, além de ter representado quase um terço do volume de negócios do setor do comércio em 2015, o comércio a retalho deteve 58,4% das empresas comerciais e empregou 8,7% da população ativa da União Europeia, o que corresponde a aproximadamente 18,8 milhões de postos de trabalho.

O volume de negócios do comércio a retalho europeu foi, em 2016, de aproximadamente 2,9 triliões de euros. Segundo este mesmo indicador, os mercados mais importantes da União Europeia foram a Alemanha (cerca de 537,5 biliões de euros), o Reino Unido (aproximadamente 480,3 biliões de euros) e a França (cerca de 440,9 biliões de euros). Além desses, também a Itália e a Espanha tiveram algum destaque, com valores da ordem dos 300 e 200 biliões de euros, respetivamente. Desses principais mercados, à exceção da Alemanha (−4,9%) e do Reino Unido (−7%), todos registaram um aumento do volume de negócios a retalho em 2016. Apesar de, em termos nominais, serem bastante inferiores, os mercados mais promissores foram o Luxemburgo, com um crescimento do volume de negócios de 11,1%, e a Roménia, com 8,8%. A distribuição detalhada do volume de negócios europeu, em 2016, pelos 28 Estados-Membros encontra-se representada na Figura 5.1.

A situação económica provocada pela crise de 2008 teve, como seria de esperar, consequências no setor do comércio, influenciando indicadores como o volume de negócios (ver Tabela 5.2). De facto, em 2009 registou-se uma redução de mais de 5% no volume de negócios europeu, seguida de uma melhoria da mesma dimensão no ano seguinte. Daí em diante, o volume de negócios não voltou a atingir valores tão baixos e apresentou uma evolução bastante positiva, registando quedas mínimas (−0,9%) somente em 2013 e 2016.

Comércio em Portugal

De acordo com dados do Eurostat e do Instituto Nacional de Estatística (INE), em 2016, o volume de negócios gerado pelas empresas de comércio em Portugal

situou-se em 128,1 biliões de euros, representando 35% do total de volume de negócios considerado no Sistema de Contas Integradas das Empresas (34,4% em 2015). No conjunto do setor do comércio, destaca-se o comércio por grosso, que concentrou 49% do volume de negócios do setor (-1,3% face a 2015), seguindo-se o comércio a retalho com 36,2% (tal como em 2015) e o comércio automóvel com 14,8% (+1,3%).

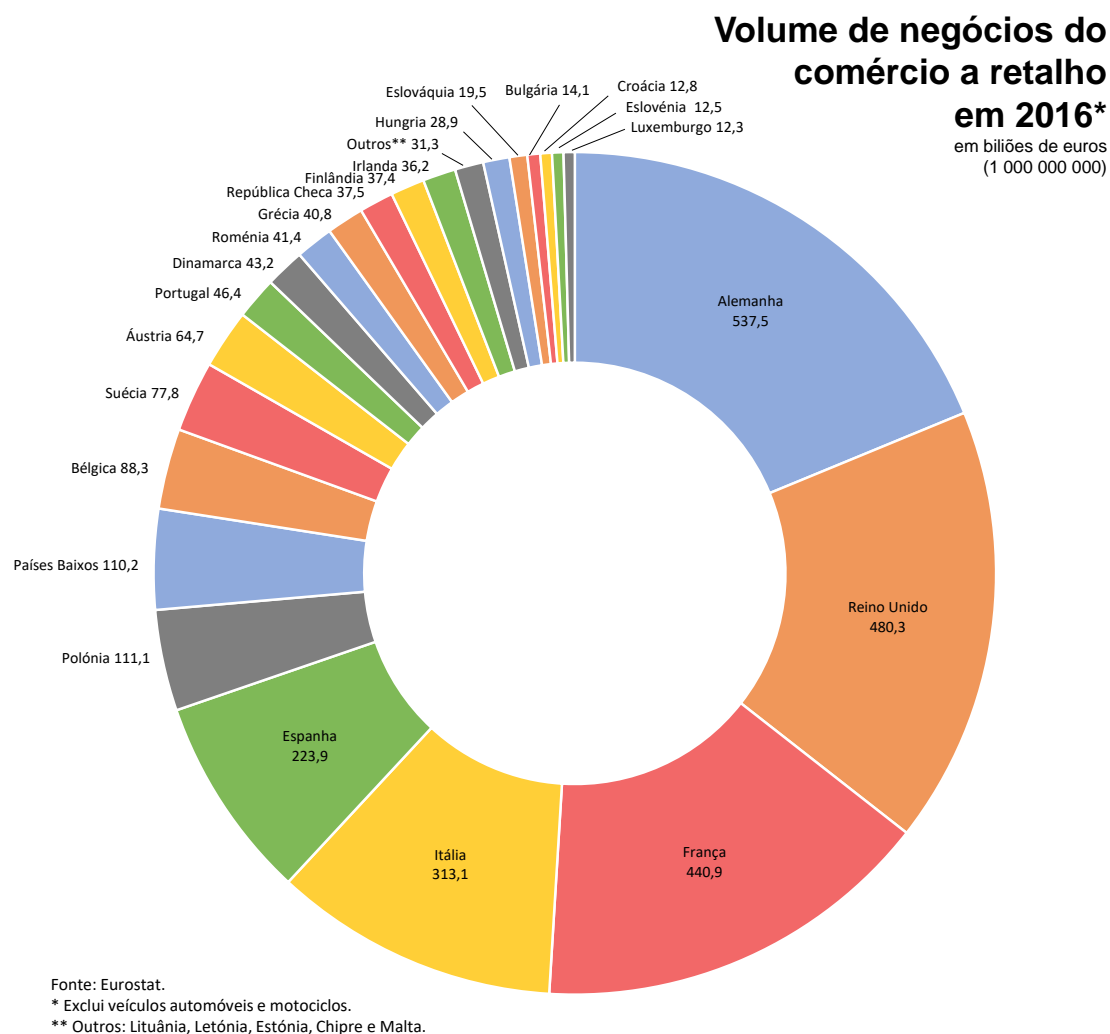


Figura 5.1: Distribuição do volume de negócios do comércio a retalho pelos 28 Estados-Membros da União Europeia (em biliões de euros), no ano de 2016.

No que respeita ao setor do comércio a retalho, este apresentou, em 2016, evoluções positivas em todos os principais indicadores, incluindo no pessoal ao serviço (+1,9%), que atingiu 431,3 mil trabalhadores (+2,7% em 2015), e no volume de negócios (+3,3%, após +2,0% em 2015).

No entanto, o cenário nem sempre foi tão positivo. Portugal foi, à semelhança dos outros países da União Europeia, afetado pela crise económica em 2009 (ver Tabela

5.1), apresentando uma taxa de crescimento do PIB negativa nesse ano ($-3,0\%$). A situação económica melhorou no ano seguinte, mas, em 2011, houve novamente uma diminuição da produção, que se prolongou até 2013, correspondendo o ano de 2012 ao pior ano da crise económica em Portugal (taxa de variação de $-4,0\%$). A partir de 2014, o cenário económico de Portugal melhorou, registando-se um aumento progressivo da taxa de crescimento do PIB de ano para ano: $+0,9\%$ em 2014, $+1,8\%$ em 2015, $+1,9\%$ em 2016 e $+2,8\%$ em 2017 (o maior crescimento económico desde 2007).

Obviamente, também o setor do comércio a retalho foi abalado pela crise económica de 2008 (ver Tabela 5.2). De facto, as vendas no retalho aumentaram de 2007 para 2008 ($+3,2\%$) mas, com a crise financeira, em 2009, as vendas caíram ($-3,5\%$). O melhoramento na Economia do país em 2010 fez com que o volume de negócios voltasse a aumentar neste setor, contudo não se mostrou suficiente para suplantar os valores anteriormente registados. Com a redução da produção em 2011, o volume de negócios diminuiu $3,7\%$, seguindo-se uma queda adicional de $4,5\%$ em 2012 (o pico da crise em Portugal). A crise neste setor foi superada em 2014, com um aumento de $2,0\%$ no volume de vendas no retalho nesse ano e no ano seguinte.

Tabela 5.2: Volume de negócios do comércio a retalho (em biliões de euros) na União Europeia e em alguns Estados-Membros, nos anos de 2007 a 2016.

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
EU-28	2500,0*	2597,1*	2462,8*	2592,0*	2683,7	2732,3	2708,1	2796,2	2879,9	2853,7
DE	402,8	420,2	460,5	474,4	503,8	498,4	491,7	526,4	565,3	537,5
ES	240,4	240,7	219,5	221,4	217,3	210,5	204,9	204,3	213,1	223,9
FR	394,5	421,2	403,5	419,3	434,2	447,7	446,0	443,4	430,5	440,9
IT	304,6	315,7	282,5	312,6	315,9	314,0	307,7	306,2	310,0	313,1
NL	97,4	102,2	99,8	102,0	104,4	104,8	102,3	102,8	107,5	110,2
PT	46,1	47,6	45,9	47,2	45,4	43,4	43,1	44,0	44,9	46,4
UK	448,5	395,2	354,6	385,9	393,2	429,9	422,9	460,4	516,3	480,3

Fonte: Eurostat.

*EU-27: antes da adesão da Croácia.

E-commerce

A introdução da *internet* no quotidiano da população europeia e o seu constante desenvolvimento deram origem a novas práticas de consumo, nomeadamente ao comércio eletrónico (*e-commerce*). De facto, embora a maioria das transações ainda ocorra *offline* (em termos nominais, o *e-commerce* representou apenas $4,8\%$ do volume de negócios a retalho europeu em 2015), o retalho *online* ganhou mais e mais

destaque nos últimos anos (ver Tabela 5.3). Esta evolução é visível por toda a União Europeia. Por exemplo, na Alemanha o comércio eletrónico foi sempre crescendo, uns anos de forma mais acentuada do que outros, chegando a apresentar um aumento de 40,5% em 2015. Outros países, como a Itália e os Países Baixos, apesar de afetados pela crise económica em 2012, com quedas nos volumes de negócios, também apresentaram crescimentos acentuados, nomeadamente em 2015, com taxas de variação de 58,3% e 63,2%, respetivamente. Também em Portugal, depois do efeito da crise ter sido atenuado, o volume de negócios do *e-commerce* registou uma grande evolução, com aumentos de 10,5% em 2014, 16,9% em 2015 e 20,1% em 2016.

Tabela 5.3: Volume de negócios do *e-commerce* no comércio a retalho (% de variação anual) na União Europeia e em alguns Estados-Membros, nos anos de 2009 a 2016.

	2009	2010	2011	2012	2013	2014	2015	2016
EU-28	+3,8	NA	NA	NA	NA	NA	NA	NA
DE	+0,4	+17,4	+15,3	+14,0	+4,9	+16,2	+40,5	+13,8
ES	-18,9	+2,5	+2,3	+0,7	+26,8	+34,5	+36,7	+51,9
FR	NA	+12,4	+6,5	+9,5	-4,8	+4,6	+6,9	+2,4
IT	+12,8	+11,5	+45,0	-10,9	+20,6	+9,3	+58,3	+24,9
NL	+42,0	+4,2	+12,7	-9,6	+20,0	-1,1	+63,2	+14,4
PT	+2,6	+4,3	-1,3	-9,8	+2,8	+10,5	+16,9	+20,1
UK	+1,3	+9,3	+16,5	+18,9	+10,8	+22,2	+14,3	-30,2

Fonte: Eurostat.

Em 2016, pelo menos 22 países apresentaram taxas de variação positivas, metade delas ultrapassando os 20%. Assim, pode dizer-se que o futuro do retalho passa pelo *e-commerce* e, por isso, o seu estudo é bastante pertinente.

5.1 Apresentação da base de dados

Os dados utilizados neste estudo, disponíveis na plataforma *online* do Eurostat¹, correspondem a índices mensais do volume de negócios (total) em dois contextos distintos do comércio a retalho – o comércio a retalho generalizado, com exceção apenas dos veículos automóveis e motociclos, e a venda a retalho por correspondência ou via *internet* (uma subcategoria) –, recolhidos entre janeiro de 2000 e fevereiro de 2018. Por se tratarem de índices, devem sempre ser comparados com o ano de base, neste caso, 2015, que corresponde ao índice 100.

¹Dados recolhidos a 4 de maio de 2018.

O índice de volume de negócios (TOVT) tem como objetivo mostrar a evolução do mercado de bens e serviços. O volume de negócios corresponde à quantia líquida das vendas e prestações de serviços (abrangendo as indemnizações compensatórias) respeitantes às atividades normais das entidades, conseqüentemente após as reduções em vendas e não incluindo nem o imposto sobre o valor acrescentado (IVA) nem outros impostos diretamente relacionados com as vendas e prestações de serviços.

De acordo com a PORDATA (Base de Dados de Portugal Contemporâneo), os principais parceiros comerciais de Portugal (PT) dentro da União Europeia são a Alemanha (DE), a Espanha (ES), a França (FR), a Itália (IT), os Países Baixos (NL) e o Reino Unido (UK), tanto a nível de importações como de exportações de bens e serviços, e, por esse motivo, esses foram os países escolhidos para efeitos comparativos.

Assim sendo, foram recolhidas no total 14 séries de índices de volumes de negócios distintas, isto é, duas séries para cada país escolhido, em que uma diz respeito ao comércio a retalho (exceto veículos automóveis e motociclos) e a outra à venda por correspondência ou via *internet* (*e-commerce*).

Em termos de ajustamento da sazonalidade, escolheram-se os dados que não sofreram qualquer tipo de transformação (*unadjusted data*), isto é, os dados brutos, uma vez que se pretende que os dados sejam o menos modificados possível. Além disso, de acordo com Chu & Zhang (2003), o ajustamento da sazonalidade pode resultar em distorções severas dos dados, dando origem a comportamentos não lineares, e num pior desempenho das previsões, pelo que tipicamente é recomendado o uso de *seasonally unadjusted data*.

Depois de reunidos, os dados foram, então, divididos em dois conjuntos, um de treino (ou estimação) e um de teste (ou previsão), a fim de avaliar a precisão dos diferentes modelos de previsão implementados. O período de treino selecionado foi de janeiro de 2000 a dezembro de 2016 (primeiras 204 observações), e o período de teste foi de janeiro de 2017 a fevereiro de 2018 (últimas 14 observações).

5.2 Análise descritiva

Como se sabe, a análise exploratória dos dados é um processo quase que imprescindível na Estatística, constituindo o ponto de partida de qualquer estudo. Nesta Secção, pretende-se realizar uma análise inicial dos dados, onde se calculam as estatísticas descritivas básicas das duas séries disponíveis para cada um dos sete países. O objetivo principal é avaliar o comportamento do índice de volume de negócios

em diferentes contextos: em duas diferentes categorias do comércio a retalho e sete países distintos.

Nas Figuras 5.2 e 5.3 encontram-se representadas as 14 séries em estudo. Comparando as séries do comércio a retalho com as do *e-commerce*, facilmente se observa que as últimas apresentam comportamentos mais díspares (entre países) do que as primeiras. Além disso, apesar de, em alguns casos, as séries do comércio a retalho apresentarem alterações na sua variabilidade, essas não são tão visíveis como nas séries do *e-commerce*. Tal pode ser explicado pelo facto de o *e-commerce* ser um tipo de comércio a retalho emergente, ainda em desenvolvimento, o que faz com que seja mais irregular e varie bastante de país para país.

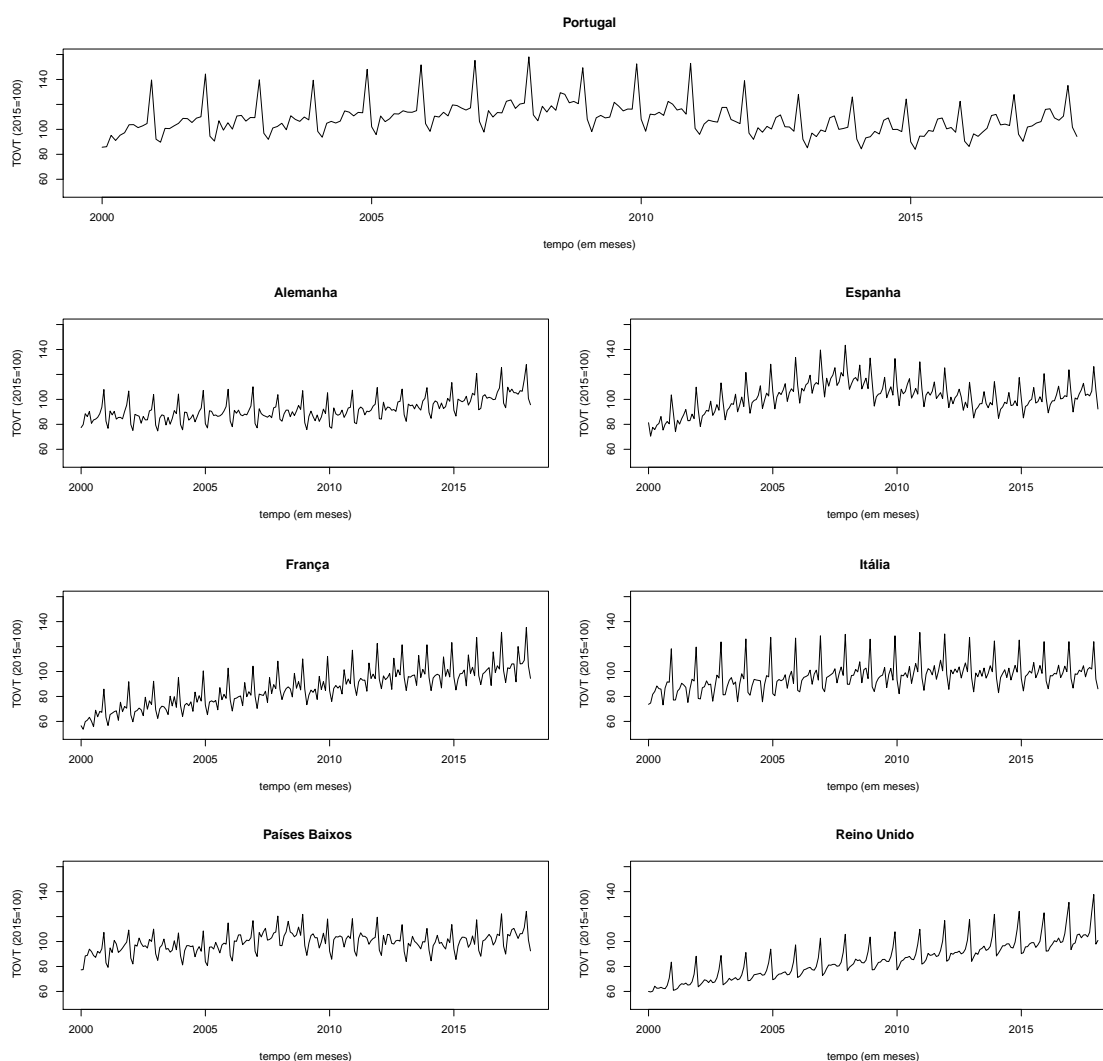


Figura 5.2: Representação gráfica das séries do comércio a retalho, com exceção de veículos automóveis e motocicletas, para os sete países em estudo, no período observado.

Complementando a informação gráfica com as correspondentes medidas descritivas (disponíveis para consulta no Apêndice B), pode observar-se que, no caso das séries do comércio a retalho, há uma forte indicação da presença de uma componente sazonal, sendo o mês de dezembro o pico do volume de negócios em qualquer um dos países em estudo (o que pode ser justificado pelas festividades dessa época). Além disso, algumas séries apresentam tendências bastantes semelhantes entre si, nomeadamente em países como a França e o Reino Unido, que são crescentes em todo o domínio temporal, e Portugal e Espanha, que apenas apresentam um período decrescente entre 2009 e 2014 (aproximadamente). As séries que apresentam menor variabilidade são as dos Países Baixos, Itália e Alemanha, possivelmente devido ao comportamento algo constante sugerido pela análise gráfica (alguns períodos que, em média, são constantes (Tabela B.3, Apêndice B)). Para as séries do comércio a retalho, os desvios padrão variam entre 8,39 e 15,60 (Tabela B.1, Apêndice B).

Como seria de esperar, a média de 2015 ronda o índice 100, independentemente do tipo de comércio e/ou do país em questão. No entanto, o comportamento à volta desse índice e, conseqüentemente, do ano base, varia bastante (Tabela B.3, Apêndice B). Por exemplo, em Portugal, cerca de 75% das observações estão acima da média de 2015, o que indica que os volumes de negócios são, usualmente, superiores aos registados nesse ano. Em contrapartida, em países como a Alemanha, a França, o Reino Unido e a Itália, pelo menos 75% das observações são inferiores ao índice 100. Já em Espanha e nos Países Baixos, os volumes de negócios são, de forma equiparada, tanto inferiores como superiores aos registados no ano base.

No que respeita às séries do *e-commerce*, tal como referido anteriormente, o comportamento varia muito de país para país. De facto, apesar de, na maioria dos países, se notar um elevado crescimento nos últimos anos, antecedido de um período mais ou menos constante, países como Portugal e Itália apresentam comportamentos bastante irregulares (Figura 5.3). Em termos de variabilidade, por exemplo, os países mais regulares apresentam uma variabilidade crescente com a tendência, enquanto que a série de Portugal apresenta uma forte dispersão em alguns anos, que vai reduzindo ao longo do tempo (volatilidade económica). A série da Itália é um caso especial, em que, apesar da tendência aparentar uma linearidade crescente, a variância vai sofrendo algumas alterações ao longo do tempo, não dependentes da componente tendência. De facto, parecem existir várias variâncias na série, ainda que localmente constantes.

Além disso, nas séries do *e-commerce*, a componente sazonal não é tão explícita nem tão regular entre países (Tabela B.4, Apêndice B). Por exemplo, nos Países

Baixos e no Reino Unido os volumes de negócios mantêm-se aproximadamente constantes, em média, nos meses de janeiro a outubro, enquanto que nos restantes países se nota uma diminuição do volume de negócios nos meses de julho e, essencialmente, de agosto. De forma geral, os meses com volumes de negócios mais elevados são novembro e dezembro, o que se pode justificar pelas festividades dessa época e pela ocorrência da *Cyber Monday*² em novembro.

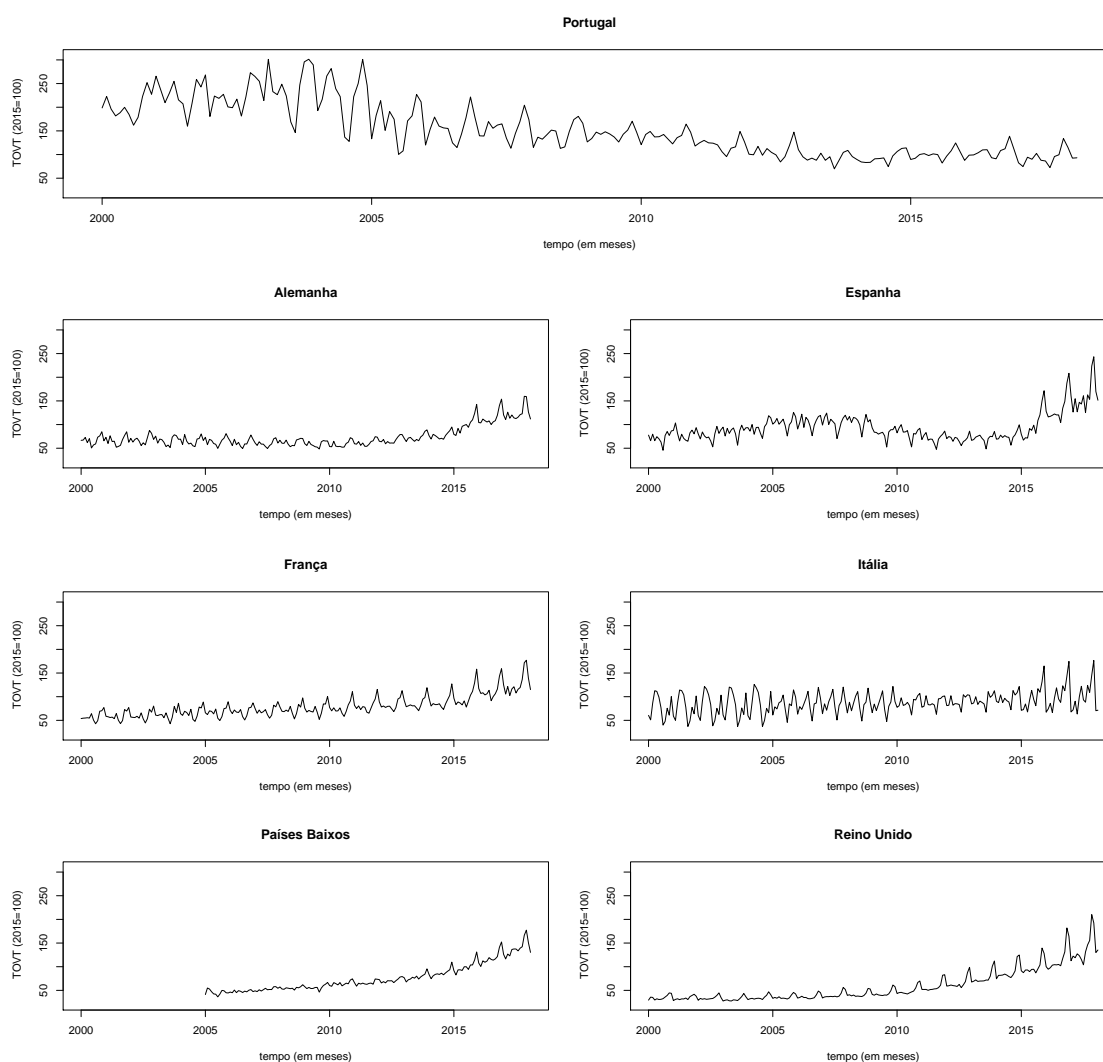


Figura 5.3: Representação gráfica das séries da venda a retalho por correspondência ou via *internet*, para os sete países em estudo, no período observado.

À semelhança do que acontece nas séries do comércio a retalho, Portugal apresenta mais de 75% das observações acima da média do ano base. Em contrapartida, em países como a Alemanha, a França, o Reino Unido e os Países Baixos, pelo menos

²Trata-se de uma extensão da *Black Friday*, direcionada apenas para o comércio *online*.

75% das observações são inferiores ao índice 100, o que indica que os volumes de negócios registados foram, na sua maioria, inferiores aos de 2015. Obviamente, esta última afirmação é imediata se se tiver em consideração que as séries apresentam tendências essencialmente crescentes pelo menos nos últimos quatro ou cinco anos.

A série do *e-commerce* nos Países Baixos é a única que apresenta valores em falta, começando apenas em janeiro de 2005 (Tabela B.2, Apêndice B).

Naturalmente, este estudo tem especial interesse no caso particular das séries de Portugal.

Relativamente à série do comércio a retalho em Portugal, a representação gráfica da Figura 5.2 sugere uma tendência decrescente no volume de negócios entre 2009 e 2014 (aproximadamente), isto é, durante o período da crise económica em Portugal, sendo crescente no restante domínio temporal.

Como já referido, esta série apresenta uma componente sazonal bem evidente. De facto, pela observação da Figura 5.4, é possível notar que o pico do consumo, em termos de volume de negócios, é o mês de dezembro (tal como acontece nos outros países em estudo, talvez devido às festividades da época). Quando em comparação com os restantes, pode ver-se que os meses de julho e agosto apresentam consumos ligeiramente superiores, possivelmente por corresponderem aos principais meses escolhidos para férias, e que se registam menores volumes de negócios em janeiro e, essencialmente, em fevereiro. Nos restantes meses, os volumes de negócios medianos mantêm-se aproximadamente constantes.

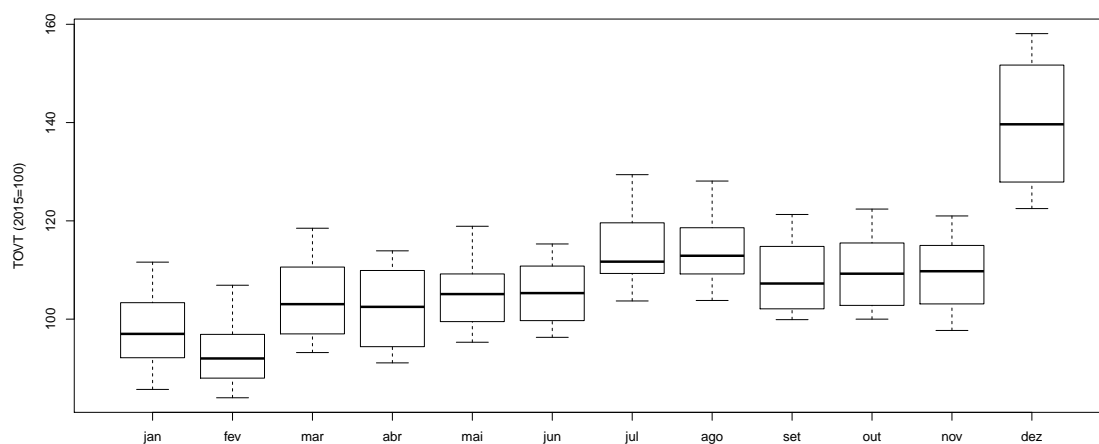


Figura 5.4: Diagramas em caixa de bigodes para a série do comércio a retalho em Portugal, dividida pelos doze meses do ano civil, no período observado.

O comportamento da série do *e-commerce* em Portugal não é tão regular (Figura 5.3). Na verdade, as observações dos primeiros anos (2000-2005), além de se caracterizarem por uma forte dispersão (volatilidade económica), correspondem a volumes de negócios que chegam a ser mais do dobro dos registados em 2015 (talvez justificados pela disponibilização da *internet* para o público em geral). Depois desse período, o consumo por correspondência ou via *internet* diminui de forma gradual, tornando-se aproximadamente constante a partir de 2012.

Ao contrário do que acontece em relação à variabilidade, que é bastante mais acentuada do que na série do comércio a retalho, na série do *e-commerce* em Portugal a existência de sazonalidade não é tão notória (Figura 5.5). De facto, com exceção dos meses de julho e agosto, que se destacam pela negativa, os volumes de negócios medianos mantêm-se aproximadamente constantes para todos os outros meses. No mês de novembro o consumo por correspondência ou via *internet* é ligeiramente superior, possivelmente devido à existência da *Cyber Monday*.

Na Figura 5.5 identificam-se três observações *outliers*, mais concretamente nos meses de janeiro de 2001, de julho de 2001 e de julho de 2002.

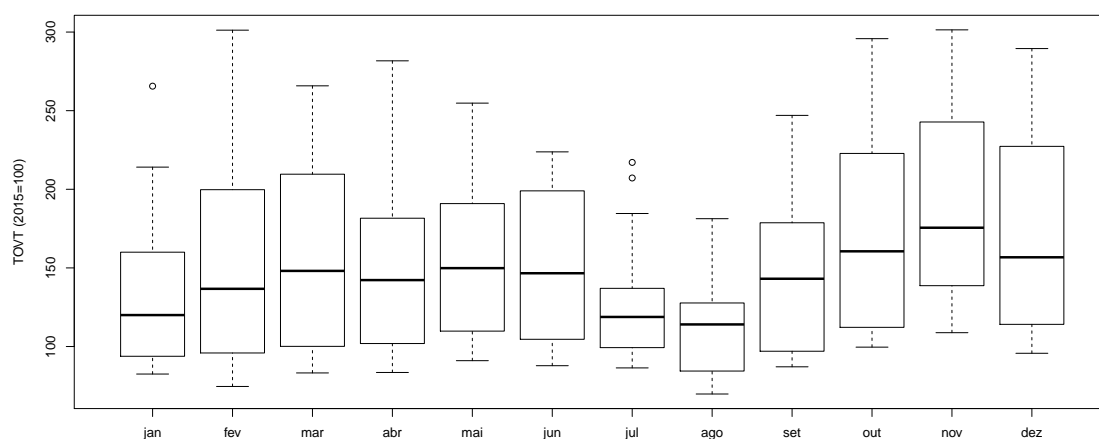


Figura 5.5: Diagramas em caixa de bigodes para a série da venda a retalho por correspondência ou via *internet* em Portugal, dividida pelos doze meses do ano civil, no período observado.

5.3 Aplicação dos métodos de previsão

Depois de se fazer uma análise exploratória às séries em estudo, o próximo passo é a aplicação das metodologias aos dados, que, neste caso, correspondem a modelos/métodos de previsão para séries temporais. Como referido no Capítulo 2, os estudos

no segmento do retalho utilizam, essencialmente, quatro metodologias: o método de Holt-Winters, o modelo ARIMA de Box-Jenkins, o modelo de regressão linear múltipla e os modelos de redes neuronais. Para efeitos comparativos, escolheram-se, de entre as quatro metodologias, as mais tradicionais (descritas, em pormenor, no Capítulo 4), deixando de parte o estudo dos modelos de redes neuronais. Esta escolha foi reforçada por Chatfield (2000), que refere que “as afirmações acerca da superioridade dos modelos de redes neuronais parecem exageradas”, uma vez que “muitas séries temporais exibem características não lineares, mas os modelos não lineares podem não fornecer melhores previsões (*out-of-sample*) do que os modelos lineares, talvez porque os últimos são menos propícios a falhas nos pressupostos”.

A seleção de um modelo SARIMA compreende as três fases da metodologia Box-Jenkins: identificação, estimação e diagnóstico. Para a identificação de um modelo SARIMA, o primeiro passo é a estacionarização da série, através de transformações tanto em relação à média como à variância. Desta forma, e como a estabilização da variância deve ser feita em primeiro lugar, procede-se à transformação logarítmica dos dados³. Quanto à ordem de diferenciação, para estabilização da média, esta é escolhida com base na análise gráfica (da série e das FAC e FACP empíricas) e nos testes de estacionariedade (ADF e KPSS).

Depois de se transformar a série em estudo numa série estacionária, identificando a ordem de diferenciação regular, d , é preciso estimar a componente sazonal. Para tal, começa-se por determinar o período sazonal, s , através da análise das FAC e FACP da série estacionária, e, de seguida, estimam-se vários modelos fazendo variar P , D e Q (as ordens da parte sazonal), usualmente entre 0 e 1. A escolha das ordens P , D e Q tem em consideração a significância dos parâmetros associados e o critério AIC. No entanto, é de salientar que, no caso do intervalo de confiança para o parâmetro associado a $P = 1$ incluir o valor 1, se deve optar por uma diferenciação sazonal, isto é, pelo modelo que considera $P = 0$ e $D = +1$.

Por fim, resta identificar as ordens p e q , comparando o comportamento das FAC e FACP empíricas com o das FAC e FACP teóricas (ver Tabela 4.1, Secção 4.1.3). De forma a realizar uma escolha mais cuidada, devem ser sempre explorados modelos “vizinhos”, sendo estes analisados tanto em relação à significância dos parâmetros como ao comportamento dos resíduos. Quando em dúvida, a escolha entre dois ou mais modelos SARIMA baseia-se no critério AIC, tendo sempre em mente que, se os AICs diferem em apenas duas unidades (2), se escolhe o modelo mais parcimonioso (com menor número de parâmetros).

³As representações gráficas das séries em estudo após esta transformação estão disponíveis no Apêndice A.

No caso dos métodos de alisamento exponencial, e uma vez que é identificada uma componente sazonal em todas as séries em estudo, procede-se à estimação dos dois modelos de Holt-Winters, aditivo e multiplicativo, e escolhe-se aquele que apresenta o menor erro quadrático médio das previsões a 1-passo.

Uma vez que estes métodos são não paramétricos, o cálculo dos intervalos de previsão requer alguns cuidados, sendo necessário determinar o erro quadrático médio a h -passos para cada horizonte temporal h ($h = 1, 2, \dots, 14$). Além disso, a validade dos intervalos de previsão está dependente da verificação do pressuposto da normalidade, que é testado através do teste de Kolmogorov-Smirnov, para os erros de previsão a 1-passo, a 2-passos, \dots , a 14-passos.

No que respeita aos modelos de regressão linear múltipla, é adotada a metodologia descrita em Alpuim & El-Shaarawi (2009), fazendo variar, exclusivamente, a estimação da componente tendência (T_t) em cada uma das séries. Em cada caso, o modelo completo é escrito como

$$Y_t = T_t + \sum_{i=1}^6 \left[\alpha_i \cos\left(\frac{2\pi it}{12}\right) + \beta_i \sin\left(\frac{2\pi it}{12}\right) \right] + \sum_{i=1}^6 \left[\gamma_i t \cos\left(\frac{2\pi it}{12}\right) + \delta_i t \sin\left(\frac{2\pi it}{12}\right) \right] + \epsilon_t, \quad (5.1)$$

onde ϵ_t é um processo gaussiano estacionário $AR(p)$,

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \dots + \phi_p \epsilon_{t-p} + a_t, \quad (5.2)$$

e a_t um ruído branco e $t = 1, 2, \dots, 204$.

Através do *output* do comando *decompose*, que decompõe a série temporal nas suas componentes (tendência, sazonalidade e aleatoriedade) usando médias móveis, são identificados alguns possíveis modelos para descrever a tendência. Quando a série temporal é, de alguma forma, (visualmente) interrompida por um (ou mais) evento(s), a escolha do(s) ponto(s) de corte resulta da comparação dos vários coeficientes de determinação ajustados provenientes da estimação de diversos modelos que consideram, dentro de um certo intervalo temporal, todos os pontos de corte possíveis.

Após a remoção das variáveis explicativas não significativas dos vários modelos completos, a escolha entre dois ou mais modelos finais é feita com base no coeficiente de determinação ajustado, corrigido de forma a ter em conta os parâmetros autorregressivos (caso se aplique). De salientar que, antes da estimação dos mode-

los de regressão, é aplicada uma transformação logarítmica aos dados de forma a estabilizar a variância e, portanto, Y_t representa o logaritmo do índice de volume de negócios (TOVT) no tempo t na equação (5.1).

As variáveis do modelo de regressão foram codificadas da seguinte forma:

- **t ou tempo**, representa o domínio temporal, codificado de 1 (janeiro de 2000) a 218 (fevereiro de 2018);
- **evento**, representa a variável binária que codifica a ocorrência ou não de um evento (i.e., assume o valor 0 nos tempos pré-evento e 1 a partir da ocorrência do mesmo). Em caso de existir mais do que um evento, as diferentes variáveis são apresentadas, por ordem de ocorrência, como evento1, evento2, etc.. Os eventos são considerados mutuamente exclusivos, ou seja, no caso da ocorrência de dois eventos, o primeiro assume o valor zero a partir do tempo em que ocorre o segundo;
- **cos***, é o cosseno de período *, onde * assume os valores 12; 6; 4; 3; 2,4 e 2. Por exemplo, cos12 é a codificação para o cosseno de período 12;
- **sen***, é o seno de período *, onde * assume os valores 12; 6; 4; 3 e 2,4;
- **t :var**, representa a interação da variável tempo com uma das restantes variáveis. Por exemplo, t :cos12 é a interação da variável tempo com o cosseno de período 12.

Além disso, é também importante referir o (**Intercept**), que representa o valor esperado da variável em estudo quando as variáveis explicativas são simultaneamente iguais a zero.

A estimação dos modelos foi realizada através do ambiente R, onde, além da utilização de funções já existentes, foi necessária a implementação de algumas novas⁴.

A análise dos resíduos, quando aplicável, pretende, essencialmente, verificar se o comportamento dos resíduos se aproxima ao de um ruído branco. Desta forma, é esperado que estejam garantidas, além da média nula e variância constante, a normalidade e a independência dos erros.

De forma a avaliar os pressupostos de média nula e variância constante, a representação gráfica dos resíduos pode ser útil. Além disso, para verificar a condição exigida à média, caso os pressupostos de independência e normalidade dos erros não sejam rejeitados, pode recorrer-se ao teste t para o valor médio.

⁴Os detalhes sobre as funções utilizadas podem ser consultados no Apêndice D.

A condição de normalidade pode ser avaliada através da análise do histograma dos resíduos, que se deve aproximar do comportamento da função densidade de uma distribuição Normal. No entanto, para uma verificação mais rigorosa, podem utilizar-se testes estatísticos, sendo o mais comum, para amostras de grandes dimensões, o teste de Kolmogorov-Smirnov, cuja hipótese nula é a normalidade dos erros.

Para a avaliação do pressuposto da independência dos erros são utilizadas diferentes metodologias dependendo do método de previsão aplicado. De facto, quando se estima um modelo SARIMA, recomenda-se a utilização de um teste de Portman-teau, sendo um dos mais utilizados o teste de Ljung-Box, que testa se as primeiras k autocorrelações são simultaneamente nulas. Desta forma, em caso de rejeição da hipótese nula conclui-se que o modelo escolhido não é apropriado. No que respeita à análise da independência para modelos de regressão linear, deve utilizar-se a estatística de Durbin-Watson, que não rejeita a hipótese de independência quando toma valores próximos de 2. A observação das FAC e FACP dos resíduos deve, em ambos os métodos de previsão, servir de complemento à informação recolhida com recurso à inferência estatística.

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

5.3.1 Caso I: Comércio a retalho em Portugal

Nesta Secção descreve-se, a título de exemplo, o processo de modelação da série do comércio a retalho em Portugal (o nosso país). Os principais resultados dos processos de modelação para as séries do comércio a retalho dos restantes países (Espanha, Alemanha, Itália, França, Reino Unido e Países Baixos) são apresentados na Secção 5.3.3.

Modelo SARIMA

O primeiro passo para a identificação de um modelo SARIMA passa pela estacionarização da série em estudo. Desta forma, depois da estabilização da variância, através da transformação logarítmica, é, então, preciso definir qual a ordem de diferenciação regular para a estabilização da média.

Após aplicar uma diferenciação de 1.^a ordem ($d = 1$), a série passa a ser estacionária em média (Figura 5.6). De forma a sustentar esta afirmação, feita com base na análise gráfica, são aplicados dois testes de estacionariedade – o teste ADF e o teste KPSS – à série após diferenciação. No teste ADF, se a hipótese nula é rejeitada conclui-se que a série dos resíduos é estacionária, enquanto que, no teste

KPSS, a rejeição da hipótese nula implica a não estacionariedade da série. A escolha do número de *lags* para o teste ADF, ou seja, do valor de p , tem por base a regra proposta por Ng & Perron (1995), descrita no Capítulo 3. O mesmo número de *lags* é utilizado para o teste KPSS.

De acordo com a distribuição associada a cada um dos testes, a estacionariedade é confirmada, em ambos os casos, se a estatística de teste for inferior ao valor crítico, que, para um nível de significância de 5%, é 0,4630 no teste KPSS e $-1,9500$ no teste ADF. Neste caso, para 13 *lags*, as estatísticas de teste são 0,0720 e $-2,9223$ para os testes KPSS e ADF, respetivamente. Assim, conclui-se, a um nível de significância de 5%, que, após uma diferenciação de 1.^a ordem, a série é estacionária em média.

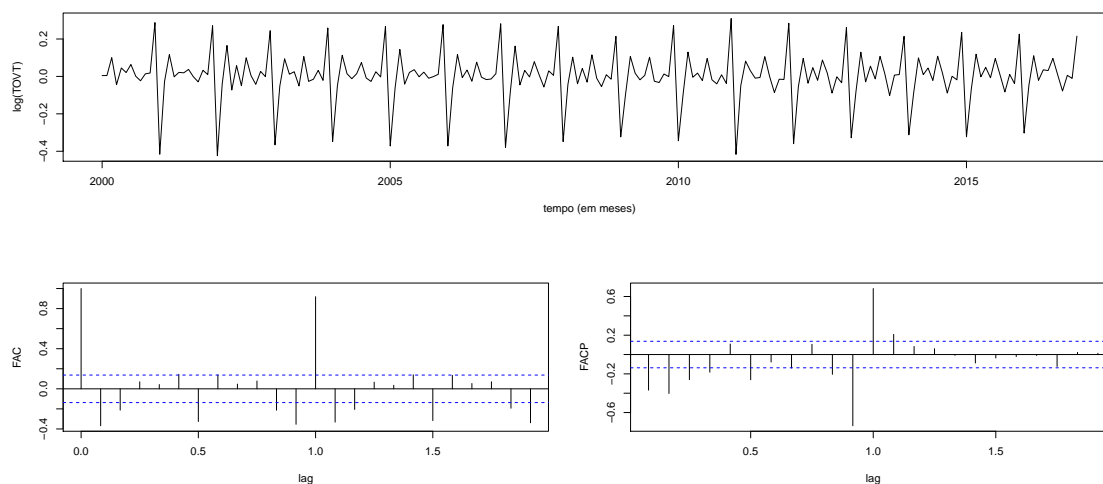


Figura 5.6: Série dos logaritmos do comércio a retalho em Portugal, após diferenciação de 1.^a ordem ($d = 1$), e respetivas FAC e FACP estimadas.

Depois de se transformar a série original numa série estacionária em média e em variância, o próximo passo é ajustar a parte sazonal do modelo, que é bem visível na Figura 5.6, tanto na representação gráfica da série como na FAC correspondente. A periodicidade aparenta ser anual, e, portanto, considera-se $s = 12$. A identificação das restantes ordens da parte sazonal (P , D e Q) é explorada através da combinação das várias possibilidades, fazendo variar P , D e Q entre 0 e 1 (Tabela 5.4).

De acordo com os resultados obtidos, o modelo que resulta no menor AIC é o que considera $P = 1$, $D = 0$ e $Q = 1$ e, portanto, estas são as ordens escolhidas para a parte sazonal do modelo. Obviamente, estas não são imutáveis e, de acordo com as necessidades futuras, podem ser alteradas (aumentadas ou reduzidas).

Tabela 5.4: Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos do comércio a retalho em Portugal.

Modelo	$\hat{\nu}_1$	$\hat{\eta}_1$	AIC
SARIMA(0, 1, 0)(1, 0, 0) ₁₂	0,9762	–	–812,45
SARIMA(0, 1, 0)(1, 0, 1) ₁₂	0,9981	–0,6707	–856,31
SARIMA(0, 1, 0)(0, 1, 0) ₁₂	–	–	–798,41
SARIMA(0, 1, 0)(0, 1, 1) ₁₂	–	–0,6734	–844,36

Na Figura 5.7 pode ver-se o comportamento dos resíduos após o ajustamento da parte sazonal.

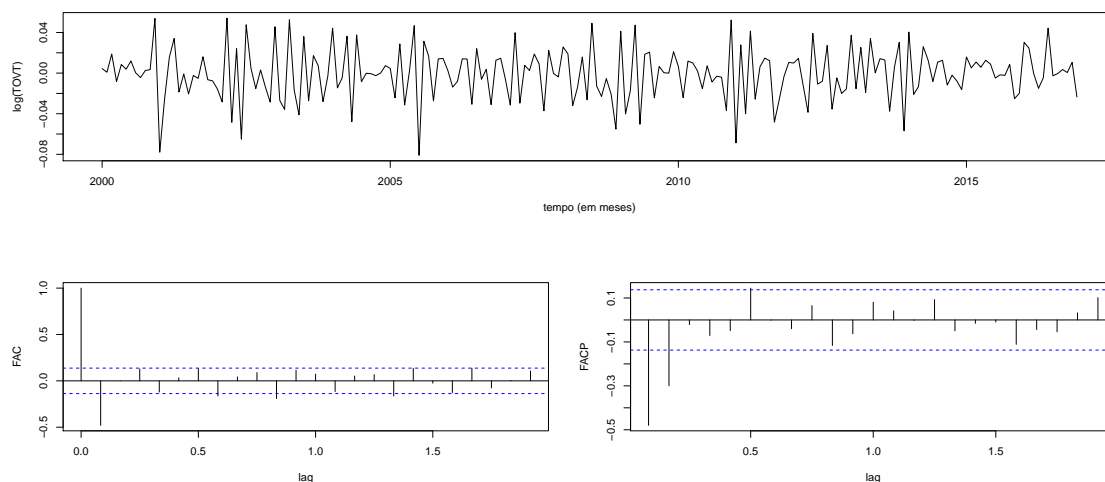


Figura 5.7: Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após diferenciação de 1.^a ordem e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas.

O último passo para a identificação de um modelo SARIMA é a escolha das ordens p e q , através da comparação das FAC e FACP empíricas com as FAC e FACP teóricas dos vários modelos conhecidos. De facto, de acordo com a Tabela 4.1 da Secção 4.1.3, as FAC e FACP da Figura 5.7 sugerem o ajustamento de um modelo AR(2), ou, alternativamente, de um MA(1), aos resíduos obtidos após a estimação da parte sazonal. No entanto, além destes modelos, são também ajustados modelos “vizinhos”, de forma a realizar uma escolha mais cuidada (ver Tabela 5.5).

Analisando a Tabela 5.5, verifica-se que, para os modelos que consideram $P = 1, D = 0$ e $Q = 1$, os AICs são bastante próximos. O modelo mais parcimonioso nessas condições, SARIMA(0, 1, 1)(1, 0, 1)₁₂, não é considerado adequado já que o intervalo de confiança para ν_1 inclui o valor 1 e, portanto, não se deve escolher $P = 1$. Em vez disso, deve considerar-se $P = 0$ e $D = 1$, o que resulta num

aumento significativo do AIC, levando à exclusão deste modelo. No que respeita ao modelo com mais parâmetros, este inclui um coeficiente não significativo e, por isso, também não deve ser considerado. Por fim, resta escolher entre três modelos com igual número de parâmetros, donde se seleciona aquele que apresenta o menor AIC. Assim, o modelo escolhido é SARIMA(2, 1, 0)(1, 0, 1)₁₂.

Tabela 5.5: Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos do comércio a retalho em Portugal.

Modelo	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\nu}_1$	$\hat{\eta}_1$	AIC
SARIMA(2, 1, 0)(1, 0, 1) ₁₂	-0,6810	-0,3283	-	-	0,9967	-0,5448	-933,72
SARIMA(2, 1, 1)(1, 0, 1) ₁₂	-0,6580	-0,3166	-0,0253*	-	0,9967	-0,5442	-931,73
SARIMA(1, 1, 1)(1, 0, 1) ₁₂	-0,2216	-	-0,4602	-	0,9965	-0,5468	-929,66
SARIMA(0, 1, 1)(1, 0, 1) ₁₂	-	-	-0,5881	-	0,9966**	-0,5681	-927,42
SARIMA(0, 1, 1)(0, 1, 1) ₁₂	-	-	-0,5806	-	-	-0,5714	-911,06
SARIMA(0, 1, 2)(1, 0, 1) ₁₂	-	-	-0,7087	0,1812	0,9965	-0,5390	-931,25

* o coeficiente não é estatisticamente significativo, para um nível de significância $\alpha = 5\%$.

** o intervalo de confiança a 95% para ν_1 contém o valor 1.

Os resultados da estimação do modelo escolhido podem ser consultados, em mais detalhe, na Tabela 5.6.

Tabela 5.6: Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos do comércio a retalho em Portugal.

Modelo final: SARIMA(2, 1, 0)(1, 0, 1) ₁₂ AIC = -933,72 $\hat{\sigma} = 0,0211$				
	ϕ_1	ϕ_2	ν_1	η_1
estimativa	-0,6810	-0,3283	0,9967	-0,5448
erro padrão	0,0676	0,0665	0,0017	0,0661

Como é possível observar, a estimativa do coeficiente ν_1 é muito próxima de 1, mas, ainda assim, as ordens da parte sazonal mantêm-se, uma vez que o valor 1 não pertence ao intervalo de confiança a 95%.

Para validar o modelo escolhido é necessário realizar uma análise dos resíduos. Idealmente, estes devem apresentar um comportamento próximo de uma distribuição Normal, de média nula e variância constante, e não apresentar correlação temporal.

O histograma da Figura 5.8 sugere que os resíduos têm uma distribuição ligeiramente assimétrica à esquerda, não sendo, contudo, suficiente para que o teste de

Kolmogorov-Smirnov rejeite a hipótese de normalidade (valor de prova de 0,2552). Além disso, de acordo com a representação gráfica da série dos resíduos (Figura 5.8), esta parece apresentar uma distribuição uniforme em torno do resíduo zero, o que leva a crer que os erros têm média nula e variância constante. De facto, o teste- t para o valor esperado confirma a não rejeição da hipótese de média nula, resultando num valor de prova de 0,6804. Assim, conclui-se que não há evidência estatística para rejeitar a hipótese de que os erros seguem uma distribuição Normal de média nula e variância constante.

Quanto à independência, o teste de Ljung-Box é aplicado à série dos resíduos, fazendo variar k entre 5 e 36 (k corresponde ao número de autocorrelações a serem testadas como um grupo). Segundo os resultados do teste, a hipótese de independência não é rejeitada para nenhum dos valores de k , apresentando valores de prova entre 0,0568 ($k = 30$) e 0,4591 ($k = 15$). Mais, as FAC e FACP estimadas dos resíduos (Figura 5.8) assemelham-se às FAC e FACP de um ruído branco e, portanto, pode admitir-se a independência dos erros.

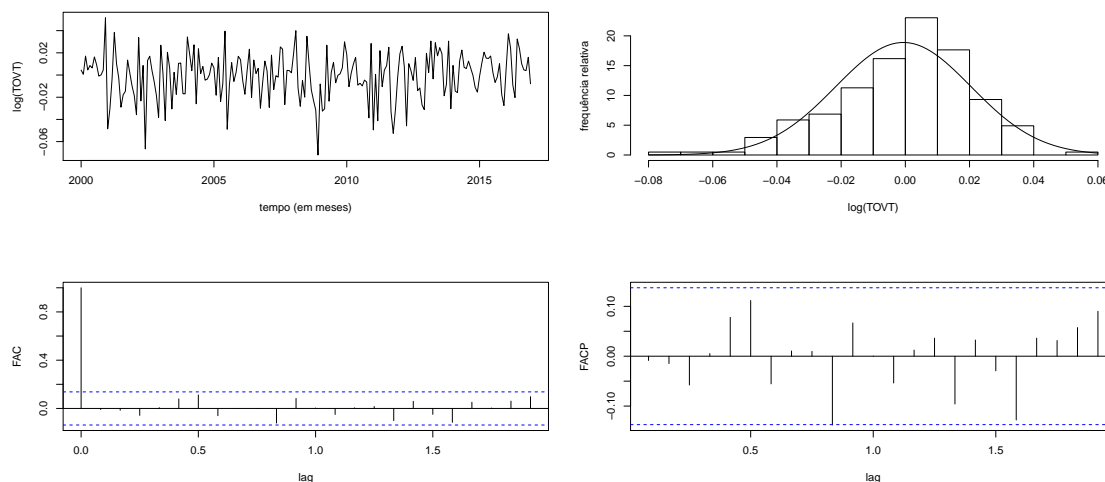


Figura 5.8: Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.

Na Figura 5.9 encontram-se representadas as previsões (no período de teste, i.e., de janeiro de 2017 a fevereiro de 2018), pontuais e intervalares, e as estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo final, nas unidades originais, sobrepostas à série em estudo.

A Figura 5.9 sugere que a qualidade preditiva do modelo é melhor na série de treino do que na série de teste, o que é comprovado através das medidas de avaliação calculadas para esses dois períodos (ver Tabela 5.34, Secção 5.4).

Relativamente aos intervalos de previsão, pode dizer-se que a sua taxa de cobertura é, neste caso, de 100%, uma vez que todas as 14 observações da série de teste pertencem ao interior dos mesmos.

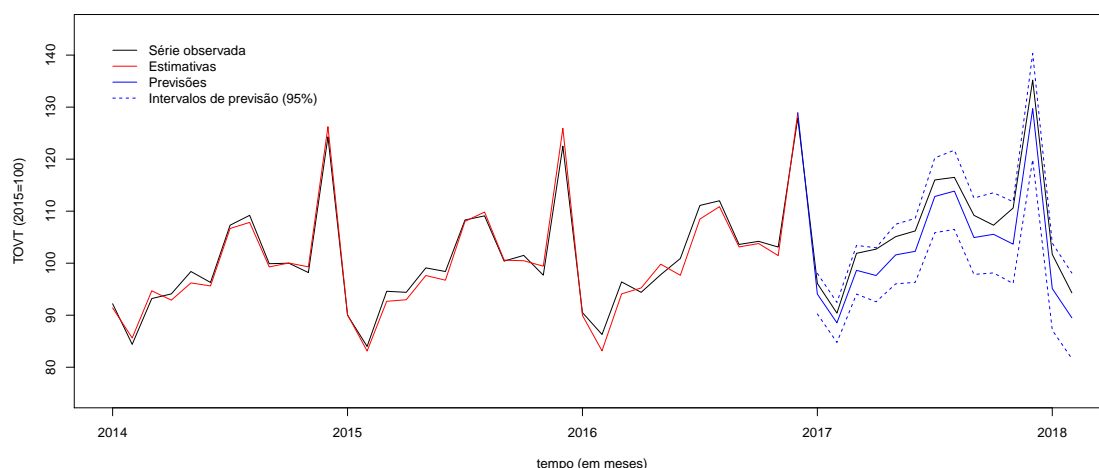


Figura 5.9: Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo SARIMA, sobrepostas à série do comércio a retalho em Portugal.

Método de Holt-Winters

A série do comércio a retalho em Portugal apresenta, à semelhança das restantes séries em estudo, uma componente sazonal bastante marcada. Como visto na Secção 4.2, o método de alisamento exponencial mais adequado para dados com sazonalidade é o método de Holt-Winters, que tanto pode ser aditivo como multiplicativo, de acordo com o tipo de sazonalidade inerente aos dados.

Após a estimação dos modelos aditivo e multiplicativo de Holt-Winters (Tabelas 5.7 e 5.8) à série do comércio a retalho em Portugal, conclui-se que o mais adequado em termos de erro quadrático médio a 1-passo é o que considera a sazonalidade na sua forma multiplicativa. De facto, tal conclusão vai ao encontro da análise gráfica (Figura 5.2, Secção 5.2), onde é possível observar que a amplitude sazonal é proporcional ao nível da série.

Pela análise das Tabelas 5.7 e 5.8, conclui-se que as estimativas para as constantes de alisamento associadas à taxa de crescimento (declive), β , apresentam valores próximos de zero, o que, tendo em consideração as equações de atualização da Secção 4.2.3, indica que o declive num dado instante fica determinado, quase na totalidade, pelo seu valor no instante anterior.

Tabela 5.7: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método multiplicativo de Holt-Winters à série do comércio a retalho em Portugal.

Modelo multiplicativo					EQM = 6,1686
$\hat{\alpha} = 0,3469$	$\hat{\beta} = 0,0610$	$\hat{\gamma} = 0,6259$	$\hat{l}_1 = 100,3229$	$\hat{b}_1 = 0,5067$	
$\hat{s}_1 = 0,8818$	$\hat{s}_2 = 0,8554$	$\hat{s}_3 = 0,9579$	$\hat{s}_4 = 0,9521$	$\hat{s}_5 = 0,9685$	
$\hat{s}_6 = 0,9834$	$\hat{s}_7 = 1,0252$	$\hat{s}_8 = 1,0220$	$\hat{s}_9 = 0,9946$	$\hat{s}_{10} = 1,0021$	
$\hat{s}_{11} = 1,0136$	$\hat{s}_{12} = 1,3434$				

Tabela 5.8: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método aditivo de Holt-Winters à série do comércio a retalho em Portugal.

Modelo aditivo					EQM = 6,7054
$\hat{\alpha} = 0,3306$	$\hat{\beta} = 0,0577$	$\hat{\gamma} = 0,6922$	$\hat{l}_1 = 100,3229$	$\hat{b}_1 = 0,5067$	
$\hat{s}_1 = -12,2962$	$\hat{s}_2 = -15,1087$	$\hat{s}_3 = -4,3795$	$\hat{s}_4 = -5,0087$	$\hat{s}_5 = -3,2920$	
$\hat{s}_6 = -1,7170$	$\hat{s}_7 = 2,5913$	$\hat{s}_8 = 2,2747$	$\hat{s}_9 = -0,5087$	$\hat{s}_{10} = 0,2622$	
$\hat{s}_{11} = 1,4538$	$\hat{s}_{12} = 35,7288$				

As estimativas/previsões, determinadas através das equações de atualização da Tabela 4.5 (Secção 4.2.3), permitem calcular as medidas de avaliação, tanto para a série de treino como para a série de teste. Como é possível observar na Tabela 5.9, de acordo com as cinco medidas de avaliação calculadas, o modelo multiplicativo é, de facto, aquele que melhor estima a série do comércio a retalho em Portugal no período de treino (janeiro de 2000 a dezembro de 2016). No entanto, para o período de teste, o modelo aditivo é o que apresenta melhores resultados.

Tabela 5.9: Medidas de avaliação dos modelos de Holt-Winters para a série do comércio a retalho em Portugal.

Modelo	EQM	REQM	EPAM	EEAM	U-Theil	
aditivo	6,7054	2,5895	1,8328	0,6046	0,1821	série de treino
multiplicativo	6,1686	2,4837	1,7152	0,5693	0,1730	
aditivo	7,3072	2,7032	1,8939	0,4391	0,2355	série de teste
multiplicativo	9,0174	3,0029	2,3260	0,5262	0,2604	

Na verdade, este tipo de resultado é bastante recorrente uma vez que “o modelo que melhor se ajusta aos dados históricos pode não ser o melhor para gerar previsões

(*out-of-sample*)” (Chatfield, 2000). Tal acontece porque, independentemente do método de previsão que é utilizado, as previsões (*out-of-sample*), geralmente, não se aproximam tanto da realidade como seria esperado pelo ajustamento do modelo à série de treino. Desta forma, a comparação da qualidade preditiva de vários métodos deve sempre ser avaliada em dados que não são incluídos na estimação do modelo (Chatfield, 2000).

Ainda assim, de forma a manter um critério de seleção para cada um dos métodos, o modelo escolhido para a modelação da série do comércio a retalho em Portugal continua a ser o multiplicativo. Além disso, tal escolha não influencia a conclusão relativamente aos três métodos de previsão (ver Tabela 5.34, Secção 5.4).

A construção dos intervalos de previsão é feita com base nas previsões pontuais e nas estimativas das variâncias dos erros de previsão calculados para os vários horizontes temporais, $h = 1, 2, \dots, 14$, sendo, neste caso, a variância para cada h estimada pelo erro quadrático médio a h -passos (Tabela 5.10). Assim, a construção do intervalo de previsão para cada horizonte temporal h segue os seguintes passos:

1. calcular as estimativas a h -passos para o modelo escolhido;
2. determinar o erro quadrático médio a h -passos, com base nos erros de previsão obtidos para esse horizonte temporal;
3. construir o intervalo de previsão a 95% usando a expressão (4.30) da Secção 4.2.5.

Para facilitar este processo foram desenvolvidas três funções no ambiente R, que podem ser consultadas no Apêndice D.

Tabela 5.10: Erro quadrático médio a h -passos para o modelo de Holt-Winters escolhido para a série do comércio a retalho em Portugal.

h	1	2	3	4	5	6	7
EQM_h	6,1686	6,9295	8,2348	10,1080	10,8606	12,3001	14,6986
h	8	9	10	11	12	13	14
EQM_h	15,8017	18,1320	20,2802	21,0151	23,6247	30,3382	33,0001

No entanto, a validade destes intervalos é dependente do cumprimento do pressuposto de normalidade dos erros de previsão para cada horizonte temporal h . Os valores de prova resultantes do teste de Kolmogorov-Smirnov, que testa a hipótese nula “os erros de previsão seguem uma distribuição Normal” contra a alternativa “os

erros de previsão não seguem uma distribuição Normal”, encontram-se na Tabela 5.11.

Analisando os resultados dos testes à normalidade, conclui-se, para um nível de significância de 5%, que apenas o intervalo para a previsão a 11-passos (que corresponde ao mês de novembro de 2017) não é válido. Apesar disso, a taxa de cobertura é de 100%, como é possível verificar na Figura 5.10.

Tabela 5.11: Teste à normalidade dos erros de previsão a h -passos para o modelo de Holt-Winters escolhido para a série do comércio a retalho em Portugal.

h	1	2	3	4	5	6	7
valor-p	0,1726	0,2259	0,0529	0,0933	0,3175	0,1483	0,1619
h	8	9	10	11	12	13	14
valor-p	0,2545	0,0937	0,1090	0,0381	0,0667	0,1277	0,3069

Como esperado, quanto maior a distância temporal da previsão maior o erro associado e, portanto, o erro quadrático médio aumenta, de forma geral, à medida que o horizonte temporal aumenta. Obviamente, isso também se reflete na amplitude dos intervalos de previsão.

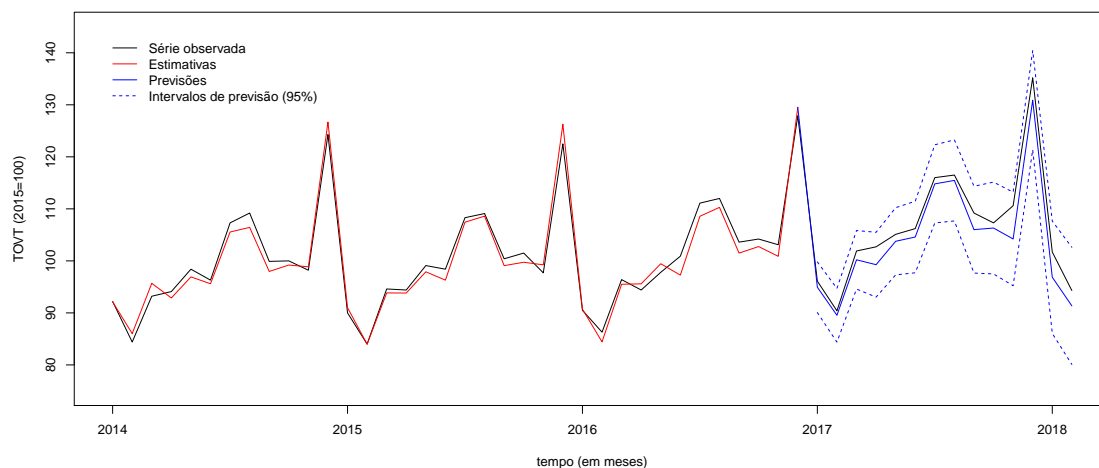


Figura 5.10: Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de Holt-Winters, sobrepostas à série do comércio a retalho em Portugal.

Modelo de regressão linear múltipla

Após a transformação da série do comércio a retalho em Portugal numa série de logaritmos, é necessário identificar qual o modelo mais adequado para descrever a tendência da mesma. De forma a facilitar esse processo é, então, utilizada uma decomposição (aditiva) da série nas suas componentes, construída através de médias móveis (comando *decompose*), e observado o comportamento da componente tendência (Figura 5.11).

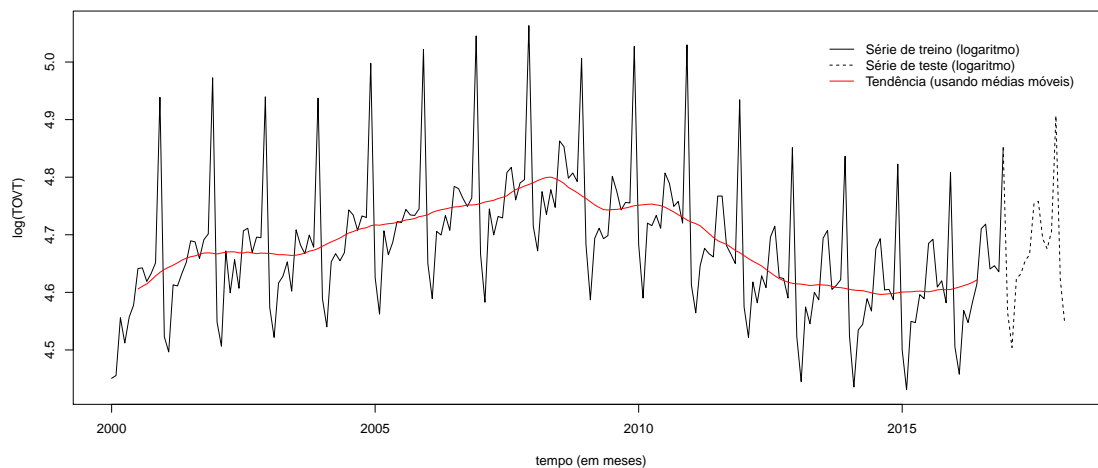


Figura 5.11: Série dos logaritmos do comércio a retalho em Portugal e respetiva componente tendência, obtida por médias móveis (usando o comando *decompose* no ambiente R).

Analisando a Figura 5.11, facilmente se conclui que a série não apresenta uma tendência linear. De facto, se se ajustar o modelo (5.1) (Secção 5.3) com uma tendência linear à série em estudo, torna-se claro, pela análise da série dos resíduos (Figura 5.12), que assumir um comportamento linear não é suficiente para a modelação da tendência. Aliás, o decaimento lento para zero apresentado pela FAC da Figura 5.12 é um indicador da não estacionariedade da série dos resíduos, que sugere, portanto, a necessidade de um outro ajustamento para a tendência.

Voltando à análise da Figura 5.11, é identificada, numa primeira instância, a ocorrência de um evento, entre janeiro de 2007 e dezembro de 2009, que resulta na divisão da tendência em duas partes: uma linear crescente e uma quadrática de concavidade voltada para cima.

Para a determinação do ponto de corte procede-se, então, à estimação de vários modelos, idênticos ao descrito na equação (5.1) (Secção 5.3), alterando apenas a

componente tendência, isto é,

$$T_t = \begin{cases} b_0 + b_1 t, & \text{se } t < t_1 \\ (b_0 + b_2) + (b_1 + b_3)t + b_4 t^2, & \text{se } t \geq t_1 \end{cases}, \quad (5.3)$$

com t_1 a variar entre 85 e 120 (codificações para janeiro de 2007 e dezembro de 2009, respetivamente). Após a estimação dos modelos, considerando diferentes datas para a ocorrência do evento, escolhe-se o ponto de corte que corresponde ao modelo com o maior coeficiente de determinação ajustado. Assim, neste caso, considera-se que ocorreu um evento a janeiro de 2008 ($R_a^2 = 0,9619$).

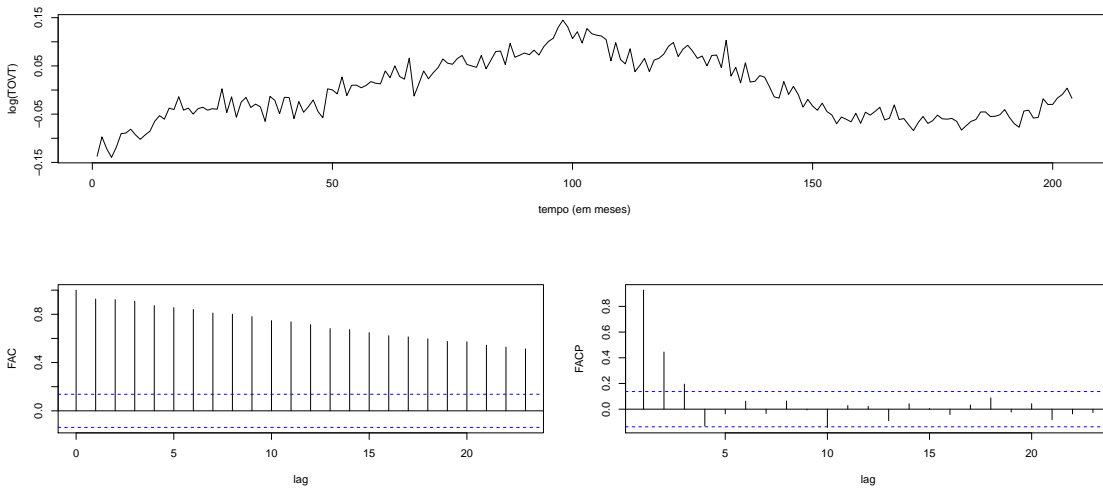


Figura 5.12: Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após ajustamento de uma tendência linear, e respetivas FAC e FACP estimadas.

No entanto, a tendência quadrática, estimada para as observações a partir de janeiro de 2008, pode, ainda, ser substituída por duas lineares, o que se traduz numa divisão em três tendências. Para tal, é, então, necessário localizar a ocorrência de um novo evento (os dois eventos são exclusivos), desta vez entre janeiro de 2013 e dezembro de 2015. Assim, considerando que o primeiro ponto de corte é em janeiro de 2008, o procedimento para determinar o segundo ponto de corte é idêntico ao ilustrado anteriormente, onde a componente tendência é dada por

$$T_t = \begin{cases} b_0 + b_1 t, & \text{se } t < t_1 \\ (b_0 + b_2) + (b_1 + b_3)t, & \text{se } t_1 \leq t < t_2 \\ (b_0 + b_2 + b_4) + (b_1 + b_3 + b_5)t, & \text{se } t \geq t_2 \end{cases}, \quad (5.4)$$

com $t_1 = 97$ (codificação para janeiro de 2008) e t_2 a variar entre 157 e 192 (codi-

ficações para janeiro de 2013 e dezembro de 2015, respetivamente). Desta forma, o modelo com o maior coeficiente de determinação ajustado é aquele que considera que o segundo evento ocorreu a abril de 2014 ($R_a^2 = 0,9694$).

Assim, pela análise da Figura 5.11, identificam-se dois modelos possíveis:

M1, com um evento em janeiro de 2008, que separa uma tendência linear de uma tendência quadrática;

M2, com dois eventos exclusivos, um em janeiro de 2008 e outro em abril de 2014, que dividem a tendência em três tendências lineares.

De notar que estes eventos vão, de certa forma, ao encontro do que se esperava, uma vez que se aproximam tanto do início como do final da crise financeira em Portugal. De facto, a crise europeia teve início em 2008, apesar de ter sido sentida com mais intensidade no ano seguinte, e, em Portugal, começou a ser ultrapassada em 2014.

Depois da estimação dos modelos M1 e M2, deve proceder-se à análise da série dos resíduos de forma a detetar algum tipo de autocorrelação, uma vez que, como descrito na Secção 4.3.5, a existência de autocorrelação pode trazer consequências ao nível da significância das variáveis e induzir em erro na sua inclusão/eliminação. Na verdade, a análise das FAC e FACP estimadas (Figura 5.13) revela, em ambos os casos, a necessidade de modelar a série dos resíduos como um modelo autorregressivo de ordem 3. No entanto, antes de se avançar para o processo de modelação é conveniente avaliar a série dos resíduos quanto à sua estacionariedade. Para tal, são aplicados dois testes – o teste ADF e o teste KPSS – à série dos resíduos, sendo a conclusão complementada com a análise gráfica da FAC correspondente.

Tendo em conta as distribuições utilizadas nestes testes, a estacionariedade é confirmada, em ambos os casos, se a estatística de teste for inferior ao valor crítico. Este valor crítico, para um nível de significância de 5%, é 0,4630 no caso do teste KPSS e $-1,9500$ para o teste ADF. Assim, considerando 12 lags, as estatísticas de teste são $-3,1826$ e $-4,2852$, para o teste ADF aplicado à série dos resíduos dos modelos M1 e M2, respetivamente, e 0,0493 e 0,0315, para o teste KPSS. Conclui-se, portanto, que as séries dos resíduos são estacionárias e que, por isso, se pode prosseguir com a modelação das mesmas.

Nas Tabelas 5.12 e 5.13 podem ver-se os resultados do ajustamento dos modelos M1 e M2, incluindo as correções feitas aos erros padrão das estimativas, às estatísticas de teste (teste- z em vez de teste- t) e aos valores de prova, tendo em consideração os modelos autorregressivos, AR(3), ajustados aos resíduos. As correções são determinadas de acordo com as equações (4.45) e (4.46), apresentadas na Secção 4.3.5.

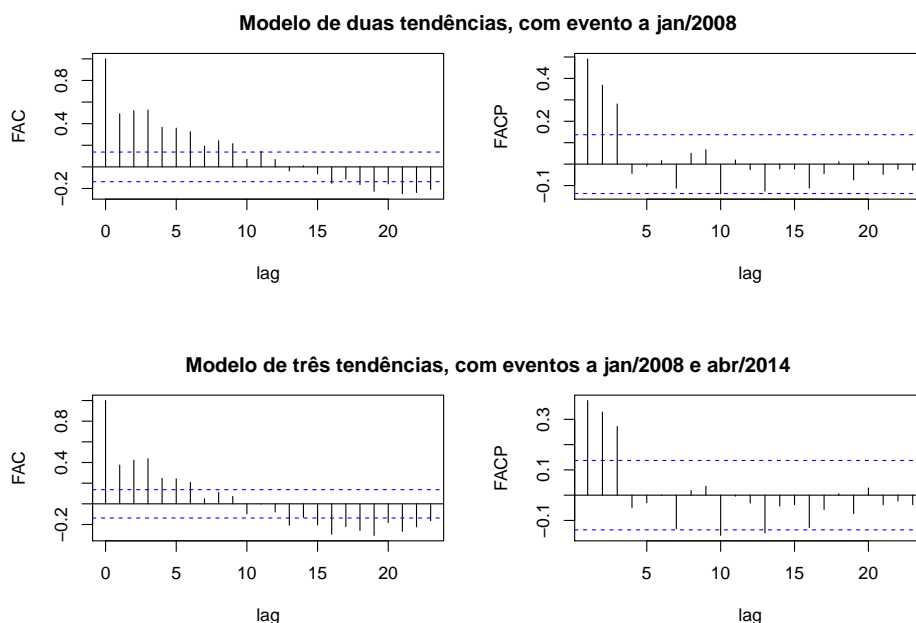


Figura 5.13: FAC e FACP obtidas para as séries dos resíduos, resultantes do ajustamento dos dois modelos de regressão (completos) à série dos logaritmos do comércio a retalho em Portugal.

Analisando as Tabelas 5.12 e 5.13, facilmente se percebe que a correção para a autocorrelação produz, efetivamente, algumas mudanças na estrutura final do modelo. Por exemplo, no modelo com um evento (Tabela 5.12), a interação da variável tempo com o cosseno de período 2,4 ($t:\cos 2,4$) passa a ser significativa assim que se tem em consideração a autocorrelação existente nos dados. O mesmo acontece no modelo com dois eventos (Tabela 5.13) relativamente às interações da variável tempo com os cossenos de períodos 2,4 e 6 ($t:\cos 2,4$ e $t:\cos 6$) e com o seno de período 4 ($t:\sin 4$). Como se sabe, tais resultados podem ter influência direta no número de variáveis dos modelos reduzidos, uma vez que, em modelos harmónicos (com erros correlacionados), quando uma curva cosseno com um certo período é significativa, esta deve ser incluída juntamente com o correspondente seno do mesmo período, e vice-versa.

Relativamente à qualidade de ajustamento, a correção para a autocorrelação traduz-se em algumas melhorias. De facto, basta comparar o coeficiente de determinação ajustado (R_a^2) com o seu correspondente corrigido (R_{ac}^2) para perceber que os modelos corrigidos explicam melhor a variabilidade dos dados (+0,0154 no modelo M1 e +0,0088 no modelo M2).

Tendo em consideração a significância das variáveis e que, em modelos harmónicos (com erros correlacionados), basta que um seno ou cosseno de um

certo período seja significativo para que o par seno/cosseno seja mantido no modelo, devem ser removidas do modelo M1 as interações da variável tempo com o cosseno de período 2 ($t:\cos 2$) e com os seno e cosseno de período 4 ($t:\cos 4$ e $t:\sin 4$).

Tabela 5.12: Resultados do ajustamento do modelo completo, com um evento, à série dos logaritmos do comércio a retalho em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,606401	0,004790	961,6934	<0,0001	0,019867	231,8570	<0,0001
tempo	0,001760	0,000086	20,5274	<0,0001	0,000304	5,7876	<0,0001
evento(*)	0,942288	0,056940	16,5487	<0,0001	0,137213	6,8673	<0,0001
cos12	0,041715	0,004624	9,0214	<0,0001	0,003521	11,8464	<0,0001
sen12	-0,053169	0,004612	-11,5283	<0,0001	0,003610	-14,7273	<0,0001
cos6	0,055925	0,004623	12,0962	<0,0001	0,002387	23,4336	<0,0001
sen6	-0,028797	0,004593	-6,2699	<0,0001	0,002369	-12,1546	<0,0001
cos4	0,072416	0,004623	15,6634	<0,0001	0,002589	27,9698	<0,0001
sen4	-0,027994	0,004589	-6,0997	<0,0001	0,002624	-10,6695	<0,0001
cos3	0,062925	0,004623	13,6106	<0,0001	0,003619	17,3852	<0,0001
sen3	0,002234	0,004588	0,4869	0,6269	0,003590	0,6223	0,5337
cos2,4	0,061611	0,004623	13,3257	<0,0001	0,003249	18,9646	<0,0001
sen2,4	0,013505	0,004588	2,9440	0,0037	0,003268	4,1333	<0,0001
cos2	0,018274	0,003257	5,6114	<0,0001	0,001964	9,3049	<0,0001
t :evento	-0,011655	0,000781	-14,9206	<0,0001	0,001951	-5,9750	<0,0001
t^2 :evento	0,000026	0,000003	10,0920	<0,0001	0,000006	4,2505	<0,0001
t :cos12	-0,000202	0,000039	-5,1947	<0,0001	0,000030	-6,8164	<0,0001
t :sen12	-0,000094	0,000039	-2,3903	0,0179	0,000030	-3,0702	0,0021
t :cos6	-0,000037	0,000039	-0,9415	0,3478	0,000020	-1,8257	0,0679
t :sen6	0,000153	0,000039	3,9132	0,0001	0,000020	7,5945	<0,0001
t :cos4	0,000012	0,000039	0,2985	0,7657	0,000022	0,5337	0,5936
t :sen4	0,000043	0,000039	1,0927	0,2760	0,000022	1,9228	0,0545
t :cos3	-0,000124	0,000039	-3,1753	0,0018	0,000030	-4,1014	<0,0001
t :sen3	0,000029	0,000039	0,7385	0,4612	0,000030	0,9538	0,3402
t :cos2,4	-0,000063	0,000039	-1,6041	0,1105	0,000027	-2,2922	0,0219
t :sen2,4	0,000042	0,000039	1,0793	0,2819	0,000027	1,5310	0,1258
t :cos2	-0,000023	0,000028	-0,8393	0,4024	0,000017	-1,3968	0,1625
R^2	0,9668	F	197,97	$g.l.$	177	$\hat{\sigma}$	0,0231
R_a^2	0,9619	$\hat{\phi}_1$	0,1997	$\hat{\phi}_2$	0,2903	$\hat{\phi}_3$	0,3054
		R_c^2	0,9806	R_{ac}^2	0,9773	$\hat{\sigma}_a$	0,0165

(*) evento ocorre em janeiro de 2008.

Tabela 5.13: Resultados do ajustamento do modelo completo, com dois eventos, à série dos logaritmos do comércio a retalho em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,606395	0,004294	1072,6568	<0,0001	0,013282	346,8066	<0,0001
tempo	0,001761	0,000077	22,9068	<0,0001	0,000210	8,3936	<0,0001
evento1(*)	0,504244	0,015643	32,2354	<0,0001	0,036212	13,9248	<0,0001
evento2(*)	-0,217896	0,072712	-2,9967	0,0031	0,123979	-1,7575	0,0788
cos12	0,041870	0,004151	10,0857	<0,0001	0,003472	12,0598	<0,0001
sen12	-0,053531	0,004135	-12,9469	<0,0001	0,003550	-15,0797	<0,0001
cos6	0,056055	0,004145	13,5231	<0,0001	0,002377	23,5867	<0,0001
sen6	-0,028893	0,004118	-7,0165	<0,0001	0,002360	-12,2422	<0,0001
cos4	0,072486	0,004145	17,4876	<0,0001	0,002575	28,1506	<0,0001
sen4	-0,028017	0,004115	-6,8083	<0,0001	0,002609	-10,7371	<0,0001
cos3	0,062966	0,004146	15,1888	<0,0001	0,003581	17,5831	<0,0001
sen3	0,002208	0,004114	0,5367	0,5922	0,003551	0,6218	0,5341
cos2,4	0,061668	0,004145	14,8766	<0,0001	0,003241	19,0271	<0,0001
sen2,4	0,013475	0,004113	3,2760	0,0013	0,003263	4,1299	<0,0001
cos2	0,018311	0,002920	6,2716	<0,0001	0,001968	9,3021	<0,0001
t:evento1	-0,004819	0,000135	-35,7658	<0,0001	0,000352	-13,6705	<0,0001
t:evento2	-0,000590	0,000393	-1,4987	0,1358	0,000685	-0,8608	0,3893
t:cos12	-0,000205	0,000035	-5,8221	<0,0001	0,000030	-6,9060	<0,0001
t:sen12	-0,000082	0,000035	-2,3209	0,0214	0,000030	-2,7189	0,0066
t:cos6	-0,000040	0,000035	-1,1470	0,2529	0,000020	-2,0003	0,0455
t:sen6	0,000157	0,000035	4,4828	<0,0001	0,000020	7,7965	<0,0001
t:cos4	0,000009	0,000035	0,2620	0,7936	0,000022	0,4217	0,6732
t:sen4	0,000044	0,000035	1,2690	0,2061	0,000022	2,0113	0,0443
t:cos3	-0,000126	0,000035	-3,5952	0,0004	0,000030	-4,2068	<0,0001
t:sen3	0,000030	0,000035	0,8601	0,3909	0,000030	1,0066	0,3141
t:cos2,4	-0,000065	0,000035	-1,8552	0,0652	0,000027	-2,3828	0,0172
t:sen2,4	0,000043	0,000035	1,2309	0,2200	0,000027	1,5669	0,1171
t:cos2	-0,000024	0,000025	-0,9901	0,3235	0,000017	-1,4734	0,1406
R^2	0,9734	F	238,80	$g.l.$	176	$\hat{\sigma}$	0,0207
R_a^2	0,9694	$\hat{\phi}_1$	0,1619	$\hat{\phi}_2$	0,2654	$\hat{\phi}_3$	0,2873
		R_c^2	0,9815	R_{ac}^2	0,9782	$\hat{\sigma}_a$	0,0161

(*) evento1 ocorre em janeiro de 2008 e evento2 em abril de 2014.

No que respeita ao modelo M2 (Tabela 5.13), existem, numa primeira instância, três variáveis que devem ser removidas: a interação da variável tempo com o coseno de período 2 ($t:\cos 2$), a variável que codifica o 2.º evento (evento2) e a sua interação com a variável tempo ($t:\text{evento2}$). No entanto, ao contrário do que acontece no modelo M1, a ordem pela qual as variáveis não significativas são removidas é importante. De facto, assim que se remove do modelo a interação $t:\text{evento2}$, a variável evento2 passa a ser estatisticamente significativa e, portanto, não deve ser retirada do mesmo. Esta situação reforça a necessidade de remover as variáveis separadamente, começando sempre pelas menos significativas (maior valor de prova), e de avaliar o efeito que a sua remoção tem nas restantes.

Com a eliminação da variável $t:\text{evento2}$ do modelo M2, é possível concluir que as primeira e terceira tendências são idênticas em termos de declive, apresentando apenas diferenças em termos de ordenada na origem. Tal significa que, a partir de abril de 2014, o índice de volume de negócios apresenta um crescimento semelhante ao observado entre janeiro de 2000 e dezembro de 2007.

Os resultados do ajustamento dos modelos finais reduzidos, com um e dois eventos, encontram-se disponíveis nas Tabelas 5.14 e 5.15, respetivamente. De notar que os dois modelos incluem, ainda, variáveis não significativas, que não são removidas de forma a garantir a validade da relação (4.44) (Secção 4.3.5).

De acordo com as Tabelas 5.14 e 5.15, o melhor modelo, em termos de coeficiente de determinação ajustado corrigido (R_{ac}^2), é aquele que divide a tendência através de dois eventos, um em janeiro de 2008 e outro em abril de 2014, sendo, portanto, o escolhido para estimar a série dos logaritmos do comércio a retalho em Portugal.

Após a escolha do modelo final, é importante verificar se este cumpre os pressupostos associados aos modelos de regressão: normalidade, independência, média nula e variância constante dos erros.

Recorrendo, numa primeira abordagem, à análise gráfica da série dos resíduos (Figura 5.14), conclui-se que o modelo final cumpre os pressupostos. De facto, nos dois primeiros gráficos é possível observar que os resíduos se distribuem uniformemente em torno de zero e se aproximam do comportamento da função densidade de uma distribuição Normal, indicando que os erros seguem uma distribuição Normal de média nula e variância constante. Além disso, a FAC dos resíduos apresenta um comportamento semelhante à FAC de um ruído branco, ou seja, com autocorrelações não significativamente diferentes de zero, o que sugere a independência dos erros.

Tabela 5.14: Resultados do ajustamento do modelo final reduzido, com um evento, à série dos logaritmos do comércio a retalho em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,606279	0,004774	964,7881	<0,0001	0,019399	237,4501	<0,0001
tempo	0,001762	0,000085	20,6149	<0,0001	0,000298	5,9057	<0,0001
evento(*)	0,941107	0,056763	16,5796	<0,0001	0,136155	6,9120	<0,0001
cos12	0,041649	0,004610	9,0336	<0,0001	0,003581	11,6303	<0,0001
sen12	-0,053151	0,004599	-11,5570	<0,0001	0,003672	-14,4752	<0,0001
cos6	0,055824	0,004609	12,1110	<0,0001	0,002430	22,9765	<0,0001
sen6	-0,028782	0,004580	-6,2849	<0,0001	0,002412	-11,9341	<0,0001
cos4	0,073591	0,002288	32,1693	<0,0001	0,001294	56,8665	<0,0001
sen4	-0,023674	0,002288	-10,3487	<0,0001	0,001300	-18,2076	<0,0001
cos3	0,062994	0,004609	13,6666	<0,0001	0,003676	17,1384	<0,0001
sen3	0,002176	0,004575	0,4758	0,6348	0,003646	0,5970	0,5505
cos2,4	0,061644	0,004610	13,3719	<0,0001	0,003282	18,7810	<0,0001
sen2,4	0,013414	0,004573	2,9330	0,0038	0,003302	4,0627	<0,0001
cos2	0,015931	0,001617	9,8500	<0,0001	0,000969	16,4409	<0,0001
t :evento	-0,011638	0,000779	-14,9457	<0,0001	0,001935	-6,0154	<0,0001
t^2 :evento	0,000026	0,000003	10,0988	<0,0001	0,000006	4,2672	<0,0001
t :cos12	-0,000202	0,000039	-5,2081	<0,0001	0,000030	-6,7019	<0,0001
t :sen12	-0,000094	0,000039	-2,4013	0,0174	0,000031	-3,0247	0,0025
t :cos6	-0,000037	0,000039	-0,9429	0,3470	0,000020	-1,7914	0,0732
t :sen6	0,000153	0,000039	3,9223	0,0001	0,000020	7,4583	<0,0001
t :cos3	-0,000124	0,000039	-3,1832	0,0017	0,000031	-4,0385	0,0001
t :sen3	0,000029	0,000039	0,7399	0,4603	0,000031	0,9386	0,3479
t :cos2,4	-0,000062	0,000039	-1,6075	0,1097	0,000028	-2,2668	0,0234
t :sen2,4	0,000042	0,000039	1,0820	0,2807	0,000028	1,5155	0,1296
R^2	0,9664	F	224,97	$g.l.$	180	$\hat{\sigma}$	0,0231
R_a^2	0,9621	$\hat{\phi}_1$	0,1998	$\hat{\phi}_2$	0,2847	$\hat{\phi}_3$	0,3040
		R_c^2	0,9800	R_{ac}^2	0,9770	$\hat{\sigma}_a$	0,0167

(*) evento ocorre em janeiro de 2008.

Tabela 5.15: Resultados do ajustamento do modelo final reduzido, com dois eventos, à série dos logaritmos do comércio a retalho em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,607472	0,004242	1086,0284	<0,0001	0,012937	356,1515	<0,0001
tempo	0,001739	0,000076	23,0001	<0,0001	0,000204	8,5454	<0,0001
evento1(*)	0,502848	0,015672	32,0849	<0,0001	0,036470	13,7878	<0,0001
evento2(*)	-0,325701	0,011359	-28,6741	<0,0001	0,028562	-11,4031	<0,0001
cos12	0,042203	0,004159	10,1471	<0,0001	0,003478	12,1345	<0,0001
sen12	-0,053926	0,004141	-13,0232	<0,0001	0,003572	-15,0984	<0,0001
cos6	0,056111	0,004159	13,4909	<0,0001	0,002390	23,4814	<0,0001
sen6	-0,029016	0,004132	-7,0230	<0,0001	0,002373	-12,2271	<0,0001
cos4	0,072531	0,004159	17,4391	<0,0001	0,002585	28,0625	<0,0001
sen4	-0,028205	0,004128	-6,8326	<0,0001	0,002618	-10,7746	<0,0001
cos3	0,063109	0,004158	15,1764	<0,0001	0,003598	17,5389	<0,0001
sen3	0,002070	0,004127	0,5014	0,6167	0,003568	0,5800	0,5619
cos2,4	0,061775	0,004159	14,8542	<0,0001	0,003285	18,8044	<0,0001
sen2,4	0,013402	0,004127	3,2478	0,0014	0,003310	4,0487	0,0001
cos2	0,015813	0,001460	10,8345	<0,0001	0,000976	16,1958	<0,0001
t :evento1	-0,004794	0,000134	-35,7197	<0,0001	0,000355	-13,5107	<0,0001
t :cos12	-0,000210	0,000035	-5,9938	<0,0001	0,000029	-7,1494	<0,0001
t :sen12	-0,000077	0,000035	-2,1870	0,0300	0,000030	-2,5452	0,0109
t :cos6	-0,000041	0,000035	-1,1619	0,2468	0,000020	-2,0221	0,0432
t :sen6	0,000158	0,000035	4,4889	<0,0001	0,000020	7,7950	<0,0001
t :cos4	0,000009	0,000035	0,2472	0,8050	0,000022	0,3979	0,6907
t :sen4	0,000047	0,000035	1,3336	0,1840	0,000022	2,1139	0,0345
t :cos3	-0,000128	0,000035	-3,6559	0,0003	0,000030	-4,2705	<0,0001
t :sen3	0,000031	0,000035	0,8975	0,3707	0,000030	1,0489	0,2942
t :cos2,4	-0,000067	0,000035	-1,8991	0,0592	0,000028	-2,4136	0,0158
t :sen2,4	0,000042	0,000035	1,2100	0,2279	0,000028	1,5249	0,1273
R^2	0,9729	F	255,98	$g.l.$	178	$\hat{\sigma}$	0,0208
R_a^2	0,9691	$\hat{\phi}_1$	0,1579	$\hat{\phi}_2$	0,2704	$\hat{\phi}_3$	0,2863
		R_c^2	0,9812	R_{ac}^2	0,9781	$\hat{\sigma}_a$	0,0162

(*) evento1 ocorre em janeiro de 2008 e evento2 em abril de 2014.

Tais conclusões vão ao encontro dos resultados inferenciais: o teste de Kolmogorov-Smirnov garante a hipótese de normalidade, com um valor de prova de 0,9171; a estatística de Durbin-Watson aproxima-se de 2 ($DW = 1,9731$), o que permite concluir que não se rejeita a independência dos erros; e o teste- t ao valor esperado apresenta um valor de prova de 0,8908, que resulta na não rejeição da hipótese de média nula.

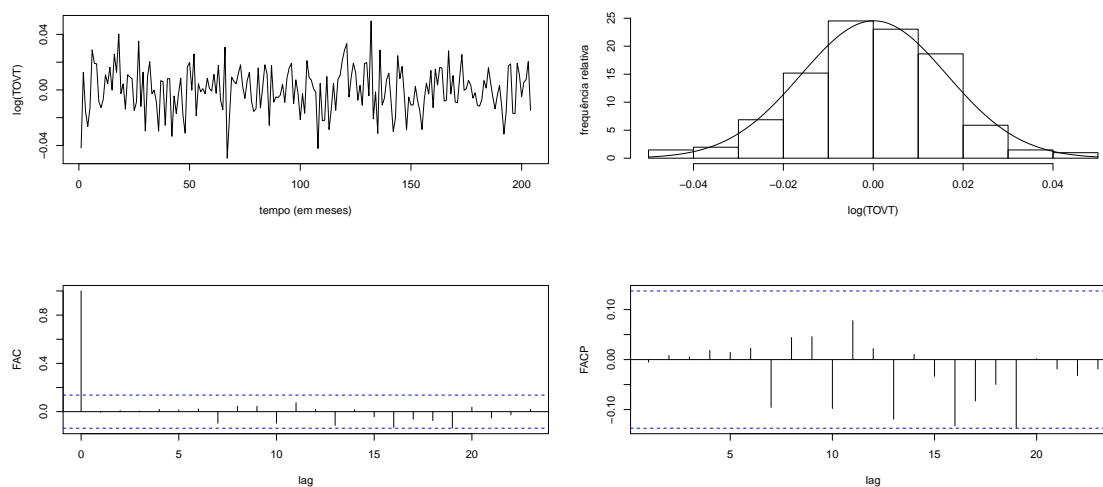


Figura 5.14: Série dos resíduos para a série dos logaritmos do comércio a retalho em Portugal, após ajustamento do modelo de regressão, e respetivo histograma, FAC e FACP estimadas.

Obviamente que, uma vez ajustado o modelo e estimados os respetivos parâmetros, é necessário reverter a transformação logarítmica de modo a obter previsões, pontuais e intervalares, nas unidades originais. Além disso, também os erros de previsão e as medidas de avaliação devem ser determinados nas unidades originais.

Comparando as medidas de avaliação calculadas para os dois modelos finais (Tabela 5.16), o modelo escolhido é também o que produz os melhores resultados, tanto na série de treino como na de teste.

Tabela 5.16: Medidas de avaliação dos modelos de regressão para a série do comércio a retalho em Portugal.

Modelo	EQM	REQM	EPAM	EEAM	U-Theil	
1 evento	3,5482	1,8837	1,3427	0,4240	0,1308	série de treino
2 eventos	3,3801	1,8385	1,2992	0,4109	0,1278	
1 evento	36,0581	6,0048	5,2687	1,1835	0,5308	série de teste
2 eventos	13,9676	3,7373	3,0549	0,6845	0,3363	

De facto, as diferenças na qualidade preditiva são mais visíveis na série de teste, onde, por exemplo, o EPAM difere em mais de 2%, rondando os 5,27% no modelo com um evento e apenas os 3,05% no modelo com dois eventos. Na verdade, estes resultados são reforçados por Cowpertwait & Metcalfe (2009), que referem que, em modelos de regressão, as previsões se devem basear num modelo linear para a tendência, estimado apenas para os dados mais recentes, em vez de considerar modelos polinomiais de ordens superiores.

Na Figura 5.15 encontram-se representadas as previsões (período de teste: janeiro de 2017 a fevereiro de 2018), pontuais e intervalares, e as estimativas pontuais (período: janeiro de 2014 a dezembro de 2016) obtidas através do modelo final, juntamente com a série original.

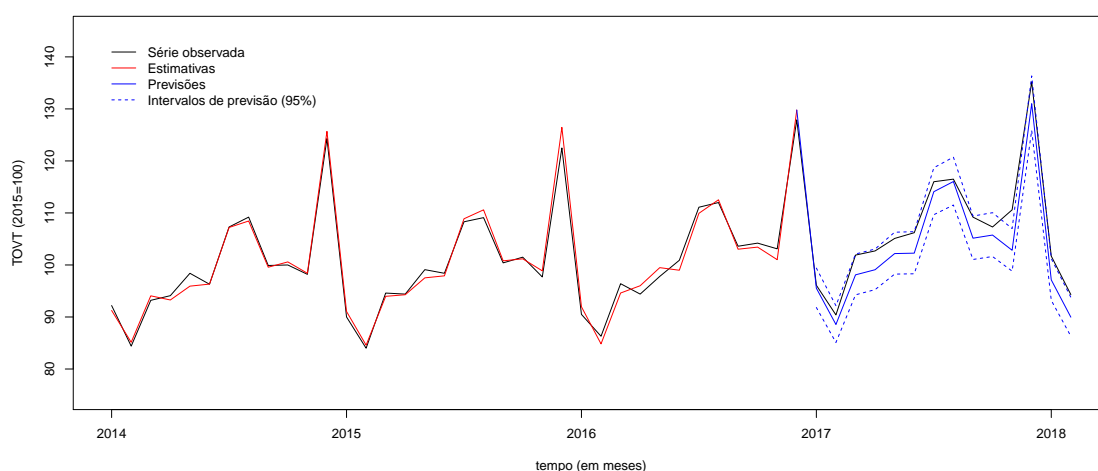


Figura 5.15: Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de regressão linear, sobrepostas à série do comércio a retalho em Portugal.

Relativamente aos intervalos de previsão, apenas 11 das 14 observações da série de teste pertencem ao interior dos mesmos, o que corresponde a uma taxa de cobertura bastante inferior à esperada (78,6% em vez de 95%).

Considerações finais

Da análise dos resultados obtidos no processo de modelação e previsão da série do comércio a retalho em Portugal pode concluir-se que:

- a amplitude dos intervalos de previsão para o modelo de regressão linear múltipla é menor do que a dos intervalos de previsão para os restantes métodos;

- em termos da taxa de cobertura, os intervalos de previsão para os modelos SARIMA e de Holt-Winters são exímios (cobertura de 100%). Em contrapartida, apenas 11 das 14 observações da série de teste pertencem ao interior dos intervalos de previsão para o modelo de regressão, o que representa uma cobertura de apenas 78,6%;
- de acordo com as cinco medidas de avaliação calculadas, o modelo mais preciso na estimação da série de treino é o de regressão linear múltipla. No entanto, o mesmo não se pode dizer relativamente à sua qualidade preditiva, sendo o modelo de Holt-Winters o mais adequado para estimar a série de teste. De facto, os modelos de regressão são mais apropriados para previsões a curto prazo, como é o caso da série de treino em que as estimações são obtidas a 1-passo, do que para previsões a longo prazo, como acontece para a série de teste.

5.3.2 Caso II: *E-commerce* em Portugal

Nesta Secção descreve-se, a título de exemplo, o processo de modelação da série do *e-commerce* em Portugal (o nosso país). Durante o processo, esta série apresenta algumas características distintas da série do comércio a retalho em Portugal, sendo, por isso, também relevante apresentar a análise detalhada do mesmo. Os principais resultados dos processos de modelação para as séries do *e-commerce* dos restantes países (Espanha, Alemanha, Itália, França, Reino Unido e Países Baixos) são apresentados na Secção 5.3.3.

Modelo SARIMA

Através da análise exploratória da série do *e-commerce* em Portugal, conclui-se que esta não é estacionária nem em média nem em variância. No entanto, para a correta identificação de um modelo SARIMA, é crucial proceder-se, num primeiro instante, à estacionarização da série. Desta forma, depois da estabilização da variância, através da transformação logarítmica, é, então, necessário definir qual a ordem de diferenciação regular que é suficiente para estabilizar a média.

Depois de ser aplicada uma diferenciação de 1.^a ordem ($d = 1$), a série dos logaritmos passa a ser estacionária em média (Figura 5.16). De facto, esta conclusão é suportada, não só pela análise gráfica, mas também pelos dois testes de estacionariedade utilizados – o teste ADF e o teste KPSS. Neste caso, para 13 lags, as estatísticas de teste são $-3,3934$ e $0,0507$ para os testes ADF e KPSS, respetivamente, o que

leva a concluir, a um nível de significância de 5%, que, após uma diferenciação de 1.^a ordem, a série é estacionária em média.

Assim que a série original é transformada numa série estacionária, o passo seguinte é ajustar a parte sazonal do modelo.

Numa primeira instância, pela análise da FAC da Figura 5.16, considera-se um período $s = 12$. No entanto, após todo o processo de identificação e estimação de um modelo com esse requisito, percebe-se, através da FAC dos resíduos do modelo final, que existe sazonalidade não explicada, sinalizada pela forte correlação que se observa no *lag* 6. Desta forma, é necessário proceder-se à escolha de um outro modelo, desta vez considerando $s = 6$.

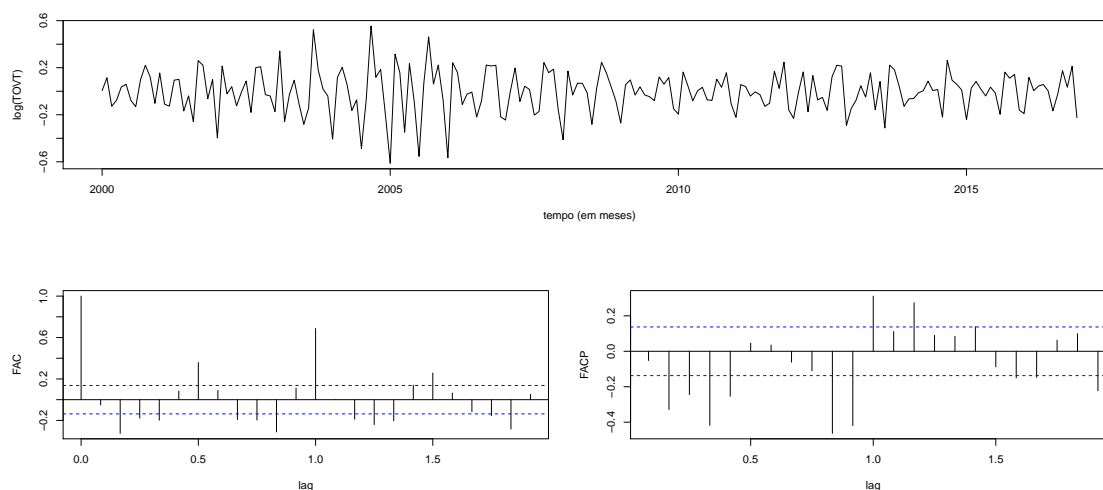


Figura 5.16: Série dos logaritmos do *e-commerce* em Portugal, após diferenciação de 1.^a ordem ($d = 1$), e respetivas FAC e FACP estimadas.

A identificação das restantes ordens da parte sazonal (P , D e Q) é explorada através da combinação das várias possibilidades, fazendo variar P , D e Q entre 0 e 2 (Tabela 5.17).

Segundo os resultados da Tabela 5.17, o modelo com o menor AIC é o que considera $P = 2$, $D = 0$ e $Q = 2$ e, portanto, estas são as ordens escolhidas para a parte sazonal. Naturalmente, estas não são definitivas e, de acordo com as necessidades futuras, podem ser modificadas (aumentadas ou reduzidas).

Na Figura 5.17 pode ver-se o comportamento dos resíduos após o ajustamento da parte sazonal.

O último passo para a identificação de um modelo SARIMA é a escolha das ordens p e q , utilizando como ferramenta as FAC e FACP empíricas. De facto, através da comparação das FAC e FACP da Figura 5.17 com as FAC e FACP

teóricas (ver Tabela 4.1, Secção 4.1.3), é possível reconhecer-se o comportamento de um modelo MA(2), ou, alternativamente, de um AR(2). No entanto, além destes modelos, são também ajustados modelos “vizinhos”, de forma a realizar uma escolha mais cuidada (ver Tabela 5.18).

Tabela 5.17: Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos do *e-commerce* em Portugal.

Modelo	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\eta}_1$	$\hat{\eta}_2$	AIC
SARIMA(0, 1, 0)(1, 0, 0) ₆	0,3595	–	–	–	–142,06
SARIMA(0, 1, 0)(1, 0, 1) ₆	0,9473*	–	–0,7395	–	–196,69
SARIMA(0, 1, 0)(0, 1, 0) ₆	–	–	–	–	–62,37
SARIMA(0, 1, 0)(0, 1, 1) ₆	–	–	–0,7893	–	–184,76
SARIMA(0, 1, 0)(1, 1, 1) ₆	–0,9387	–	0,5340	–	–234,11
SARIMA(0, 1, 0)(2, 0, 1) ₆	–0,2353	0,7209	0,7299	–	–264,67
SARIMA(0, 1, 0)(2, 0, 2) ₆	–0,0911	0,8939	0,4192	–0,4125	–277,25
SARIMA(0, 1, 0)(1, 1, 2) ₆	–0,9928	–	0,3509	–0,4849	–261,82

* o intervalo de confiança para ν_1 contém o valor 1.

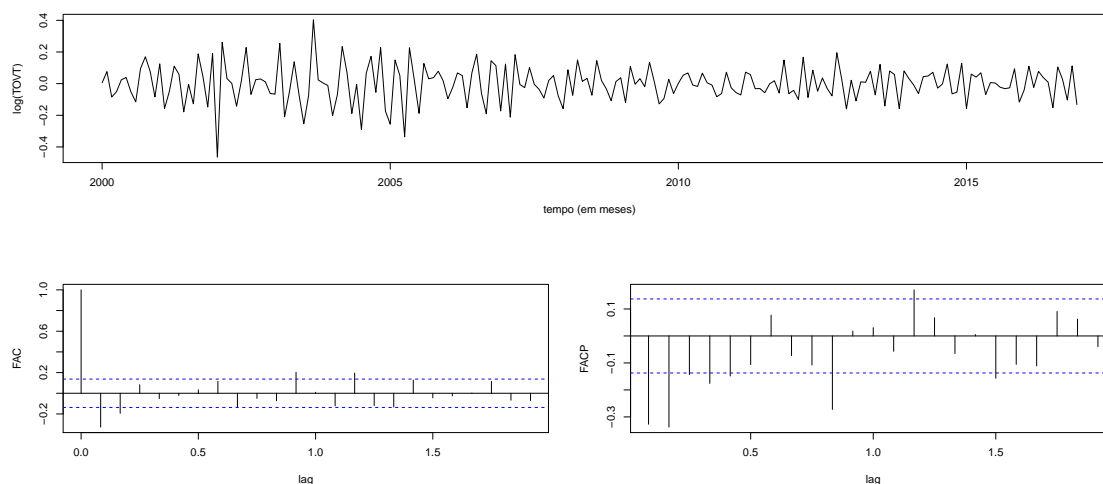


Figura 5.17: Série dos resíduos para a série dos logaritmos do *e-commerce* em Portugal, após diferenciação de 1.^a ordem e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas.

Analisando a Tabela 5.18, verifica-se que o modelo cuja parte regular é apenas autorregressiva não é tão adequado, em termos de AIC, como os restantes. A escolha entre os outros modelos foi realizada, não com base no AIC, por serem muito próximos, mas sim no número de parâmetros. De facto, de entre os três

modelos, opta-se, então, pelo mais parcimonioso, que não apresenta uma parte regular mista, ou seja, o modelo SARIMA(0, 1, 2)(2, 0, 2)₆.

Tabela 5.18: Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos do *e-commerce* em Portugal.

Modelo	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}_1$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\eta}_1$	$\hat{\eta}_2$	AIC
SARIMA(2, 1, 0)(2, 0, 2) ₆	-0,4840	-0,3622	-	-0,0594	0,9125	0,3939	-0,3344	-325,59
SARIMA(1, 1, 1)(2, 0, 2) ₆	0,2224	-	-0,8355	-0,0870	0,8893	0,4268	-0,3316	-339,16
Modelo	$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\eta}_1$	$\hat{\eta}_2$	AIC
SARIMA(0, 1, 2)(2, 0, 2) ₆	-	-0,5826	-0,1992	-0,0881	0,8879	0,4172	-0,3389	-340,27
SARIMA(1, 1, 2)(2, 0, 2) ₆	-0,8057	0,1732*	-0,7121	-0,0829	0,8892	0,4465	-0,2805	-341,61

* o coeficiente não é estatisticamente significativo, para um nível de significância $\alpha = 5\%$.

Os resultados da estimação deste modelo podem ser consultados, em mais detalhe, na Tabela 5.19.

Tabela 5.19: Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos do *e-commerce* em Portugal.

Modelo final(*): SARIMA(0, 1, 2)(2, 0, 2) ₆							
						AIC = -340,27	$\hat{\sigma} = 0,0976$
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\eta}_1$	$\hat{\eta}_2$	
estimativa	-0,5826	-0,1992	-0,0881	0,8879	0,4172	-0,3389	
erro padrão	0,0758	0,0799	0,0383	0,0391	0,0948	0,0966	

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

De forma a validar o modelo escolhido deve realizar-se uma análise dos resíduos (ver Figura 5.18). Idealmente, estes devem apresentar um comportamento próximo de uma distribuição Normal, de média nula e variância constante, e não ser correlacionados no tempo.

De acordo com a representação gráfica da série dos resíduos (Figura 5.18), considera-se que esta apresenta uma distribuição (relativamente) uniforme em torno do resíduo zero, o que sugere que os erros têm média nula e variância constante. Além disso, o histograma da Figura 5.18 sugere que, apesar da existência de algumas observações discrepantes, os resíduos têm uma distribuição Normal, o que é comprovado pela aplicação do teste de Kolmogorov-Smirnov, com um valor de prova de 0,7183. Admite-se, assim, que os erros seguem uma distribuição Normal de média nula e variância constante.

Quanto à independência, o teste de Ljung-Box é aplicado à série dos resíduos, fazendo variar k entre 7 e 36 (lembre-se que k corresponde ao número de autocorrelações a serem testadas como um grupo). A hipótese de independência dos erros é rejeitada para todos os valores de k , à exceção do $k = 7$. Apesar disso, o comportamento das FAC e FACP dos resíduos (Figura 5.18) assemelha-se ao de um ruído branco, com correlações significativas, da ordem dos 0,2, apenas para o lag 11.

Apesar da conclusão sobre o pressuposto da independência não ser a mais favorável, em comparação com outros modelos, este é aquele que, ainda assim, apresenta melhores resultados.

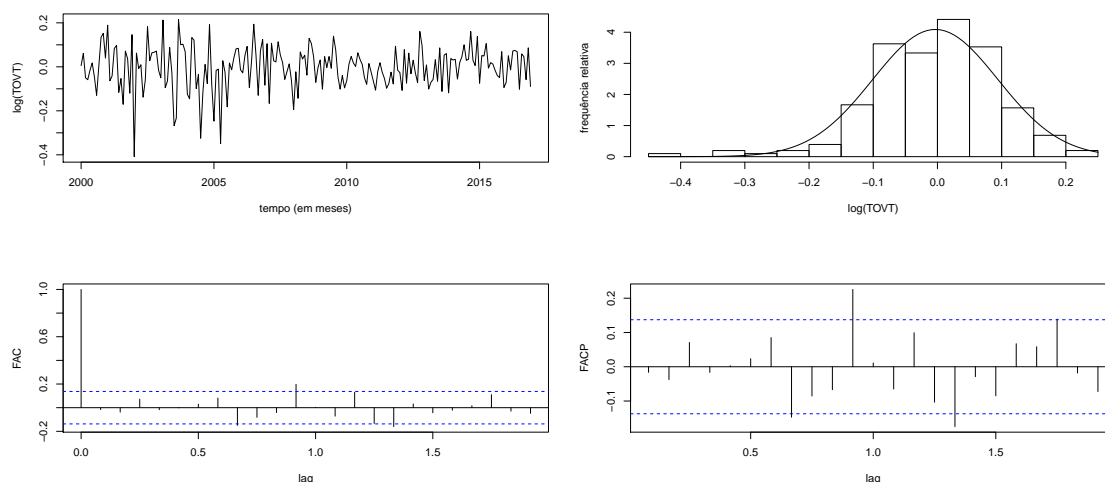


Figura 5.18: Série dos resíduos para a série dos logaritmos do *e-commerce* em Portugal, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.

Na Figura 5.19 encontram-se representadas as previsões (no período de teste: janeiro de 2017 a fevereiro de 2018), pontuais e intervalares, e as estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo final, nas unidades originais, sobrepostas à série em estudo.

Ainda que a Figura 5.19 sugira que a qualidade preditiva do modelo é melhor na série de treino do que na série de teste, tal não é comprovado pelas medidas de avaliação calculadas para esses dois períodos (ver Tabela 5.35, Secção 5.4). De facto, embora isso se verifique para a maior parte do período de treino, este modelo não é muito preciso a estimar os primeiros 5-6 anos da série do *e-commerce* em Portugal. Possivelmente, estes resultados estão diretamente relacionados com a elevada variabilidade apresentada pela série, ainda que transformada, durante esse período. Assim, com base na análise, conclui-se que os modelos SARIMA não são os mais adequados para lidar com este tipo de séries.

Relativamente aos intervalos de previsão, apenas 12 das 14 observações da série de teste pertencem ao interior dos mesmos, o que corresponde a uma taxa de cobertura um pouco inferior à esperada (85,7% em vez de 95%).

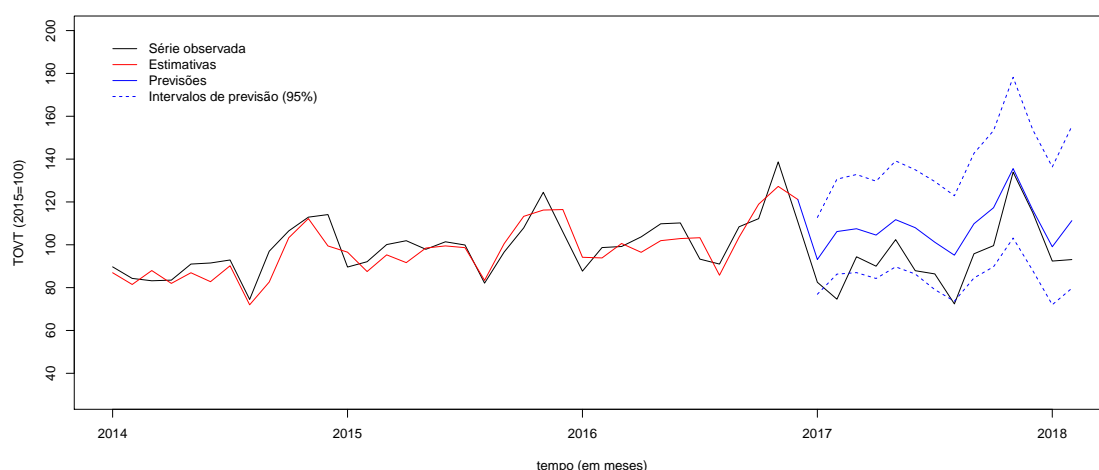


Figura 5.19: Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo SARIMA, sobrepostas à série do *e-commerce* em Portugal.

Método de Holt-Winters

Uma vez que também é possível reconhecer uma componente sazonal na série do *e-commerce* em Portugal, a escolha do melhor método de alisamento exponencial volta a recair entre os dois métodos de Holt-Winters disponíveis: o aditivo e o multiplicativo. As estimativas das constantes de alisamento e os valores usados na inicialização, tanto do método multiplicativo como do aditivo, podem ser consultados nas Tabelas 5.20 e 5.21, respetivamente.

Tabela 5.20: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método multiplicativo de Holt-Winters à série do *e-commerce* em Portugal.

Modelo multiplicativo					EQM = 420,3237
$\hat{\alpha} = 0,1823$	$\hat{\beta} = 0,0589$	$\hat{\gamma} = 0,6781$	$\hat{l}_1 = 203,2920$	$\hat{b}_1 = 2,2286$	
$\hat{s}_1 = 1,1915$	$\hat{s}_2 = 1,0632$	$\hat{s}_3 = 0,9320$	$\hat{s}_4 = 1,0118$	$\hat{s}_5 = 1,1139$	
$\hat{s}_6 = 0,9367$	$\hat{s}_7 = 0,8972$	$\hat{s}_8 = 0,7744$	$\hat{s}_9 = 0,8494$	$\hat{s}_{10} = 1,0460$	
$\hat{s}_{11} = 1,1568$	$\hat{s}_{12} = 1,0271$				

Tabela 5.21: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método aditivo de Holt-Winters à série do *e-commerce* em Portugal.

Modelo aditivo					EQM = 452,3238
$\hat{\alpha} = 0,1765$	$\hat{\beta} = 0,0648$	$\hat{\gamma} = 0,6588$	$\hat{l}_1 = 203,2920$	$\hat{b}_1 = 2,2286$	
$\hat{s}_1 = 42,2531$	$\hat{s}_2 = 13,6990$	$\hat{s}_3 = -15,7135$	$\hat{s}_4 = 2,2865$	$\hat{s}_5 = 25,6698$	
$\hat{s}_6 = -14,9552$	$\hat{s}_7 = -21,7302$	$\hat{s}_8 = -47,7510$	$\hat{s}_9 = -32,2427$	$\hat{s}_{10} = 9,2656$	
$\hat{s}_{11} = 33,6698$	$\hat{s}_{12} = 5,5490$				

À semelhança do que acontece com a série do comércio a retalho em Portugal, o modelo que apresenta o menor erro quadrático médio a 1-passo é o que considera a sazonalidade multiplicativa e, portanto, esse é o escolhido. Além disso, analisando a Tabela 5.22, conclui-se, com base nas cinco medidas de avaliação, que o modelo multiplicativo é o que produz menores erros, tanto na estimação de observações passadas como na previsão de valores futuros.

Tabela 5.22: Medidas de avaliação dos modelos de Holt-Winters para a série do *e-commerce* em Portugal.

Modelo	EQM	REQM	EPAM	EEAM	U-Theil	
aditivo	452,3238	21,2679	9,1284	0,7888	0,6311	série de treino
multiplicativo	420,3237	20,5018	8,4697	0,7405	0,5866	
aditivo	298,9753	17,2909	17,9859	1,1051	1,0825	série de teste
multiplicativo	265,2426	16,2863	16,7628	1,0299	1,0215	

Os erros quadráticos médios, calculados para os 14 horizontes temporais, necessários para a construção dos intervalos de previsão estão resumidos na Tabela 5.23.

Tabela 5.23: Erro quadrático médio a h -passos para o modelo de Holt-Winters escolhido para a série do *e-commerce* em Portugal.

h	1	2	3	4	5	6	7
EQM_h	420,3237	458,9360	470,3099	500,2939	515,3850	540,0650	616,2973
h	8	9	10	11	12	13	14
EQM_h	681,5444	716,6324	749,2076	776,7510	819,6080	1111,9670	1141,6730

De notar que a ordem de grandeza destes erros é bastante superior à dos erros quadráticos médios utilizados para o cálculo dos intervalos de previsão na série do comércio a retalho em Portugal. De facto, isso resulta em intervalos de previsão

com grandes amplitudes, daí a taxa de cobertura de 100% (ver Figura 5.20). No entanto, a normalidade dos erros de previsão é, neste caso, rejeitada, a um nível de significância de 5%, para todos os passos h , com $h = 1, 2, \dots, 14$. Assim sendo, conclui-se que os intervalos de previsão construídos segundo a expressão (4.30) não são os mais adequados para este tipo de dados.

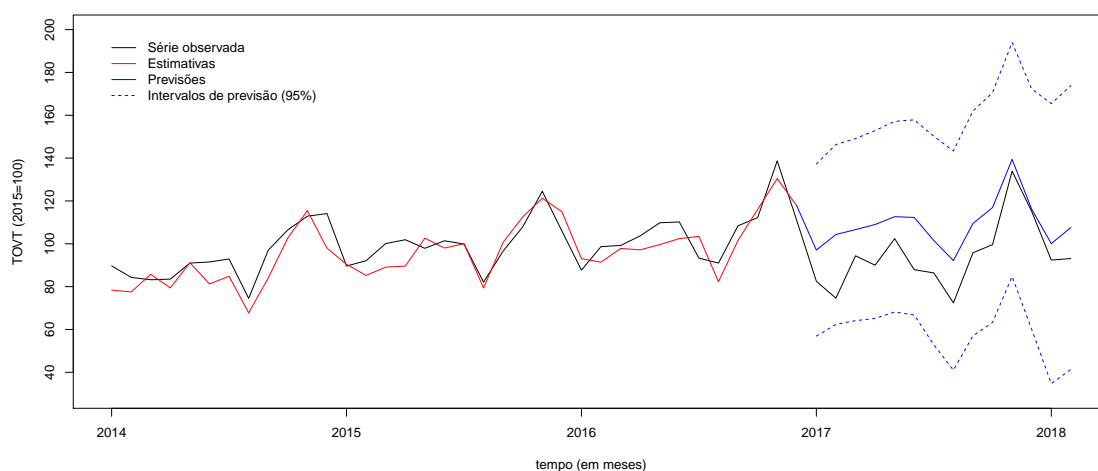


Figura 5.20: Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de Holt-Winters, sobrepostas à série do *e-commerce* em Portugal.

Modelo de regressão linear múltipla

À semelhança do que acontece no Caso I (Secção 5.3.1), depois de se transformar a série original numa série de logaritmos, é utilizada uma decomposição (aditiva) como ferramenta para indicar qual (ou quais) o(s) modelo(s) mais adequado(s) para descrever a tendência da série (Figura 5.21).

Analisando a Figura 5.21, considera-se que existem três modelos que podem servir para estimar a componente tendência. A sugestão mais simples corresponde ao modelo linear, que apresenta um coeficiente de determinação ajustado de 0,8498. Os outros modelos registam a ocorrência de um ou dois eventos, identificados nos períodos de janeiro de 2003 a dezembro de 2004 e de janeiro de 2013 a dezembro de 2014: um deles divide a tendência em duas partes, uma linear e outra quadrática, através da ocorrência de um evento a julho de 2004 ($R_a^2 = 0,8977$), enquanto que o outro separa a tendência em três tendências lineares, com eventos a julho de 2004 e a fevereiro de 2013 ($R_a^2 = 0,9232$).

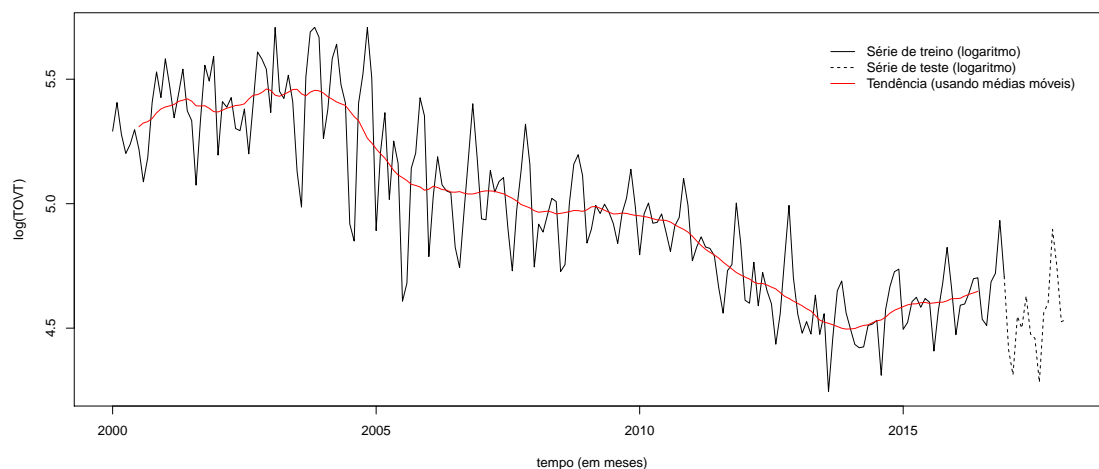


Figura 5.21: Série dos logaritmos das vendas por correspondência ou via *internet* em Portugal e respetiva componente tendência, obtida por médias móveis (usando o comando *decompose* no ambiente R).

Depois de se estimarem os três modelos, deve, então, verificar-se a existência, ou não, de correlação na série dos resíduos, nunca esquecendo de antes comprovar a estacionariedade da mesma. Os resultados da aplicação dos testes de estacionariedade, ADF e KPSS, encontram-se na Tabela 5.24.

Tabela 5.24: Resultados da aplicação dos testes de estacionariedade às séries dos resíduos dos três modelos de regressão ajustados.

<i>lags</i>	Teste ADF		<i>lags</i>	Teste KPSS		
	ET	VC (5%)		ET	VC (5%)	
14	-2,2814	-1,9500	14	0,0880	0,4630	Modelo linear
14	-1,9607	-1,9500	14	0,0776	0,4630	Modelo com 1 evento
13	-2,8551	-1,9500	13	0,0565	0,4630	Modelo com 2 eventos

De facto, e uma vez que as estatísticas de teste (ET) são inferiores aos valores críticos (VC), a estacionariedade das séries dos resíduos fica garantida. No entanto, é preciso notar que o modelo com um evento se encontra, de acordo com o teste ADF, na fronteira entre a estacionariedade e a não-estacionariedade (estatística de teste muito próxima do valor crítico). Além disso, analisando as séries dos resíduos, tanto do modelo com tendência linear como do modelo com duas tendências, estas não aparentam ser estacionárias. Ainda assim, assumindo como verdadeiras as conclusões dos testes de estacionariedade, prossegue-se com o ajustamento da correlação em todos os modelos sugeridos.

As FACP das séries dos resíduos (Figura 5.22) apresentam, nos três casos, correlações significativas para os *lags* 1, 3, 5 e 6, apesar da correlação do *lag* 3 ser muito pouco significativa para o modelo que divide a tendência em três tendências lineares.

Ajusta-se, apesar disso, um modelo autorregressivo de ordem 3, AR(3), a cada uma das séries dos resíduos. As FAC e FACP das séries dos resíduos obtidas após esse ajustamento revelam correlações significativas para o *lag* 6, sugerindo que a componente sazonal não é totalmente explicada através das ondas sinusoidais. Possivelmente, para modelar essa sazonalidade, seria necessário complementar o modelo com outros ciclos (de período 6), ou recorrer a outras metodologias. Como esse não é o objetivo desta dissertação, prossegue-se, então, a modelação, ignorando a existência de correlação no *lag* 6.

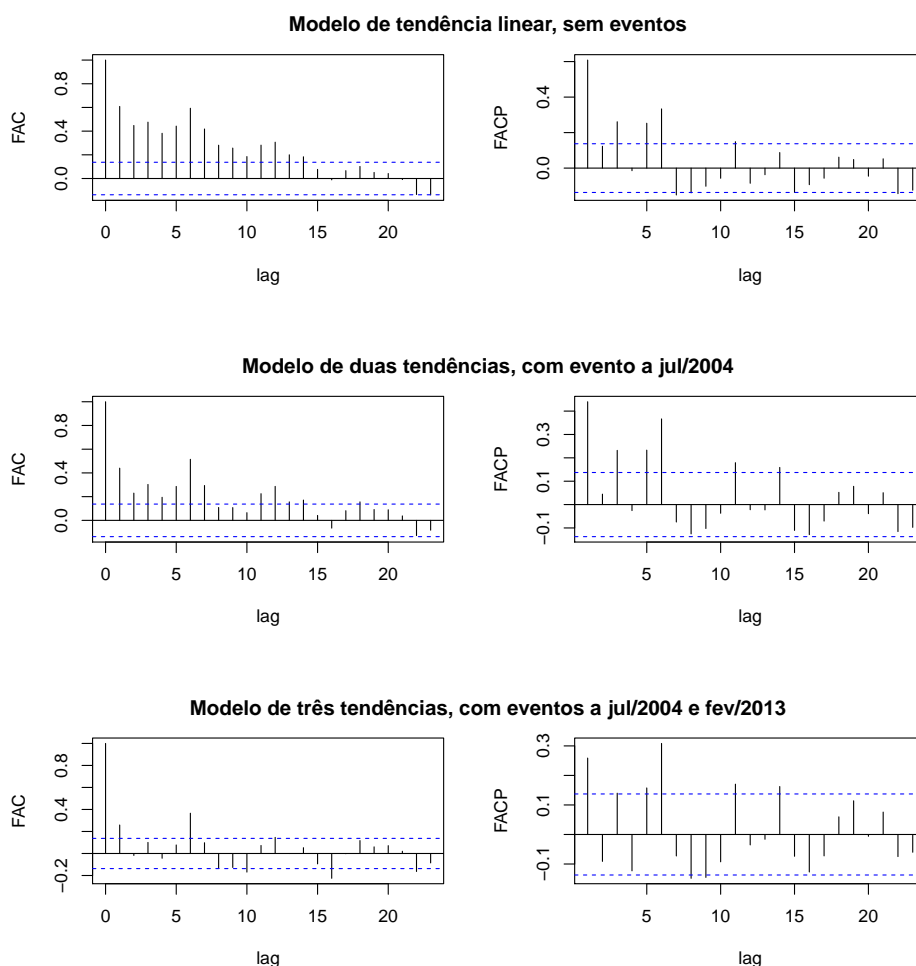


Figura 5.22: FAC e FACP obtidas para as séries dos resíduos, resultantes do ajustamento dos três modelos de regressão (completos) à série dos logaritmos do e-commerce em Portugal.

Após a correção para a autocorrelação, os coeficientes de determinação ajustados são também alterados, tomando os valores 0,9133, 0,9210 e 0,9292 para os modelos com uma, duas e três tendências, respetivamente. Naturalmente, nem todas as variáveis são estatisticamente significativas e, portanto, os modelos devem ser reduzidos.

No modelo de tendência linear, as variáveis a remover são o cosseno de período 2 ($\cos 2$) e todas as interações da variável tempo com senos e cossenos, com exceção dos períodos 2 e 12.

Para o modelo com um evento, a essas variáveis juntam-se, numa primeira análise, a variável tempo, que corresponde ao declive da 1.^a tendência, e o termo quadrático da 2.^a tendência (t^2 :evento). Durante o processo de eliminação das variáveis não significativas (por ordem decrescente de valor de prova), a remoção do termo quadrático traduz-se numa transformação do modelo que descreve a tendência, passando este a ser composto por duas tendências lineares. Além disso, a eliminação desta variável provoca alterações na significância da variável evento – a constante associada à nova tendência linear do modelo –, que também é removida. Mais, de acordo com o teste ADF, quando se retira o termo quadrático, a série dos resíduos deixa de ser estacionária. No entanto, como a estatística de teste é aproximadamente igual a $-1,95$ (valor crítico do teste ADF) e a FAC apresenta um comportamento idêntico ao do modelo com uma tendência linear, prossegue-se com a estimação.

No que respeita ao modelo que divide a tendência em três lineares, as variáveis não significativas são, inicialmente, o cosseno de período 2, a interação da variável tempo com o 2.^o evento (t :evento2) e todas as interações entre a variável tempo e os senos e cossenos também removidas dos outros modelos. Através da remoção da variável t :evento2, é possível concluir que, à semelhança do que acontece na série do comércio a retalho em Portugal, as primeira e terceira tendências são idênticas em termos de declive, o que significa que, a partir de fevereiro de 2013, o índice de volume de negócios apresenta um crescimento semelhante ao observado entre janeiro de 2000 e junho de 2004. Ao longo do processo, o modelo utilizado para modelar a correlação dos resíduos, AR(3), deixa de ser adequado, o que, tendo em consideração a respetiva FACP da Figura 5.22, já era expectável. Desta forma, a estimação da correlação dos resíduos passa a ser feita por um modelo autorregressivo de ordem 1, AR(1). A interação entre a variável tempo e o cosseno de período 2 deixa de ser significativa, sendo, por isso, eliminada.

Os resultados do ajustamento dos modelos finais reduzidos podem ser consultados nas Tabelas 5.25, 5.26 e 5.27. Estes modelos incluem, ainda, variáveis não significativas, que não são removidas a fim de garantir a validade da relação (4.44) (Secção 4.3.5).

De acordo com as Tabelas 5.25, 5.26 e 5.27, o melhor modelo, em termos de coeficiente de determinação ajustado, corrigido para a autocorrelação (R_{ac}^2), é aquele que divide a tendência através de dois eventos, um em julho de 2004 e outro em fevereiro de 2013, sendo, por isso, o escolhido para estimar a série dos logaritmos do *e-commerce* em Portugal.

Tabela 5.25: Resultados do ajustamento do modelo final reduzido, com tendência linear, à série dos logaritmos do *e-commerce* em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	5,508968	0,018924	291,1044	<0,0001	0,059830	92,0775	<0,0001
tempo	-0,005102	0,000160	-31,8639	<0,0001	0,000484	-10,5377	<0,0001
cos12	0,141599	0,026855	5,2727	<0,0001	0,025885	5,4703	<0,0001
sen12	0,042124	0,026666	1,5797	0,1158	0,026510	1,5890	0,1121
cos6	0,039069	0,013319	2,9335	0,0038	0,008959	4,3608	<0,0001
sen6	-0,137951	0,013326	-10,3523	<0,0001	0,008958	-15,3989	<0,0001
cos4	-0,038987	0,013316	-2,9279	0,0038	0,009751	-3,9984	0,0001
sen4	-0,041302	0,013316	-3,1018	0,0022	0,009797	-4,2156	<0,0001
cos3	0,038234	0,013315	2,8715	0,0046	0,009373	4,0794	<0,0001
sen3	-0,023049	0,013314	-1,7312	0,0850	0,009370	-2,4598	0,0139
cos2,4	0,021808	0,013315	1,6378	0,1031	0,006518	3,3457	0,0008
sen2,4	-0,025861	0,013315	-1,9422	0,0536	0,006508	-3,9736	0,0001
$t:\cos12$	-0,000488	0,000226	-2,1556	0,0324	0,000218	-2,2423	0,0249
$t:\sen12$	-0,000419	0,000226	-1,8484	0,0661	0,000224	-1,8688	0,0616
$t:\cos2$	-0,000074	0,000080	-0,9310	0,3530	0,000033	-2,2209	0,0264
R^2	0,8639	F	85,66	$g.l.$	189	$\hat{\sigma}$	0,1344
R_a^2	0,8538	$\hat{\phi}_1$	0,4867	$\hat{\phi}_2$	-0,0098	$\hat{\phi}_3$	0,2735
		R_c^2	0,9200	R_{ac}^2	0,9127	$\hat{\sigma}_a$	0,0992

Nota: o modelo revela sazonalidade não explicada.

Comparando as medidas de avaliação calculadas, nas unidades originais, para os três modelos finais (Tabela 5.28), o modelo escolhido é, de facto, o que apresenta melhores previsões para a série de treino. No entanto, para a série de teste, este modelo é o que pior se comporta. Possivelmente, esta discrepância na qualidade

preditiva, relativamente aos outros dois modelos, é justificada pela tendência apresentada pela série de treino (ver Figura 5.21). De facto, a partir de 2017 nota-se uma tendência ligeiramente decrescente, que, naturalmente, não pode ser bem explicada pela tendência crescente utilizada, no modelo escolhido, para explicar o comportamento a partir de fevereiro de 2013. Contudo, em situações reais, as previsões são feitas com base nos dados existentes e, este tipo de alterações na tendência não pode ser adivinhado. Além disso, independentemente do modelo de regressão escolhido, este é, de acordo com as medidas de avaliação calculadas, e apesar de existir sazonalidade por explicar, o melhor método para estimar e prever esta série (ver Tabela 5.35, Secção 5.4). Assim sendo, mantém-se o critério de seleção e opta-se, efetivamente, pelo modelo com menor coeficiente de determinação ajustado corrigido.

Tabela 5.26: Resultados do ajustamento do modelo final reduzido, com um evento, à série dos logaritmos do *e-commerce* em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	5,415596	0,014037	385,8088	<0,0001	0,033171	163,2630	<0,0001
cos12	0,143807	0,023268	6,1804	<0,0001	0,024635	5,8375	<0,0001
sen12	0,037063	0,023107	1,6040	0,1104	0,025240	1,4685	0,1420
cos6	0,037262	0,011539	3,2292	0,0015	0,008806	4,2315	<0,0001
sen6	-0,134885	0,011544	-11,6844	<0,0001	0,008806	-15,3173	<0,0001
cos4	-0,038340	0,011537	-3,3232	0,0011	0,009563	-4,0091	0,0001
sen4	-0,041915	0,011537	-3,6331	0,0004	0,009611	-4,3614	<0,0001
cos3	0,036454	0,011536	3,1600	0,0018	0,009172	3,9744	0,0001
sen3	-0,022020	0,011535	-1,9090	0,0578	0,009171	-2,4010	0,0163
cos2,4	0,022442	0,011537	1,9453	0,0532	0,006448	3,4804	0,0005
sen2,4	-0,026013	0,011537	-2,2548	0,0253	0,006439	-4,0398	0,0001
t :evento	-0,004511	0,000120	-37,6202	<0,0001	0,000274	-16,4674	<0,0001
t :cos12	-0,000500	0,000196	-2,5497	0,0116	0,000207	-2,4187	0,0156
t :sen12	-0,000391	0,000196	-1,9933	0,0477	0,000213	-1,8365	0,0663
t :cos2	-0,000078	0,000069	-1,1230	0,2629	0,000033	-2,3392	0,0193
R^2	0,8978	F	118,59	$g.l.$	189	$\hat{\sigma}$	0,1165
R_a^2	0,8902	$\hat{\phi}_1$	0,4275	$\hat{\phi}_2$	-0,0300	$\hat{\phi}_3$	0,2466
		R_c^2	0,9270	R_{ac}^2	0,9203	$\hat{\sigma}_a$	0,0948

Nota: o modelo revela sazonalidade não explicada.

(*) evento ocorre em julho de 2004.

Tabela 5.27: Resultados do ajustamento do modelo final reduzido, com dois eventos, à série dos logaritmos do *e-commerce* em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	5,307259	0,022421	236,7095	<0,0001	0,028061	189,1352	<0,0001
tempo	0,003505	0,000659	5,3202	<0,0001	0,000808	4,3374	<0,0001
evento1(*)	0,180195	0,041852	4,3056	<0,0001	0,051202	3,5193	0,0004
evento2(*)	-1,368550	0,102934	-13,2954	<0,0001	0,125539	-10,9014	<0,0001
cos12	0,146772	0,019316	7,5983	<0,0001	0,022691	6,4682	<0,0001
sen12	0,035706	0,019244	1,8554	0,0651	0,022406	1,5936	0,1110
cos6	0,033539	0,009588	3,4982	0,0006	0,009868	3,3987	0,0007
sen6	-0,132439	0,009590	-13,8095	<0,0001	0,009889	-13,3926	<0,0001
cos4	-0,040568	0,009576	-4,2366	<0,0001	0,008673	-4,6773	<0,0001
sen4	-0,041932	0,009577	-4,3786	<0,0001	0,008716	-4,8108	<0,0001
cos3	0,034884	0,009577	3,6426	0,0003	0,007856	4,4402	<0,0001
sen3	-0,022509	0,009573	-2,3513	0,0198	0,007863	-2,8626	0,0042
cos2,4	0,022013	0,009576	2,2989	0,0226	0,007396	2,9762	0,0029
sen2,4	-0,026765	0,009572	-2,7962	0,0057	0,007369	-3,6323	0,0003
t :evento1	-0,008520	0,000732	-11,6452	<0,0001	0,000901	-9,4558	<0,0001
t :cos12	-0,000565	0,000163	-3,4673	0,0007	0,000190	-2,9733	0,0029
t :sen12	-0,000308	0,000163	-1,8851	0,0610	0,000190	-1,6208	0,1051
R^2	0,9304	F	156,14	$g.l.$	187	$\hat{\sigma}$	0,0967
R_a^2	0,9244	$\hat{\phi}_1$	0,2356	$\hat{\phi}_2$	NA	$\hat{\phi}_3$	NA
		R_c^2	0,9343	R_{ac}^2	0,9282	$\hat{\sigma}_a$	0,0899

(*) evento1 ocorre em julho de 2004 e evento2 em fevereiro de 2013.

Nota: o modelo revela sazonalidade não explicada.

Tabela 5.28: Medidas de avaliação dos modelos de regressão para a série do *e-commerce* em Portugal.

Modelo	EQM	REQM	EPAM	EEAM	U-Theil	
linear	325,8555	18,0515	7,5118	0,6505	0,5409	
1 evento	293,4152	17,1294	7,2552	0,6214	0,5292	série de treino
2 eventos	258,4669	16,0769	6,8676	0,5938	0,4926	
linear	175,6450	13,2531	9,6883	0,6793	0,7203	
1 evento	152,0478	12,3308	8,8766	0,6237	0,6691	série de teste
2 evento	222,5327	14,9175	14,2797	0,8822	0,9513	

Após a escolha do modelo final, é importante verificar se este cumpre os pressupostos associados aos modelos de regressão linear: normalidade, independência, média nula e variância constante dos erros.

Na verdade, de acordo com a análise da Figura 5.23, os pressupostos de normalidade e de média nula aparentam ser válidos, sendo a não rejeição da hipótese de normalidade corroborada pelo resultado do teste de Kolmogorov-Smirnov (valor de prova de 0,5791). No entanto, relativamente à variância dos erros, a análise gráfica sugere que esta não é constante, apresentando valores mais elevados para os primeiros 5-6 anos de observações, possivelmente devido à existência de volatilidade económica nesse período.

Devido à existência de correlação significativa no *lag* 6, o pressuposto da independência não é verificado. Apesar disso, a estatística de Durbin-Watson ainda se aproxima do valor 2 (1,9589), sugerindo o contrário.

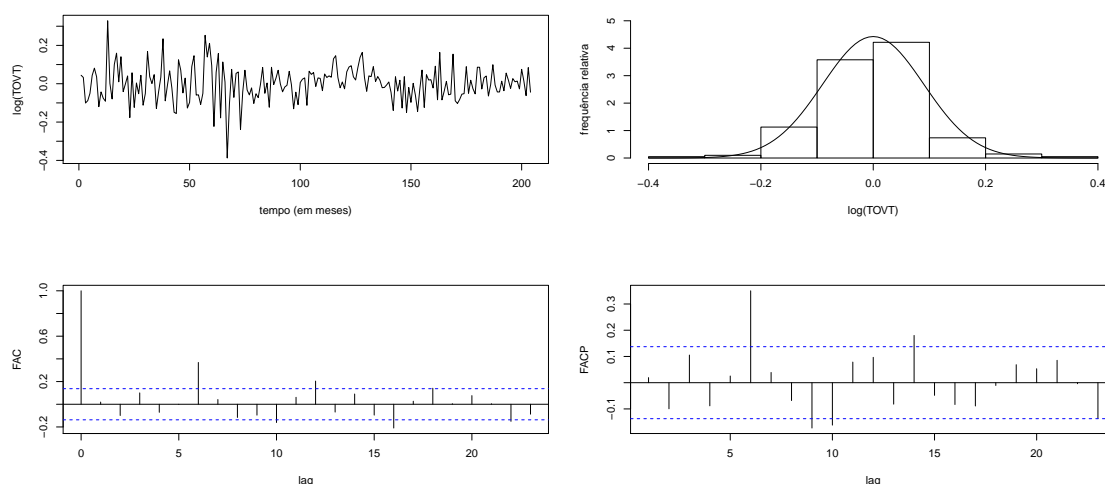


Figura 5.23: Série dos resíduos para a série dos logaritmos do *e-commerce* em Portugal, após ajustamento do modelo de regressão, e respetivo histograma, FAC e FACP estimadas.

Na Figura 5.24 encontram-se representadas as previsões (período de teste: janeiro de 2017 a fevereiro de 2018), pontuais e intervalares, e as estimativas pontuais (período: janeiro de 2014 a dezembro de 2016) obtidas através do modelo final, nas unidades originais, juntamente com a série em estudo.

Quanto às previsões pontuais, é possível observar que estas são sobrestimadas, o que se relaciona com o modelo adotado para descrever a tendência. De facto, como referido anteriormente, a tendência crescente usada para estimar o último período da série de treino não se prolonga para a série de teste.

Relativamente aos intervalos de previsão, apenas 10 das 14 observações da série de teste pertencem ao interior dos mesmos, o que corresponde a uma taxa de cobertura bastante inferior à esperada (71,4% em vez de 95%).

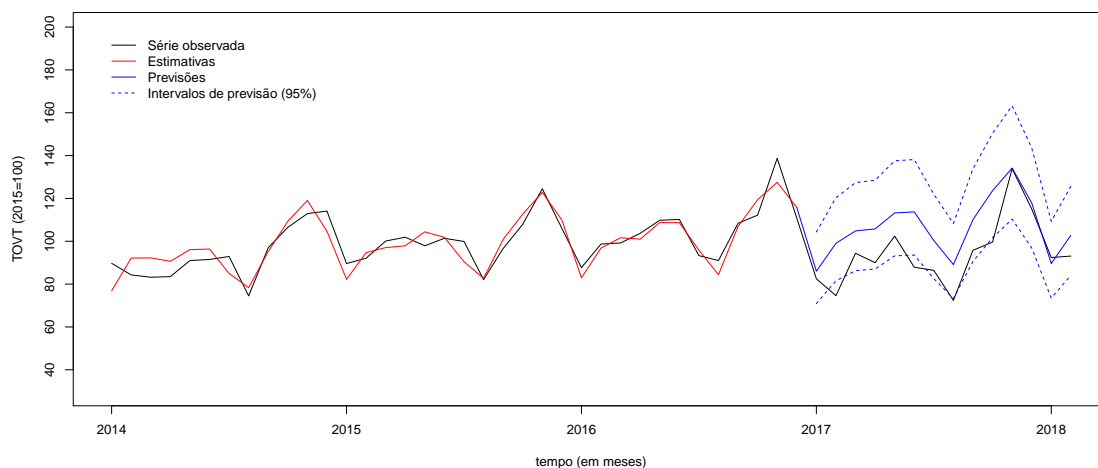


Figura 5.24: Previsões (no período de teste), pontuais e intervalares (95%), e estimativas pontuais (no período de janeiro de 2014 a dezembro de 2016) obtidas através do modelo de regressão linear, sobrepostas à série do *e-commerce* em Portugal.

Considerações finais

Da análise dos resultados obtidos no processo de modelação e previsão da série do *e-commerce* em Portugal pode concluir-se que:

- apesar de ser aplicada uma transformação logarítmica aos dados, a presença de volatilidade económica nos primeiros cinco anos da série temporal não permite a estabilização total da variância (apenas com a logaritmização dos dados);
- uma vez que as metodologias utilizadas não são as mais adequadas para lidar com séries com estas características, a volatilidade existente acaba por influenciar os resultados obtidos, dando origem a previsões pouco precisas para o domínio temporal em que esta se observa. Na verdade, todas as medidas de avaliação que consideram erros de previsão *in-sample*, nomeadamente as medidas de avaliação para a série de treino e os erros quadráticos médios a h -passos, são enviesadas pelos elevados erros registados no período em que existe volatilidade económica. Assim, neste caso, talvez a modelação/previsão apresentasse melhores resultados se fossem suprimidas as observações anteriores a 2005;
- em termos da taxa de cobertura, os intervalos de previsão para o modelo de

Holt-Winters são os únicos que incluem as 14 observações da série de teste. No entanto, a ordem de grandeza dos erros quadráticos médios a h -passos é bastante superior à dos erros calculados, nas mesmas circunstâncias, para a série do comércio a retalho em Portugal, dando origem a intervalos de previsão com grandes amplitudes, daí a cobertura de 100% (devido à volatilidade). Os outros intervalos de previsão incluem 10 e 12 observações da série de teste, para os modelos de regressão linear e SARIMA, respetivamente, o correspondente a taxas de cobertura de 71,4% e 85,7%;

- os erros de previsão a h -passos utilizados para a construção dos intervalos de previsão (modelo de Holt-Winters) não cumprem o pressuposto de normalidade e, portanto, os intervalos construídos segundo a expressão (4.30) (Secção 4.2.5) não são os mais adequados (estatisticamente corretos) para este tipo de dados. De forma a contornar esta falha, uma alternativa passa por estudar métodos de reamostragem para a construção dos intervalos de previsão (abordagem já considerada em Cordeiro (2011));
- de acordo com as cinco medidas de avaliação calculadas, o modelo mais preciso é, tanto para a série de treino como para a de teste, o modelo de regressão linear múltipla (apesar de não cumprir o pressuposto de independência, com correlação significativa no *lag* 6);
- no modelo de regressão linear, a componente sazonal não é totalmente explicada através das ondas sinusoidais. Contudo, as eventuais periodicidades escondidas nos dados podem ser detetadas através da análise espectral, nomeadamente, através do periodograma, que não é mais do que uma forma alternativa para se estimar a função densidade espectral.

5.3.3 Análise global

Após ter sido exemplificado o processo de construção/seleção dos modelos, através dos casos particulares das séries do comércio a retalho e do *e-commerce* em Portugal (Secções 5.3.1 e 5.3.2), resta analisar os resultados obtidos para as várias séries em estudo. Os modelos selecionados encontram-se resumidos nas Tabelas 5.29 a 5.33 e podem ser consultados, em mais detalhe, no Apêndice C.

A seleção do modelo de Holt-Winters mais adequado para explicar cada uma das séries (Tabela 5.29), de entre os dois disponíveis modelar a sazonalidade – aditivo e multiplicativo –, é realizada, exclusivamente, com base no erro quadrático médio a

1-passo (*in-sample*). No entanto, como esperado, estas escolhas vão, de certa forma, ao encontro do que é observado nas Figuras 5.2 e 5.3 (Secção 5.2). De facto, às séries cuja amplitude sazonal é proporcional ao nível da série é ajustado um modelo multiplicativo (sazonalidade multiplicativa), e às restantes, cuja amplitude sazonal é independente do nível da série, é ajustado um modelo aditivo (sazonalidade aditiva).

Tabela 5.29: Estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters às séries em estudo.

Comércio a retalho				País	E-commerce			
$\hat{\gamma}$	$\hat{\beta}$	$\hat{\alpha}$	Modelo		Modelo	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
0,3024	0,0336	0,2087	aditivo	DE	multiplicativo	0,3605	0,0437	0,6417
0,4829	0,0749	0,3641	multiplicativo	ES	multiplicativo	0,5371	0,0343	0,9737
0,5433	0,0329	0,3365	multiplicativo	FR	multiplicativo	0,3571	0,0145	0,9728
0,8463	0,0157	0,1798	aditivo	IT	aditivo	0,0000	0,0000	1,0000
0,2854	0,2238	0,2168	aditivo	NL	multiplicativo	0,2346	0,0729	0,7062
0,6259	0,0610	0,3469	multiplicativo	PT	multiplicativo	0,1823	0,0589	0,6781
0,4881	0,0000	0,3093	multiplicativo	UK	multiplicativo	0,4942	0,0368	1,0000

Pela análise da Tabela 5.29, conclui-se que, em alguns casos, as estimativas para as constantes de alisamento apresentam valores próximos de zero, chegando mesmo, por vezes, a atingir esse extremo. De acordo com as equações de atualização da Secção 4.2.3, quando $\hat{\beta} \approx 0$, o declive num dado instante fica determinado, quase na totalidade, pelo seu valor no instante anterior e, portanto, se ocorre a igualdade ($\hat{\beta} = 0$), o declive é constante. Da mesma forma, se $\hat{\alpha} = 0$, o nível é também constante. Por exemplo, no caso particular da série do comércio a retalho no Reino Unido, o declive é estimado como constante ($\hat{\beta} = 0$) e igual a 0,2271.

Por outro lado, algumas das estimativas para as constantes de alisamento γ são próximas do outro extremo, o valor 1, o que indica que o fator sazonal para um dado instante t praticamente não depende do fator sazonal obtido para o instante $t - 12$ ($s = 12$ é o período sazonal). Assim, se $\hat{\gamma} = 1$, pode dizer-se que os fatores sazonais são, à partida, todos diferentes entre si (caso $\hat{\gamma} = 0$, os fatores sazonais iniciais não seriam corrigidos ao longo da previsão, sendo, por isso, iguais de 12 em 12 observações). Por exemplo, no caso da série do *e-commerce* em Itália, tanto o nível como o declive são constantes ($\hat{\alpha} = \hat{\beta} = 0$) e, portanto, as únicas oscilações são provocadas pelos fatores sazonais, que, uma vez que $\hat{\gamma} = 1$, são determinados sem qualquer influência do passado. De facto, pela análise gráfica da mesma série (Figura 5.3, Secção 5.2), é possível observar que, ao longo da sua tendência (linear), a amplitude sazonal vai sofrendo várias alterações, não permitindo que a componente

sazonal seja explicada com recurso a valores do passado.

No que respeita aos modelos SARIMA (Tabelas 5.30 e 5.31), a seleção dos mesmos segue a metodologia Box-Jenkins, baseando-se a escolha entre dois ou mais modelos candidatos no critério AIC.

Tabela 5.30: Estimativas dos parâmetros dos modelos SARIMA ajustados às séries dos logaritmos do comércio a retalho.

País	Modelo	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}_1$	$\hat{\nu}_1$	$\hat{\eta}_1$	$\hat{\eta}_2$
DE*	SARIMA(2, 1, 1)(0, 1, 1) ₁₂	-0,6687	-0,5167	-0,3182	-	-0,6963	-
ES*	SARIMA(2, 1, 0)(0, 1, 1) ₁₂	-0,7339	-0,3786	-	-	-0,6376	-
FR*	SARIMA(2, 1, 0)(0, 1, 1) ₁₂	-0,7112	-0,4841	-	-	-0,5893	-
IT	SARIMA(0, 1, 1)(1, 0, 1) ₁₂	-	-	-0,7051	0,9894	-0,2578	-
NL*	SARIMA(2, 1, 0)(0, 1, 2) ₁₂	-0,8869	-0,6425	-	-	-0,5899	-0,2967
PT	SARIMA(2, 1, 0)(1, 0, 1) ₁₂	-0,6810	-0,3283	-	0,9967	-0,5448	-
UK	SARIMA(0, 1, 1)(0, 1, 1) ₁₂	-	-	-0,5942	-	-0,7162	-

* Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela 5.31: Estimativas dos parâmetros dos modelos SARIMA ajustados às séries dos logaritmos do e-commerce.

País	Modelo	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}_1$	$\hat{\nu}_1$	$\hat{\eta}_1$	$\hat{\eta}_2$
DE*	SARIMA(2, 1, 0)(1, 0, 1) ₁₂	-0,8498	-0,5279	-	0,9778	-0,4668	-
ES	SARIMA(2, 1, 0)(0, 1, 1) ₁₂	-0,4071	-0,3521	-	-	-0,6698	-
FR*	SARIMA(2, 1, 0)(1, 0, 2) ₁₂	-0,6361	-0,3927	-	0,9895	-0,1663	-0,2344
IT*	SARIMA(1, 1, 1)(1, 0, 0) ₁₂	0,2478	-	-0,9879	0,9382	-	-
País	Modelo	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\eta}_1$	$\hat{\eta}_2$
NL	SARIMA(0, 1, 1)(1, 0, 0) ₁₂	-0,6575	-	0,7105	-	-	-
PT*	SARIMA(0, 1, 2)(2, 0, 2) ₆	-0,5826	-0,1992	-0,0881	0,8879	0,4172	-0,3389
UK	SARIMA(0, 1, 1)(1, 0, 1) ₁₂	-0,5564	-	0,9840	-	-0,4695	-

* Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Para a maioria dos modelos escolhidos, os resíduos apresentam um comportamento idêntico ao de um ruído branco, exceção feita para alguns (ver Tabelas 5.30 e 5.31), de acordo com o teste de Ljung-Box, relativamente à independência. De facto, pela análise inferencial, a independência dos erros é rejeitada, mas, nesses casos, as FAC e FACP dos resíduos apresentam comportamentos semelhantes às FAC e FACP de um ruído branco, com correlações significativas, de forma geral, não superiores a 0,2.

O único modelo SARIMA que não se ajusta adequadamente aos dados é o que modela a série do *e-commerce* em Itália, uma vez que a série dos resíduos apresenta volatilidade económica. Essa característica da série dos resíduos resulta no incumprimento do pressuposto de normalidade dos erros (e, naturalmente, de homocedasticidade da variância) e, por isso, conclui-se que a metodologia Box-Jenkins não é a melhor para modelar uma série com um comportamento idêntico ao da série do *e-commerce* em Itália, isto é, com várias variâncias, ainda que localmente constantes. As mudanças na variância desta série poderiam ser estudadas, por exemplo, através da análise de *change-points* (na variância) (mais informações sobre a análise de *change-points* em Costa et al. (2016)).

A metodologia associada à construção dos modelos de regressão linear múltipla (Tabelas 5.32 e 5.33) é, de entre todas as metodologias estudadas, a mais subjetiva. De facto, apesar de existir um critério de seleção comum a todas as séries (escolher o modelo com o maior coeficiente de determinação ajustado), a determinação do modelo para estimação da componente tendência é feita apenas com base na análise gráfica, não tendo sido comprovada por nenhum tipo de análise inferencial. Na verdade, as mudanças observadas na tendência podem até não ser significativas e, neste caso, a análise de *change-points* (na média) seria uma ferramenta útil para validar as escolhas feitas.

Em alguns modelos, nomeadamente para as séries do comércio a retalho para Espanha, Portugal e Alemanha, a substituição de uma tendência quadrática por duas lineares resulta numa melhoria significativa. Como referido anteriormente, de acordo com Cowpertwait & Metcalfe (2009), em modelos de regressão, as previsões devem basear-se num modelo linear para a tendência, estimado apenas para os dados mais recentes, em vez de em modelos polinomiais de ordens superiores (como é o caso de um modelo quadrático). De facto, além de poderem produzir melhores previsões, duas tendências lineares são mais facilmente interpretáveis do que uma tendência quadrática.

Relativamente aos resíduos dos modelos de regressão linear, estes cumprem, de forma geral, os pressupostos de normalidade, média nula e variância constante dos erros. Quanto ao pressuposto de independência, este só não é verificado nos modelos que apresentam sazonalidade por explicar, ou seja, nas séries de Itália, e nas séries do *e-commerce* em França e em Portugal.

Nas séries com sazonalidade não explicada, os *lags* sazonais podem camuflar a existência de um outro qualquer padrão autorregressivo. Por exemplo, nas séries de Itália, as FAC e FACP não permitem identificar correlações significativas em *lags*

não sazonais, pelo que se assume a não existência de autocorrelação. No entanto, as eventuais periodicidades escondidas nos dados podem ser detetadas através da análise espectral, nomeadamente, através do periodograma, que não é mais do que uma forma alternativa para se estimar a função densidade espectral.

Tabela 5.32: Estrutura dos modelos de regressão ajustados às séries dos logaritmos do comércio a retalho.

País	Estrutura do modelo
DE	Tendência: duas tendências lineares Evento: fevereiro/2009 Estrutura de autocorrelação: AR(3) Variáveis removidas: $t:\cos 3$; $t:\sen 3$; $t:\cos 2,4$; $t:\sen 2,4$; $t:\cos 2$
ES	Tendência: três tendências lineares Eventos: março/2008 e setembro/2012 Estrutura de autocorrelação: AR(3) Variáveis removidas: evento2; $\cos 2$; $t:\cos 4$; $t:\sen 4$; $t:\cos 2$
FR	Tendência: duas tendências lineares Evento: novembro/2008 Estrutura de autocorrelação: AR(3) Variáveis removidas: $t:\cos 3$; $t:\sen 3$; $t:\cos 2$
IT	Tendência: quatro tendências lineares Eventos: outubro/2008, abril/2012 e setembro/2014 Estrutura de autocorrelação: nenhuma Revela sazonalidade não explicada (lag 12) Variáveis removidas: $t:\text{evento1}$
NL	Tendência: uma tendência linear seguida de duas quadráticas Eventos: julho/2002 e abril/2008 Estrutura de autocorrelação: AR(3) Variáveis removidas: $t:\cos 12$; $t:\sen 12$; $t:\cos 4$; $t:\sen 4$; $t:\cos 3$; $t:\sen 3$; $t:\cos 2,4$; $t:\sen 2,4$; $t:\cos 2$
PT	Tendência: três tendências lineares Eventos: janeiro/2008 e abril/2014 Estrutura de autocorrelação: AR(3) Variáveis removidas: $t:\text{evento2}$; $t:\cos 2$
UK	Tendência: duas tendências lineares Evento: outubro/2008 Estrutura de autocorrelação: AR(3) Variáveis removidas: $t:\cos 2,4$; $t:\sen 2,4$

Tabela 5.33: Estrutura dos modelos de regressão ajustados às séries dos logaritmos do *e-commerce*.

País	Estrutura do modelo
DE	Tendência: três tendências lineares Eventos: outubro/2011 e maio/2015 Estrutura de autocorrelação: AR(3) Variáveis removidas: cos3; sen3
ES	Tendência: uma tendência linear seguida de uma quadrática Eventos: abril/2008 Estrutura de autocorrelação: AR(3) Variáveis removidas: $t:\cos 12$; $t:\sin 12$; $t:\cos 3$; $t:\sin 3$; $t:\cos 2$
FR	Tendência: duas tendências lineares Evento: agosto/2015 Estrutura de autocorrelação: AR(3) Revela sazonalidade não explicada (<i>lag</i> 12) Variáveis removidas: evento, $t:\cos 2,4$; $t:\sin 2,4$
IT	Tendência: linear Evento: sem evento Estrutura de autocorrelação: nenhuma Revela sazonalidade não explicada (<i>lag</i> 12) Variáveis removidas: cos2; $t:\cos 4$; $t:\sin 4$; $t:\cos 2,4$; $t:\sin 2,4$; $t:\cos 2$
NL	Tendência: duas tendências lineares Eventos: outubro/2011 Estrutura de autocorrelação: AR(1) Variáveis removidas: cos3; sen3; cos2,4; sen2,4; cos2; $t:\cos 2,4$; $t:\sin 2,4$; $t:\cos 2$
PT	Tendência: três tendências lineares Eventos: julho/2004 e fevereiro/2013 Estrutura de autocorrelação: começa em AR(3), mas termina em AR(1) Revela sazonalidade não explicada (<i>lag</i> 6) Variáveis removidas: cos2; $t:\text{evento}2$; $t:\cos 6$; $t:\sin 6$; $t:\cos 4$; $t:\sin 4$; $t:\cos 3$; $t:\sin 3$; $t:\cos 2,4$; $t:\sin 2,4$; $t:\cos 2$
UK	Tendência: quadrática Evento: sem evento Estrutura de autocorrelação: AR(2) Variáveis removidas: cos2; $t:\cos 3$; $t:\sin 3$; $t:\cos 2,4$; $t:\sin 2,4$; $t:\cos 2$

5.4 Comparação dos métodos de previsão

Depois da aplicação dos diferentes métodos de previsão aos dados, o último passo é identificar qual deles é o mais adequado à situação em estudo. Tal como referido na Secção 4.4, não existe nenhuma medida que seja consensual para ser aplicada

a todas as situações de previsão e, portanto, é aconselhada a utilização de vários critérios para avaliar a qualidade preditiva de um modelo/método de previsão. Para efeitos comparativos são, então, utilizadas cinco medidas de avaliação: uma medida dependente da escala, EQM, e a sua correspondente na mesma escala dos dados, REQM; uma medida percentual, EPAM; uma medida escalada, EEAM; e uma medida alternativa, a estatística U de Theil. Além de serem calculadas para a série de teste, através dos 14 erros de previsão, estas medidas são também determinadas para a série de treino, fazendo uso dos resíduos do modelo em questão. Os resultados obtidos para as séries do comércio a retalho e do *e-commerce* podem ser consultados nas Tabelas 5.34 e 5.35, respetivamente.

Os cinco critérios de avaliação produzem conclusões idênticas, exceção apenas para as séries do *e-commerce* em França e Itália, onde o EPAM e o EEAM fornecem uma conclusão distinta das restantes medidas.

Da análise das Tabelas 5.34 e 5.35, pode ver-se que o modelo que melhor explica o comportamento dos dados (série de treino) é, de forma geral, o de regressão linear múltipla. No entanto, no que diz respeito à previsão (série de teste), esse é o modelo que apresenta piores resultados. Esta conclusão vai ao encontro do que se esperava, uma vez que os modelos de regressão linear se constroem, principalmente, com base em componentes determinísticas, assumindo que o comportamento da série se mantém imutável no futuro, e, portanto, não são tão aconselhados para previsões a longo prazo. De facto, Chu & Zhang (2003) referem, a respeito dos modelos de regressão, que “se as componentes do modelo estão em constante mudança, como acontece em muitas séries temporais económicas, os modelos determinísticos podem não ser apropriados” e desaconselham o uso de funções trigonométricas para modelar dados do segmento do retalho.

De acordo com os resultados obtidos para a série de teste, a escolha do melhor método de previsão não é consensual. Efetivamente, embora o modelo SARIMA seja considerado o método mais preciso para cinco das sete séries do comércio a retalho, no que respeita às séries do *e-commerce*, não existe um método que se destaque.

Além disso, comparando a ordem de grandeza das medidas de avaliação calculadas para as séries do comércio a retalho com as das séries do *e-commerce*, facilmente se conclui que, de modo geral, as últimas apresentam valores bem mais elevados, possivelmente devido à sua variabilidade.

Tendo em consideração a ordem de grandeza e a inconstância dos resultados obtidos para as séries do *e-commerce*, talvez nenhum dos três métodos utilizados seja, na verdade, apropriado para modelar séries com tais características.

Tabela 5.34: Medidas de avaliação calculadas para as séries do comércio a retalho, no período de treino (janeiro de 2000 a dezembro de 2016) e no período de teste (janeiro de 2017 a fevereiro de 2018), com base nos resultados obtidos na aplicação dos três métodos de previsão em estudo.

País	Modelo	Série de treino						Série de teste					
		EQM	REQM	EPAM	EEAM	U-Theil	U-Theil	EQM	REQM	EPAM	EEAM	U-Theil	U-Theil
DE	SARIMA	3,5712	1,8898	1,6216	0,6074	0,2194	0,2194	5,1472	2,2687	1,7410	0,6084	0,2418	0,2418
	Holt-Winters	4,3834	2,0937	1,8743	0,6940	0,2412	0,2412	5,3537	2,3138	1,7875	0,6320	0,2586	0,2586
	Regressão	2,5064	1,5832	1,3982	0,5250	0,1803	0,1803	10,5210	3,2436	2,4915	0,8847	0,3354	0,3354
ES	SARIMA	4,4821	2,1171	1,4633	0,3425	0,1885	0,1885	3,2586	1,8052	1,4236	0,5928	0,1848	0,1848
	Holt-Winters	5,5033	2,3459	1,7175	0,4096	0,2088	0,2088	4,6211	2,1497	1,6898	0,6957	0,2122	0,2122
	Regressão	2,6505	1,6280	1,2161	0,2839	0,1452	0,1452	7,5084	2,7402	2,2001	0,9375	0,2749	0,2749
FR	SARIMA	2,2018	1,4838	1,3272	0,4049	0,1164	0,1164	4,0874	2,0217	1,4727	0,4407	0,1330	0,1330
	Holt-Winters	2,5204	1,5876	1,4447	0,4665	0,1238	0,1238	4,8154	2,1944	1,6128	0,4859	0,1458	0,1458
	Regressão	1,5561	1,2474	1,1998	0,3632	0,0967	0,0967	5,2393	2,2890	1,6903	0,5057	0,1477	0,1477
IT	SARIMA	4,5915	2,1428	1,6587	0,7731	0,1541	0,1541	2,0575	1,4344	1,2119	2,1901	0,1368	0,1368
	Holt-Winters	4,4141	2,1010	1,6224	0,7578	0,1506	0,1506	1,5008	1,2251	1,0207	1,8622	0,1176	0,1176
	Regressão	2,5751	1,6047	1,3372	0,6130	0,1152	0,1152	6,2539	2,5008	2,2676	4,0183	0,2382	0,2382
NL	SARIMA	3,8463	1,9612	1,5251	0,4989	0,2119	0,2119	3,0495	1,7463	1,3131	0,3596	0,1884	0,1884
	Holt-Winters	6,5603	2,5613	2,0494	0,6911	0,2746	0,2746	5,7452	2,3969	1,8355	0,4966	0,2683	0,2683
	Regressão	2,9245	1,7101	1,3593	0,4430	0,1850	0,1850	7,2928	2,7005	2,0227	0,5680	0,2970	0,2970
PT	SARIMA	5,6398	2,3748	1,6504	0,5196	0,1654	0,1654	18,1862	4,2645	3,6993	0,8317	0,3735	0,3735
	Holt-Winters	6,1686	2,4837	1,7152	0,5693	0,1730	0,1730	9,0174	3,0029	2,3260	0,5262	0,2604	0,2604
	Regressão	3,3801	1,8385	1,2992	0,4109	0,1278	0,1278	13,9676	3,7373	3,0549	0,6845	0,3363	0,3363
UK	SARIMA	1,2944	1,1377	0,9407	0,3274	0,1368	0,1368	4,2604	2,0641	1,7064	0,4869	0,1855	0,1855
	Holt-Winters	1,4084	1,1868	1,0176	0,3550	0,1426	0,1426	4,4513	2,1098	1,7335	0,5024	0,1912	0,1912
	Regressão	0,8294	0,9107	0,7869	0,2707	0,1101	0,1101	6,8705	2,6212	1,9693	0,5850	0,2479	0,2479

Tabla 5.35: Medidas de avaliação calculadas para as séries do e-commerce, no período de treino (janeiro de 2000 a dezembro de 2016) e no período de teste (janeiro de 2017 a fevereiro de 2018), com base nos resultados obtidos na aplicação dos três métodos de previsão em estudo.

País	Modelo	Série de treino					Série de teste				
		EQM	REQM	EPAM	EEAM	U-Theil	EQM	REQM	EPAM	EEAM	U-Theil
DE	SARIMA	20,5765	4,5361	4,5594	0,5672	0,4825	155,6450	12,4758	8,8266	2,9299	0,8354
	Holt-Winters	20,9007	4,5717	4,8669	0,5757	0,5249	75,4853	8,6882	6,1750	2,0460	0,5906
	Regressão	8,2339	2,8695	3,2099	0,3813	0,3309	43,1640	6,5699	3,9823	1,3737	0,4453
ES	SARIMA	62,7431	7,9211	6,3205	0,5082	0,4922	419,5723	20,4835	11,2356	0,9700	0,6031
	Holt-Winters	77,9908	8,8312	7,3078	0,5697	0,5731	429,7333	20,7300	11,5468	0,9759	0,6546
	Regressão	40,4500	6,3600	5,5710	0,4347	0,4103	1228,3860	35,0483	19,8395	1,8136	0,9960
FR	SARIMA	11,1286	3,3360	3,2371	0,6147	0,2752	128,5986	11,3401	7,0722	0,7973	0,6248
	Holt-Winters	13,1270	3,6231	3,4753	0,6464	0,2937	123,8519	11,1289	7,3656	0,8230	0,6245
	Regressão	5,6399	2,3749	2,6109	0,4757	0,2035	127,5319	11,2930	6,4791	0,7310	0,6365
IT	SARIMA	75,5831	8,6939	5,9372	0,9926	0,3452	16,0817	4,0102	4,0541	2,1521	0,1470
	Holt-Winters	82,6041	9,0887	5,6365	0,9156	0,3717	7,2666	2,6957	2,5647	1,3238	0,1046
	Regressão	105,8345	10,2876	8,9337	1,4799	0,3996	279,1053	16,7064	14,0082	8,0584	0,6264
NL	SARIMA	14,1926	3,7673	4,3249	0,3967	0,7361	41,0015	6,4032	3,7112	0,2725	0,5151
	Holt-Winters	12,5451	3,5419	4,0262	0,3683	0,7329	19,2825	4,3912	2,3663	0,1796	0,3553
	Regressão	8,2689	2,8756	3,4450	0,3129	0,5672	92,9048	9,6387	6,3278	0,4578	0,8248
PT	SARIMA	319,4194	17,8723	7,6068	0,6588	0,5258	257,6532	16,0516	16,2262	0,9871	1,0355
	Holt-Winters	420,3237	20,5018	8,4697	0,7405	0,5866	265,2426	16,2863	16,7628	1,0299	1,0215
	Regressão	258,4669	16,0769	6,8676	0,5938	0,4926	222,5327	14,9175	14,2797	0,8822	0,9513
UK	SARIMA	9,1149	3,0191	3,6277	0,3424	0,3860	135,2233	11,6286	7,1636	0,5709	0,5411
	Holt-Winters	10,0033	3,1628	3,8005	0,3566	0,4053	161,6122	12,7127	8,9067	0,7392	0,5815
	Regressão	7,5265	2,7434	2,9159	0,2818	0,3173	253,0859	15,9087	9,8223	0,8348	0,6659

Ainda assim, considera-se que o modelo SARIMA é o que tende a apresentar melhor desempenho, uma vez que é o método mais preciso para a previsão de sete das catorze séries, seguindo-se o método de Holt-Winters, com resultados favoráveis para cinco delas. Também segundo Alon et al. (2001), o modelo de Box-Jenkins é, dentro dos métodos tradicionais, visto como a melhor opção. Nesse estudo, os modelos de regressão são, uma vez mais, considerados os menos adequados e, quando as condições económicas são relativamente estáveis, o método de Holt-Winters mostra-se uma escolha viável.

Avaliada a precisão das previsões pontuais, é também importante perceber qual a eficácia das previsões intervalares. De facto, em teoria, os intervalos de previsão são calculados a uma confiança de 95%, o que significa que 95% dos mesmos deve incluir a observação verdadeira. Desta forma, as previsões intervalares mais eficazes são aquelas cuja taxa de cobertura efetiva mais se aproxima de 95%. No entanto, é importante referir que os intervalos de previsão são determinados com base na série de teste que, para cada série temporal em estudo, contém apenas 14 observações e, portanto, a análise das taxas de cobertura dos intervalos deve ser feita com algum cuidado. Além disso, as taxas de cobertura podem também ser elevadas para um determinado método por este produzir intervalos de previsão com uma grande amplitude, como acontece no caso do método de Holt-Winters aplicado à série do *e-commerce* em Portugal. Assim, a análise deve ser feita de forma global, considerando que a avaliação das taxas de cobertura dos intervalos de previsão é apenas mais um indicador de desempenho.

Na Tabela 5.36 podem ser consultadas as taxas de cobertura efetivas dos intervalos de previsão dos três métodos, obtidas para cada série individual, para cada tipo de série e para o conjunto de todas as séries.

Analisando a Tabela 5.36, percebe-se, novamente, uma diferença entre os resultados para as séries do comércio a retalho e para as do *e-commerce*. Na verdade, para as últimas, apenas os intervalos dos modelos SARIMA apresentam uma taxa de cobertura próxima da teórica (94,9%).

Tanto para as séries do comércio a retalho como para as do *e-commerce*, os intervalos de previsão obtidos para o modelo de regressão são também os menos precisos, com taxas de cobertura de apenas 85,7% e 76,5%, respetivamente.

Relativamente aos intervalos de previsão do método de Holt-Winters, estes não parecem ser os mais corretos, apesar de apresentarem taxas de cobertura razoáveis. De facto, para a maioria dos casos, o pressuposto de normalidade dos erros de previsão, calculados para cada horizonte temporal h , não é verificado, o que leva

a concluir que os intervalos de previsão determinados através da expressão (4.30) (Secção 4.2.5) não são os melhores para este tipo de dados. Talvez o facto do pressuposto ser validado em praticamente todos os casos das séries do comércio a retalho (exceção apenas para 6 dos 98 intervalos calculados) seja a justificação para a obtenção de resultados tão favoráveis para esse tipo de série.

A melhor cobertura é garantida pelos intervalos de previsão dos modelos SARIMA, sendo a que mais se aproxima da teórica (95%), tanto para o conjunto de todas as séries como para cada um dos dois tipos de série em estudo.

Tabela 5.36: Taxas de cobertura (%) efetiva dos intervalos de previsão a 95%.

		SARIMA	Holt-Winters	Regressão
Comércio a retalho	DE	100,0	100,0	85,7
	ES	100,0	100,0	85,7
	FR	100,0	92,9	92,9
	IT	100,0	100,0	92,9
	NL	100,0	100,0	92,9
	PT	100,0	100,0	78,6
	UK	85,7	85,7	71,4
			98,0	96,9
E-commerce	DE	100,0	92,9	100,0
	ES	100,0	71,4	50,0
	FR	92,9	64,3	71,4
	IT	100,0	100,0	85,7
	NL	100,0	100,0	100,0
	PT	85,7	100,0	71,4
	UK	85,7	64,3	57,1
			94,9	84,7
TOTAL		97,4	90,8	81,1

Consequentemente, de acordo com a precisão das previsões pontuais e com a eficácia dos intervalos de previsão, considera-se que, de entre as três metodologias utilizadas, a mais adequada para fazer previsão em dados do segmento do retalho é a metodologia Box-Jenkins. Ainda assim, o método de Holt-Winters parece ser uma alternativa viável, sendo, no entanto, recomendada a utilização de outro tipo de abordagem para o cálculo dos respetivos intervalos de previsão.

Capítulo 6

Conclusão

Este trabalho centrou-se na análise de séries temporais económicas, em particular do segmento do retalho, através do estudo de séries mensais de índices de volumes de negócios (TOVT) observadas em sete países distintos (Portugal, Espanha, França, Itália, Alemanha, Países Baixos e Reino Unido), entre janeiro de 2000 e fevereiro de 2018. O objetivo principal deste estudo foi avaliar a precisão de vários métodos de previsão de séries temporais, aplicados a dados do segmento do retalho, de forma a identificar as metodologias mais adequadas. Para isso realizou-se, então, um estudo comparativo da capacidade preditiva dos modelos SARIMA, dos modelos de decomposição clássica associados a modelos de regressão linear múltipla e dos métodos de alisamento exponencial (Holt-Winters), através da avaliação de cinco medidas distintas (EQM, REQM, EPAM, EEAM e estatística U-Theil). Foram também calculados intervalos de previsão para cada uma das metodologias e comparadas as taxas de cobertura efetivas dos mesmos.

Da análise efetuada foi possível concluir que os modelos de regressão linear múltipla foram, no que diz respeito à previsão de dados do segmento do retalho, os que apresentaram menor precisão. Apesar disso, a escolha do “melhor” método de previsão não foi imediata. Enquanto que para as séries do comércio a retalho o modelo SARIMA foi identificado como o modelo mais preciso para cinco das sete séries, no que respeita às séries do *e-commerce* nenhum método se destacou.

Relativamente aos intervalos de previsão, verificou-se que a melhor cobertura foi garantida pelos intervalos dos modelos SARIMA, sendo a que mais se aproxima da teórica (95%). Os intervalos de previsão obtidos para os modelos de regressão linear múltipla foram também os menos adequados, com taxas de cobertura bastante inferiores a 95%. Apesar de apresentarem taxas de cobertura razoáveis, os intervalos de previsão do método de Holt-Winters não são estatisticamente rigorosos, uma vez

que, na sua maioria, falham o pressuposto da normalidade dos erros de previsão.

Em suma, à semelhança do que é concluído por Alon et al. (2001) e Chu & Zhang (2003), este estudo indicou que os modelos de regressão linear múltipla não são os mais recomendados para a previsão de dados do segmento do retalho, apresentando, de forma geral, os piores resultados. De entre as três metodologias utilizadas, a de Box-Jenkins é identificada como a mais adequada (precisa) para determinar previsões, tanto pontuais como intervalares, em dados do segmento do retalho. O método de Holt-Winters é também indicado como uma alternativa viável à previsão destas séries, embora os seus intervalos de previsão não sejam estatisticamente rigorosos.

6.1 Sugestões para trabalho futuro

Com o estudo desenvolvido surgiu um conjunto de questões, que não puderam ser resolvidas no âmbito deste trabalho.

Assim, para futura investigação, é de considerar a realização do mesmo estudo recorrendo a dados com ajustamento de calendário, de forma a aferir se, nesse caso, os resultados diferem. De facto, de acordo com Caiado (2016), “em muitas situações, as previsões de vendas dependem do número de dias úteis de cada semana ou mês, ou do número de fins de semana de cada mês” e, portanto, estas “devem ser ajustadas de modo a captar essas variações, caso contrário, podem ser enviesadas”.

Relativamente às metodologias adotadas, recomendam-se também algumas alterações. Efetivamente, para o método de Holt-Winters, as constantes de alisamento são determinadas por otimização do erro de previsão a 1-passo, através da função *HoltWinters* no ambiente R. No entanto, uma vez que o horizonte de previsão é, neste caso, superior a 1, sugere-se que as constantes de alisamento sejam otimizadas com base no erro de previsão a h -passos, para cada passo h . Também os intervalos de previsão propostos, pelo não cumprimento do pressuposto de normalidade exigido aos erros de previsão a h -passos, não se revelam os mais adequados para este tipo de dados. De forma a contornar essa situação, recomenda-se o estudo de métodos de reamostragem (*bootstrap*) para calcular previsões intervalares (abordagem já considerada em Cordeiro (2011)).

Quanto aos modelos de regressão linear múltipla, sugere-se a utilização de métodos de *change-points* para verificar, através de inferência estatística, as mudanças na média e, quando aplicável, na variância (e.g., série do *e-commerce* em Itália). Na verdade, os eventos considerados são determinados apenas com base na análise gráfica que pode ser um pouco subjetiva e levar à consideração de mudanças (eventos)

não significativas. Além disso, algumas séries dos resíduos indicam a existência de periodicidades escondidas (sazonalidade não explicada), que podem ser detetadas através da análise espectral, nomeadamente através do periodograma.

Apesar dos resultados terem por base a análise de 14 séries temporais com comportamentos distintos, a aplicação de uma metodologia de validação cruzada, como a adotada em Chu & Zhang (2003), pode também ser útil para testar a robustez dos modelos, de forma a perceber se os resultados obtidos podem ser replicados para outras séries ou se só se verificam para estes casos concretos.

Por fim, é importante referir que, sempre que foi aplicada uma transformação logarítmica aos dados, as previsões geradas pelos modelos (de regressão e SARIMA) foram revertidas para as unidades originais através da exponenciação. No entanto, para variâncias dos resíduos (diga-se σ^2) elevadas, essa restituição não deve ser feita de forma tão simples. De facto, admitindo que os erros da série logaritmizada seguem uma distribuição Normal (e, conseqüentemente, a série logaritmizada também), a série original tem uma distribuição Log-Normal e, portanto, as previsões resultantes da exponenciação estão enviesadas. Para evitar essa situação, deve aplicar-se um fator de correção de $e^{\frac{1}{2}\sigma^2}$, passando as previsões nas unidades originais a serem calculadas por $\exp\left[\hat{Y}_t + \frac{1}{2}\sigma^2\right]$ (mais detalhes em Murteira et al. (1993), Cowpertwait & Metcalfe (2009)).

Bibliografia

- Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F., Preen, D., Stewart, L., & Fatovich, D. (2015). A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in western australia. *Journal of Biomedical Informatics*, *57*, 62–73.
- Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, *7*, 136–144.
- Akaike, H. (1974). A new look at statistical model identification. *IEEE transactions on Automatic Control*, *19*, 716–723.
- Alon, I., Qi, M., & Sadowski, R. (2001). Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, *8*, 147–156.
- Alpuim, T. (1998). *Séries Temporais*. Associação dos Estudantes da Faculdade de Ciências de Lisboa, 2^a ed.
- Alpuim, T., & El-Shaarawi, A. (2008). On the efficiency of regression analysis with ar(p) errors. *Journal of Applied Statistics*, *35*, 717–737.
- Alpuim, T., & El-Shaarawi, A. (2009). Modeling monthly temperature data in lisbon and prague. *Environmetrics*, *20*, 835–852.
- Armstrong, J. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Klumer.
- Bolin, D., Guttorp, P., Januzzi, A., Jones, D., Novak, M., Podschwit, H., Richardson, L., Särkkä, A., Sowder, C., & Zimmerman, A. (2014). Statistical prediction of global sea level from global temperature. *Statistica Sinica*.
- Box, G., & Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.

- Box, G., Jenkins, G., & Reinsel, G. (2013). *Time Series Analysis: forecasting and control*. John Wiley and Sons.
- Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2016). *Time series analysis: forecasting and control*. New Jersey: John Wiley and Sons, 5^a ed.
- Brown, R. (1959). *Statistical forecasting for inventory control*. New York: McGraw-Hill.
- Brown, R. (1963). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, NJ: Prentice-Hall.
- Caiado, J. (2016). *Métodos de Previsão em Gestão – Com Aplicações em Excel*. Lisboa: Edições Sílabo, 2^a ed.
- Chatfield, C. (2000). *Time-Series Forecasting*. Chapman and Hall/CRC.
- Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*. Chapman and Hall/CRC, 5^a ed.
- Chu, C.-W., & Zhang, G. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics*, 86, 217–231.
- Chu, F.-L. (2009). Forecasting tourism demand with arma-based methods. *Tourism Management*, 30, 740–751.
- Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5, 559–583.
- Cleveland, W., & Terpenning, I. (1992). Graphical methods for seasonal adjustment. *Journal of the American Statistical Association*, 77, 52–62.
- Cordeiro, C. (2011). *Métodos de Reamostragem em Modelos de Previsão*. Ph.D. thesis, Universidade Técnica de Lisboa.
- Costa, M., Gonçalves, A. M., & Teixeira, L. (2016). Change-point detection in environmental time series based on the informational approach. *Electronic Journal of Applied Statistical Analysis*, 9, 267–296.
- Cowpertwait, P., & Metcalfe, A. (2009). *Introductory Time Series with R*. New York: Springer.

- Crabtree, B., Ray, S., Schmidt, P., O'Connor, P., & Schmidt, D. (1990). The individual over time: Time series applications in health care research. *Journal of Clinical Epidemiology*, *43*, 241–260.
- De Gooijer, J., & Hyndman, R. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*, 443–473.
- Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*, 427–431.
- Dritsaki, C. (2015). Forecasting real gdp rate through econometric models: An empirical study from greece. *Journal of International Business and Economics*, *3*, 13–19.
- Enders, W. (2015). *Applied Econometric Time Series*. Alabama: John Wiley and Sons, 4^a ed.
- Engle, R. (1982). Autoregressive conditional heterocedasticity with estimates of the variance of united kingdom inflations. *Econometrica*, *50*, 987–1007.
- Gardner, E. (2006). Exponential smoothing: The state of the art - part ii. *International Journal of Forecasting*, *22*, 637–666.
- Gardner, E., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, *31*, 1237–1247.
- Holt, C. (1957). Forecasting trends and seasonal by exponentially weighted averages. *Office of Naval Research Memorandum*, *20*.
- Hyndman, R. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, *4*, 43–46.
- Hyndman, R., & Athanasopoulos, G. (2014). *Forecasting: Principles and Practice*. OTexts.
URL <http://otexts.org/fpp/>
- Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.
- Hyndman, R., Koehler, A., Ord, J., & Snyder, R. (2008). *Forecasting with Exponential Smoothing. The State Space Approach*. Springer.

- Hyndman, R., Koehler, A., Snyder, R., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454.
- Jebb, A., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in Psychology*, *6*, 727.
- Kim, H., Song, Y., Yi, J., Chung, W., & Nam, C. (2004). Changes in mortality after the recent economic crisis in south korea. *Annals of Epidemiology*, *14*, 442–446.
- Kirchgässner, G., & Wolters, J. (2008). *Introduction to modern time series analysis*. Springer – Verlag.
- Kuvulmaz, J., Usanmaz, S., & Engin, S. (2005). Time-series forecasting by means of linear and nonlinear models. In Springer (Ed.) *MICAI 2005: Advances in Artificial Intelligence*, vol. 3789, (pp. 504–513).
- Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*, 159–178.
- Makridakis, S., Wheelwright, S., & Hyndman, R. (1998). *Forecasting: Methods and Applications*. New York: John Wiley and Sons, 3^a ed.
- Marcellino, M., Stock, J., & Watson, M. (2003). Macroeconomic forecasting in the euro area: country specific versus euro wide information. *European Economic Review*, *47*, 1–18.
- Murteira, B., Muller, D., & Turkman, K. (1993). *Análise de Sucessões Cronológicas*. Lisboa: McGraw-Hill.
- Ng, S., & 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*, 268–281.
- Pacelli, V. (2012). Forecasting exchange rates: A comparative analysis. *International Journal of Business and Social Science*, *3*, 145–156.
- Pan, Y. (2013). Predicting aggregate retail sales using hybrid ARIMA. *Proceedings of the 7th Global Business and Social Science Research Conference*.

- Pegels, C. (1969). Exponential forecasting: Some new variations. *Management Science*, *15*, 311–315.
- Persons, W. (1919). Indices of business conditions. *Review of Economic Statistics*, *1*, 5–107.
- Phillips, P., & Perron, P. (1988). Testing for unit roots in time series regression. *Biometrika*, *75*, 335–346.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL <https://www.R-project.org/>
- Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, *34*, 151–163.
- Said, S., & Dickey, D. (1984). Testing for unit roots in autoregressive moving-average models with unknown order. *Biometrika*, *71*, 599–607.
- Santos, A., Costa Jr., N., & Coelho, L. (2007). Computational intelligence approaches and linear models in case studies of forecasting exchange rates. *Expert Systems with Applications*, *33*, 816–823.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
- Schwert, W. (1989). Test for unit roots: A monte carlo investigation. *Journal of Business and Economic Statistics*, *7*, 147–159.
- Sfetsos, A., & Siriopoulos, C. (2005). Time series forecasting of averaged data with efficient use of information. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, *35*, 738–745.
- Shumway, R., & Stoffer, D. (2006). *Time Series Analysis and Its Applications With R Examples*. New York: Springer, 2^a ed.
- Song, H., & Li, G. (2008). Tourism demand modelling and forecasting – a review of recent research. *Tourism Management*, *29*, 203–220.
- Stanton, J. (2001). Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, *9*(3).

- Stock, J., & Watson, M. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In O. O. U. Press (Ed.) *Cointegration, causality, and forecasting: A festschrift in honor of Clive W. J. Granger*, (pp. 1–44).
- Taylor, J. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19, 715–756.
- Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*. Oxford, UK: Oxford University Press.
- Tsay, R. (2000). Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association*, 95, 638–643.
- Tsay, R. (2010). *Analysis of Financial Time Series*. New Jersey: John Wiley and Sons, 3^a ed.
- Veiga, C., Veiga, C., Puchalski, W., Coelho, L., & 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.
- Vile, J., Gillard, J., Harper, P., & Knight, V. (2016). Time-dependent stochastic methods for managing and scheduling emergency medical services. *Operations Research for Health Care*, 8, 42–52.
- Wagner, A., Soumerai, S., Zhang MS, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 27, 299–309.
- Wang, S. (2006). *Exponential Smoothing for Forecasting and Bayesian Validation of Computer Models*. Ph.D. thesis, Georgia Institute of Technology.
- Wang, T. (2016). Forecast of economic growth by time series and scenario planning method – a case study of shenzhen. *Modern Economy*, 7, 212–222.
- Winters, P. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324–342.
- Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell.

- Ye, L., Yang, G., Van Ranst, E., & Tang, H. (2013). Time-series modeling and prediction of global monthly absolute temperature for environmental decision making. *Advances in Atmospheric Sciences*, *30*, 382–396.
- Yokum, J., & Armstrong, J. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting*, *11*, 591–597.
- Yule, G. (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*, 1–63.
- Yule, G. (1927). On the method of investigating periodicities in disturbed series, with special reference to wölfer's sunspot numbers. *Philosophical Transactions of the Royal Society London, Series A*, *226*, 267–298.
- Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175.
- Zhang, G., Patuwo, B., & Hu, M. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*, 35–62.
- Zhang, G., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, *160*, 501–514.

Bibliografia

Apêndice A

Representação gráfica (logaritmo)

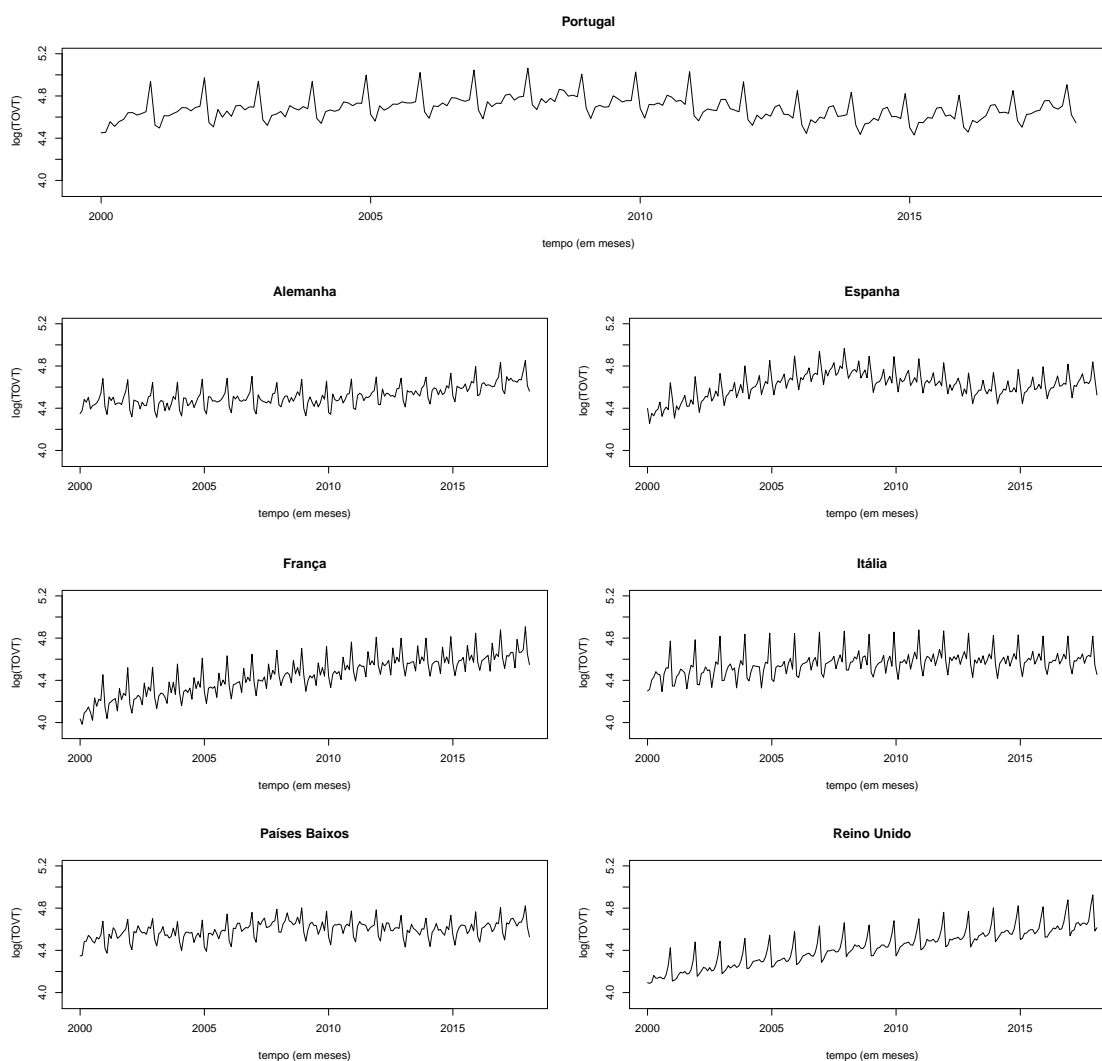


Figura A.1: Representação gráfica das séries dos logaritmos do comércio a retalho, com exceção de veículos automóveis e motocicletas, para os sete países em estudo, no período observado.

Apêndice A. Representação gráfica (logaritmo)

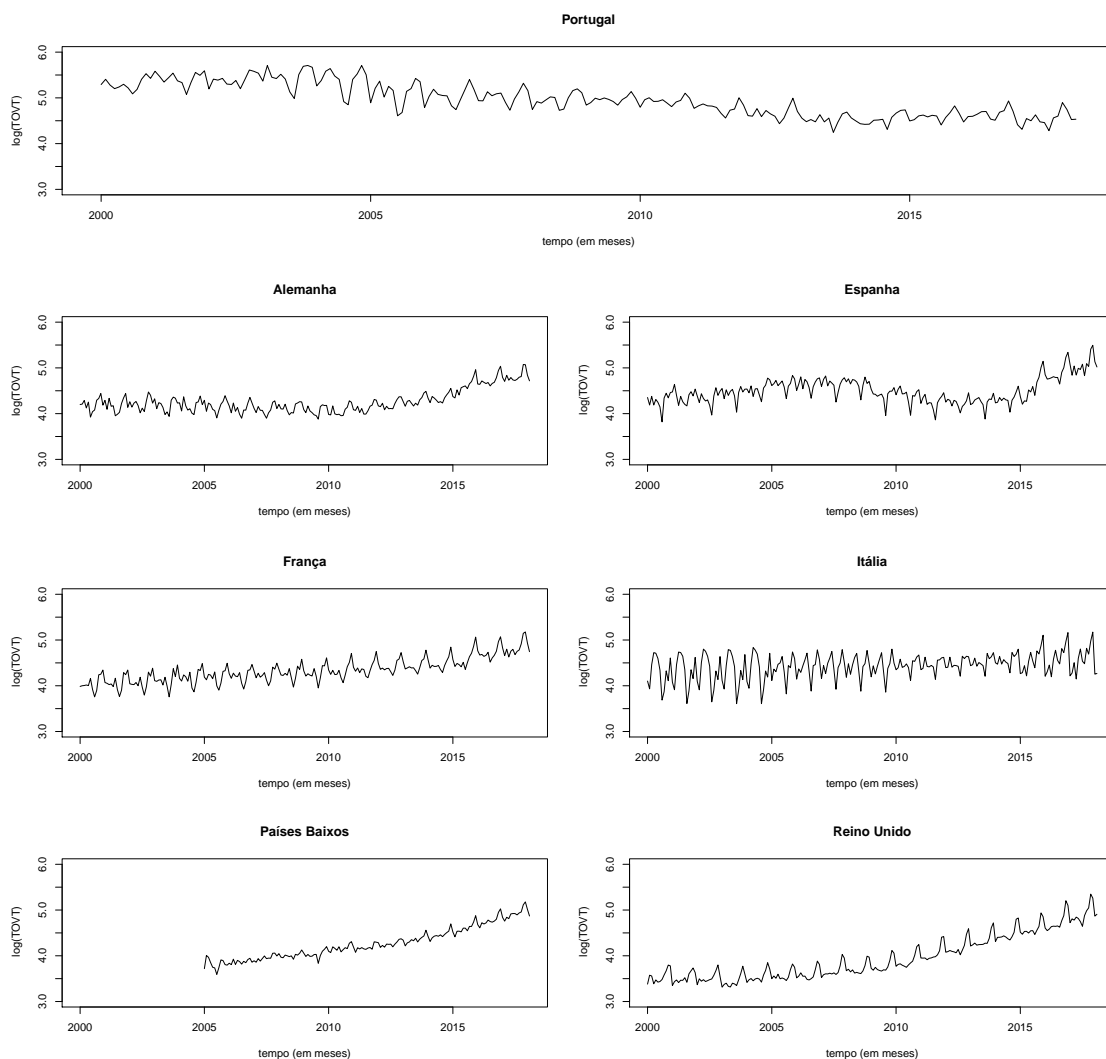


Figura A.2: Representação gráfica das séries dos logaritmos da venda a retalho por correspondência ou via *internet*, para os sete países em estudo, no período observado.

Apêndice B

Medidas descritivas dos dados

Tabela B.1: Medidas descritivas dos índices de volume de negócios, registados no comércio a retalho (com exceção de veículos automóveis e motociclos).

	país						
	DE	ES	FR	IT	NL	PT	UK
início	jan/00	jan/00	jan/00	jan/00	jan/00	jan/00	jan/00
fim	fev/18	fev/18	fev/18	fev/18	fev/18	fev/18	fev/18
n.º obs.	218	218	218	218	218	218	218
mínimo	74,70	70,50	53,70	73,20	77,30	84,00	59,60
1.º quartil	86,13	94,38	75,43	89,80	94,10	99,93	73,70
média	92,06	102,17	86,94	96,45	99,27	108,58	85,49
mediana	90,70	102,05	86,25	95,75	99,30	107,30	85,10
3.º quartil	97,08	108,80	96,80	100,48	104,43	113,80	95,30
máximo	127,90	143,40	135,30	131,30	124,20	158,10	137,70
variância	87,02	156,14	243,51	131,74	70,46	179,85	222,99
desvio padrão	9,33	12,50	15,60	11,48	8,39	13,41	14,93

Apêndice B. Medidas descritivas dos dados

Tabela B.2: Medidas descritivas dos índices de volume de negócios, registados na venda a retalho por correspondência ou via *internet*.

	país						
	DE	ES	FR	IT	NL	PT	UK
início	jan/00	jan/00	jan/00	jan/00	jan/05	jan/00	jan/00
fim	fev/18	fev/18	fev/18	fev/18	fev/18	fev/18	fev/18
n.º obs.	218	218	218	218	158	218	218
mínimo	48,40	45,60	42,80	36,90	36,30	69,80	27,50
1.º quartil	59,43	73,95	64,40	73,33	54,40	104,60	35,00
média	73,70	93,90	79,19	87,71	76,26	151,88	59,75
mediana	68,15	86,90	75,70	86,60	66,30	138,90	43,95
3.º quartil	77,40	105,78	88,25	102,30	91,48	187,45	78,68
máximo	159,70	243,30	177,00	176,60	177,30	301,40	210,20
variância	463,74	868,85	542,33	550,63	865,83	3206,35	1197,58
desvio padrão	21,53	29,48	23,29	23,47	29,43	56,62	34,61

Tabela B.3: Médias anuais e médias de meses homólogos dos índices de volume de negócios, registados no comércio a retalho, com exceção de veículos automóveis e motociclos.

	país							
	DE	ES	FR	IT	NL	PT	UK	
média anual	2000	87,20	81,10	63,68	86,33	89,88	100,56	64,68
	2001	88,35	86,60	68,91	88,25	94,34	106,43	67,36
	2002	86,45	91,90	71,47	90,13	96,90	107,00	70,17
	2003	85,62	97,15	73,30	92,04	95,24	106,59	71,97
	2004	87,45	102,56	76,41	92,53	93,81	110,79	75,34
	2005	88,98	106,87	78,30	94,33	95,33	113,10	76,29
	2006	89,88	112,48	81,28	96,78	100,24	115,98	79,60
	2007	88,48	118,17	84,63	98,30	104,17	118,17	82,72
	2008	90,49	116,57	88,20	99,83	106,96	121,36	85,12
	2009	87,10	108,24	86,90	96,93	101,58	115,49	86,01
	2010	89,31	108,48	90,47	99,51	101,51	116,26	88,18
	2011	91,69	105,93	95,19	101,42	102,11	109,47	91,34
	2012	93,62	100,81	97,98	101,49	100,63	103,52	93,24
	2013	94,90	97,18	98,63	99,60	97,90	101,23	95,47
	2014	96,38	97,68	98,46	99,39	98,48	99,79	98,83
	2015	100,01	100,00	100,01	100,01	99,98	100,00	99,69
	2016	102,90	102,37	102,52	99,60	102,14	102,35	103,13
	2017	107,21	105,37	106,48	100,74	106,13	108,10	107,32
	2018	98,20	100,00	99,45	90,00	96,90	98,00	99,40
média de meses homólogos*	jan	84,74	103,87	83,41	88,22	91,14	98,36	77,66
	fev	81,57	89,38	74,97	83,00	85,86	92,65	79,22
	mar	93,31	98,18	83,15	93,28	98,99	104,10	80,57
	abr	92,46	98,28	84,44	94,69	99,63	102,64	82,83
	mai	91,96	100,75	84,93	97,12	103,62	105,53	83,61
	jun	88,96	102,68	85,62	95,19	102,03	104,86	83,77
	jul	91,14	109,89	76,99	99,66	99,22	114,22	84,73
	ago	88,89	97,14	97,11	87,46	95,88	114,26	82,63
	set	90,19	99,36	85,33	95,96	98,33	108,51	82,87
	out	95,32	103,34	90,37	101,77	102,08	109,52	86,20
	nov	96,77	99,89	86,54	96,45	100,54	108,98	94,48
	dez	110,37	123,90	111,26	125,82	115,17	140,77	108,07

* média de meses homólogos corresponde à média de todas as observações registadas para um determinado mês.

Apêndice B. Medidas descritivas dos dados

Tabela B.4: Médias anuais e médias de meses homólogos dos índices de volume de negócios, registados na venda a retalho por correspondência ou via *internet*.

	país							
	DE	ES	FR	IT	NL	PT	UK	
média anual	2000	67,11	72,34	58,07	76,88	NA	201,23	34,99
	2001	66,08	80,36	59,10	77,73	NA	229,85	33,84
	2002	68,61	76,78	60,03	79,59	NA	221,89	34,03
	2003	66,24	85,74	62,67	79,58	NA	241,36	31,63
	2004	65,48	94,19	66,33	81,28	NA	225,20	35,06
	2005	65,31	104,81	67,23	82,78	45,63	170,46	36,01
	2006	62,17	104,32	68,58	84,97	47,92	157,03	36,88
	2007	59,93	104,42	70,51	86,37	52,25	155,90	40,11
	2008	62,95	106,65	73,11	85,48	55,22	143,83	41,56
	2009	59,00	82,80	73,64	84,60	56,62	143,13	44,51
	2010	58,56	82,95	76,37	87,34	64,56	139,34	49,68
	2011	61,93	70,97	81,01	87,46	65,49	120,71	58,18
	2012	66,92	69,65	84,13	89,85	70,98	107,42	66,05
	2013	72,75	71,07	86,31	92,53	78,18	92,98	77,00
	2014	77,22	75,68	89,49	97,11	86,78	93,43	88,63
	2015	100,00	100,01	99,98	100,01	100,01	100,00	99,77
	2016	114,05	135,65	114,38	103,83	118,50	105,31	117,18
	2017	124,67	160,68	126,63	104,27	138,52	94,58	138,28
	2018	119,15	160,65	126,95	70,90	140,45	92,75	132,65
média de meses homólogos*	jan	75,62	95,87	82,95	68,91	77,68	134,88	55,64
	fev	68,75	93,18	74,22	72,39	72,12	149,75	57,66
	mar	76,12	96,41	76,29	86,81	72,99	157,28	54,07
	abr	70,04	85,53	73,38	91,19	71,32	151,44	53,89
	mai	69,96	93,27	72,88	99,45	73,27	156,62	53,57
	jun	64,88	91,43	79,86	103,60	74,64	149,68	53,11
	jul	64,48	85,71	66,84	82,35	73,52	129,42	52,95
	ago	65,34	67,25	62,38	56,57	71,64	113,79	53,26
	set	75,01	92,00	71,74	85,62	75,82	147,62	57,99
	out	80,53	99,89	88,44	90,90	77,92	170,08	64,38
	nov	88,70	108,88	92,95	102,66	84,52	191,44	82,08
	dez	85,09	117,31	108,45	114,00	89,85	171,57	78,76

* média de meses homólogos corresponde à média de todas as observações registadas para um determinado mês.

Apêndice C

Aplicação dos métodos de previsão

Modelos SARIMA: comércio a retalho

Tabela C.1: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho na Alemanha.

Modelo final ^(*) : SARIMA(2, 1, 1)(0, 1, 1) ₁₂ AIC = -903,65 $\hat{\sigma} = 0,0216$				
	ϕ_1	ϕ_2	θ_1	η_1
estimativa	-0,6687	-0,5167	-0,3182	-0,6963
erro padrão	0,1006	0,0830	0,1164	0,0598

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela C.2: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em Espanha.

Modelo final ^(*) : SARIMA(2, 1, 0)(0, 1, 1) ₁₂ AIC = -919,23 $\hat{\sigma} = 0,0210$			
	ϕ_1	ϕ_2	η_1
estimativa	-0,7339	-0,3786	-0,6376
erro padrão	0,0683	0,0676	0,0722

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Apêndice C. Aplicação dos métodos de previsão

Tabela C.3: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em França.

Modelo final(*): SARIMA(2, 1, 0)(0, 1, 1) ₁₂ AIC = -979,38 $\hat{\sigma} = 0,0180$			
	ϕ_1	ϕ_2	η_1
estimativa	-0,7112	-0,4841	-0,5893
erro padrão	0,0644	0,0651	0,0683

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela C.4: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em Itália.

Modelo final: SARIMA(0, 1, 1)(1, 0, 1) ₁₂ AIC = -915,76 $\hat{\sigma} = 0,0225$			
	θ_1	ν_1	η_1
estimativa	-0,7051	0,9894	-0,2578
erro padrão	0,0591	0,0047	0,0794

Tabela C.5: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho nos Países Baixos.

Modelo final(*): SARIMA(2, 1, 0)(0, 1, 2) ₁₂ AIC = -920,53 $\hat{\sigma} = 0,0202$				
	ϕ_1	ϕ_2	η_1	η_2
estimativa	-0,8869	-0,6425	-0,5899	-0,2967
erro padrão	0,0585	0,0568	0,0902	0,0868

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela C.6: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho em Portugal.

Modelo final: SARIMA(2, 1, 0)(1, 0, 1) ₁₂ AIC = -933,72 $\hat{\sigma} = 0,0211$				
	ϕ_1	ϕ_2	ν_1	η_1
estimativa	-0,6810	-0,3283	0,9967	-0,5448
erro padrão	0,0676	0,0665	0,0017	0,0661

Tabela C.7: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do comércio a retalho no Reino Unido.

Modelo final: SARIMA(0, 1, 1)(0, 1, 1) ₁₂ AIC = -1104,60 $\hat{\sigma} = 0,0129$		
	θ_1	η_1
estimativa	-0,5942	-0,7162
erro padrão	0,0597	0,0628

Modelos Holt-Winters: comércio a retalho

Tabela C.8: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho na Alemanha.

Modelo aditivo					EQM = 4,3834
$\hat{\alpha} = 0,2087$	$\hat{\beta} = 0,0336$	$\hat{\gamma} = 0,3024$	$\hat{l}_1 = 87,2678$	$\hat{b}_1 = 0,1001$	
$\hat{s}_1 = -5,5146$	$\hat{s}_2 = -11,4562$	$\hat{s}_3 = 2,4354$	$\hat{s}_4 = -0,6896$	$\hat{s}_5 = 1,9646$	
$\hat{s}_6 = -4,2562$	$\hat{s}_7 = -3,9729$	$\hat{s}_8 = -3,3646$	$\hat{s}_9 = -1,9313$	$\hat{s}_{10} = 0,8187$	
$\hat{s}_{11} = 5,9562$	$\hat{s}_{12} = 20,0104$				

Tabela C.9: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em Espanha.

Modelo multiplicativo					EQM = 5,5033
$\hat{\alpha} = 0,3641$	$\hat{\beta} = 0,0749$	$\hat{\gamma} = 0,4829$	$\hat{l}_1 = 80,9233$	$\hat{b}_1 = 0,4557$	
$\hat{s}_1 = 1,0563$	$\hat{s}_2 = 0,8748$	$\hat{s}_3 = 0,9783$	$\hat{s}_4 = 0,9378$	$\hat{s}_5 = 0,9833$	
$\hat{s}_6 = 1,0165$	$\hat{s}_7 = 1,0596$	$\hat{s}_8 = 0,9205$	$\hat{s}_9 = 0,9721$	$\hat{s}_{10} = 0,9949$	
$\hat{s}_{11} = 0,9673$	$\hat{s}_{12} = 1,2387$				

Apêndice C. Aplicação dos métodos de previsão

Tabela C.10: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em França.

Modelo multiplicativo					EQM = 2,5204
$\hat{\alpha} = 0,3365$	$\hat{\beta} = 0,0329$	$\hat{\gamma} = 0,5433$	$\hat{l}_1 = 63,4961$	$\hat{b}_1 = 0,4436$	
$\hat{s}_1 = 0,9467$	$\hat{s}_2 = 0,8419$	$\hat{s}_3 = 0,9663$	$\hat{s}_4 = 0,9777$	$\hat{s}_5 = 0,9908$	
$\hat{s}_6 = 0,9955$	$\hat{s}_7 = 0,8704$	$\hat{s}_8 = 1,0695$	$\hat{s}_9 = 0,9818$	$\hat{s}_{10} = 1,0406$	
$\hat{s}_{11} = 1,0232$	$\hat{s}_{12} = 1,2955$				

Tabela C.11: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em Itália.

Modelo aditivo					EQM = 4,4141
$\hat{\alpha} = 0,1798$	$\hat{\beta} = 0,0157$	$\hat{\gamma} = 0,8463$	$\hat{l}_1 = 86,4834$	$\hat{b}_1 = 0,1541$	
$\hat{s}_1 = -10,7736$	$\hat{s}_2 = -10,8153$	$\hat{s}_3 = -3,7694$	$\hat{s}_4 = -1,9278$	$\hat{s}_5 = 2,3431$	
$\hat{s}_6 = 0,8431$	$\hat{s}_7 = -0,7403$	$\hat{s}_8 = -13,6861$	$\hat{s}_9 = -0,2986$	$\hat{s}_{10} = 4,4014$	
$\hat{s}_{11} = 3,9181$	$\hat{s}_{12} = 30,5056$				

Tabela C.12: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho nos Países Baixos.

Modelo aditivo					EQM = 6,5603
$\hat{\alpha} = 0,2168$	$\hat{\beta} = 0,2238$	$\hat{\gamma} = 0,2854$	$\hat{l}_1 = 89,6482$	$\hat{b}_1 = 0,3861$	
$\hat{s}_1 = -9,6632$	$\hat{s}_2 = -13,5924$	$\hat{s}_3 = 1,6701$	$\hat{s}_4 = -2,4174$	$\hat{s}_5 = 6,7868$	
$\hat{s}_6 = 3,6868$	$\hat{s}_7 = -1,4590$	$\hat{s}_8 = -3,2674$	$\hat{s}_9 = 1,1910$	$\hat{s}_{10} = -0,5715$	
$\hat{s}_{11} = 2,5410$	$\hat{s}_{12} = 15,0951$				

Tabela C.13: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho em Portugal.

Modelo multiplicativo				EQM = 6,1686
$\hat{\alpha} = 0,3469$	$\hat{\beta} = 0,0610$	$\hat{\gamma} = 0,6259$	$\hat{l}_1 = 100,3229$	$\hat{b}_1 = 0,5067$
$\hat{s}_1 = 0,8818$	$\hat{s}_2 = 0,8554$	$\hat{s}_3 = 0,9579$	$\hat{s}_4 = 0,9521$	$\hat{s}_5 = 0,9685$
$\hat{s}_6 = 0,9834$	$\hat{s}_7 = 1,0252$	$\hat{s}_8 = 1,0220$	$\hat{s}_9 = 0,9946$	$\hat{s}_{10} = 1,0021$
$\hat{s}_{11} = 1,0136$	$\hat{s}_{12} = 1,3434$			

Tabela C.14: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do comércio a retalho no Reino Unido.

Modelo multiplicativo				EQM = 1,4084
$\hat{\alpha} = 0,3093$	$\hat{\beta} = 0,0000$	$\hat{\gamma} = 0,4881$	$\hat{l}_1 = 64,2994$	$\hat{b}_1 = 0,2271$
$\hat{s}_1 = 0,9261$	$\hat{s}_2 = 0,9302$	$\hat{s}_3 = 0,9419$	$\hat{s}_4 = 0,9743$	$\hat{s}_5 = 0,9916$
$\hat{s}_6 = 0,9793$	$\hat{s}_7 = 0,9793$	$\hat{s}_8 = 0,9652$	$\hat{s}_9 = 0,9581$	$\hat{s}_{10} = 0,9948$
$\hat{s}_{11} = 1,0842$	$\hat{s}_{12} = 1,2750$			

Modelos de regressão linear: comércio a retalho

Tabela C.15: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho na Alemanha.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,454630	0,004063	1096,4144	<0,0001	0,007667	581,0344	<0,0001
tempo	0,000379	0,000064	5,9082	<0,0001	0,000112	3,3698	0,0008
evento(*)	-0,207794	0,013205	-15,7358	<0,0001	0,021170	-9,8154	<0,0001
cos12	0,037075	0,004191	8,8471	<0,0001	0,003071	12,0743	<0,0001
sen12	-0,019404	0,004169	-4,6542	<0,0001	0,003147	-6,1654	<0,0001
cos6	0,031445	0,004190	7,5048	<0,0001	0,002259	13,9179	<0,0001
sen6	-0,057058	0,004161	-13,7133	<0,0001	0,002243	-25,4349	<0,0001
cos4	0,054234	0,004189	12,9455	<0,0001	0,002851	19,0198	<0,0001
sen4	-0,026269	0,004159	-6,3166	<0,0001	0,002871	-9,1508	<0,0001
cos3	0,037027	0,002078	17,8146	<0,0001	0,003123	11,8573	<0,0001
sen3	0,008748	0,002077	4,2109	<0,0001	0,003123	2,8010	0,0051
cos2,4	0,035383	0,002078	17,0316	<0,0001	0,001838	19,2505	<0,0001
sen2,4	0,013051	0,002076	6,2861	<0,0001	0,001834	7,1171	<0,0001
cos2	0,007100	0,001468	4,8350	<0,0001	0,000947	7,4968	<0,0001
t :evento	0,001521	0,000102	14,9706	<0,0001	0,000175	8,6800	<0,0001
t :cos12	-0,000116	0,000035	-3,2801	0,0012	0,000026	-4,4842	<0,0001
t :sen12	-0,000003	0,000035	-0,0931	0,9260	0,000027	-0,1237	0,9015
t :cos6	-0,000012	0,000035	-0,3533	0,7243	0,000019	-0,6550	0,5125
t :sen6	0,000067	0,000035	1,8899	0,0604	0,000019	3,5036	0,0005
t :cos4	-0,000083	0,000035	-2,3458	0,0201	0,000024	-3,4451	0,0006
t :sen4	-0,000011	0,000035	-0,2997	0,7647	0,000024	-0,4362	0,6627
R^2	0,9532	F	186,37	$g.l.$	183	$\hat{\sigma}$	0,0210
R_a^2	0,9481	$\hat{\phi}_1$	-0,0758	$\hat{\phi}_2$	0,1095	$\hat{\phi}_3$	0,4837
		R_c^2	0,9640	R_{ac}^2	0,9594	$\hat{\sigma}_a$	0,0174

(*) evento ocorre em fevereiro de 2009.

Tabela C.16: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em Espanha.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,375235	0,003705	1181,0080	<0,0001	0,008091	540,7528	<0,0001
tempo	0,004421	0,000065	67,9277	<0,0001	0,000133	33,2753	<0,0001
evento1(*)	0,567825	0,020534	27,6535	<0,0001	0,034323	16,5435	<0,0001
cos12	0,027420	0,003652	7,5087	<0,0001	0,003448	7,9515	<0,0001
sen12	-0,010726	0,003629	-2,9558	0,0035	0,003529	-3,0397	0,0024
cos6	0,069378	0,003648	19,0203	<0,0001	0,002482	27,9509	<0,0001
sen6	-0,008524	0,003622	-2,3535	0,0197	0,002464	-3,4599	0,0005
cos4	0,035833	0,001811	19,7916	<0,0001	0,001338	26,7844	<0,0001
sen4	-0,004787	0,001811	-2,6440	0,0089	0,001344	-3,5623	0,0004
cos3	0,041833	0,003647	11,4693	<0,0001	0,003634	11,5109	<0,0001
sen3	0,047572	0,003620	13,1410	<0,0001	0,003606	13,1908	<0,0001
cos2,4	0,043352	0,003648	11,8841	<0,0001	0,003045	14,2378	<0,0001
sen2,4	0,014146	0,003618	3,9103	0,0001	0,003062	4,6201	<0,0001
t:evento1	-0,006491	0,000172	-37,6546	<0,0001	0,000308	-21,0864	<0,0001
t:evento2(*)	-0,003204	0,000050	-63,6159	<0,0001	0,000099	-32,4411	<0,0001
t:cos12	-0,000088	0,000031	-2,8618	0,0047	0,000029	-3,0385	0,0024
t:sen12	-0,000103	0,000031	-3,3547	0,0010	0,000030	-3,4687	0,0005
t:cos6	-0,000043	0,000031	-1,4107	0,1600	0,000021	-2,0777	0,0377
t:sen6	0,000042	0,000031	1,3704	0,1723	0,000021	2,0168	0,0437
t:cos3	-0,000069	0,000031	-2,2580	0,0251	0,000030	-2,2928	0,0219
t:sen3	-0,000046	0,000031	-1,4998	0,1354	0,000030	-1,5228	0,1278
t:cos2,4	-0,000051	0,000031	-1,6433	0,1021	0,000026	-1,9746	0,0483
t:sen2,4	-0,000002	0,000031	-0,0692	0,9449	0,000026	-0,0826	0,9342
R^2	0,9810	F	424,21	$g.l.$	181	$\hat{\sigma}$	0,0183
R_a^2	0,9787	$\hat{\phi}_1$	0,1412	$\hat{\phi}_2$	0,1658	$\hat{\phi}_3$	0,2588
		R_c^2	0,9841	R_{ac}^2	0,9819	$\hat{\sigma}_a$	0,0157

(*) evento1 ocorre em março de 2008 e evento2 em setembro de 2012.

Apêndice C. Aplicação dos métodos de previsão

Tabela C.17: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em França.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,151694	0,004340	956,6638	<0,0001	0,023045	180,1528	<0,0001
tempo	0,003168	0,000070	45,0106	<0,0001	0,000301	10,5389	<0,0001
evento(*)	0,129318	0,013251	9,7589	<0,0001	0,048495	2,6666	0,0077
cos12	0,047147	0,004414	10,6813	<0,0001	0,002936	16,0594	<0,0001
sen12	-0,048679	0,004392	-11,0847	<0,0001	0,002999	-16,2294	<0,0001
cos6	0,036908	0,004413	8,3632	<0,0001	0,002049	18,0150	<0,0001
sen6	-0,040796	0,004382	-9,3092	<0,0001	0,002028	-20,1164	<0,0001
cos4	0,078206	0,004413	17,7234	<0,0001	0,002465	31,7265	<0,0001
sen4	0,007086	0,004380	1,6176	0,1075	0,002493	2,8419	0,0045
cos3	0,046139	0,002190	21,0681	<0,0001	0,001943	23,7480	<0,0001
sen3	-0,009548	0,002190	-4,3606	<0,0001	0,001944	-4,9129	<0,0001
cos2,4	0,023569	0,004412	5,3415	<0,0001	0,002648	8,9022	<0,0001
sen2,4	0,046183	0,004379	10,5470	<0,0001	0,002639	17,4976	<0,0001
cos2	0,037174	0,001548	24,0079	<0,0001	0,000719	51,7001	<0,0001
t :evento	-0,001380	0,000106	-13,0188	<0,0001	0,000450	-3,0637	0,0022
t :cos12	-0,000089	0,000037	-2,3864	0,0180	0,000025	-3,5860	0,0003
t :sen12	0,000070	0,000037	1,8685	0,0633	0,000025	2,7411	0,0061
t :cos6	-0,000030	0,000037	-0,8049	0,4220	0,000017	-1,7369	0,0824
t :sen6	0,000130	0,000037	3,5019	0,0006	0,000017	7,5599	<0,0001
t :cos4	-0,000024	0,000037	-0,6502	0,5164	0,000021	-1,1658	0,2437
t :sen4	0,000064	0,000037	1,7260	0,0861	0,000021	3,0512	0,0023
t :cos2,4	-0,000098	0,000037	-2,6337	0,0092	0,000022	-4,3941	<0,0001
t :sen2,4	0,000031	0,000037	0,8417	0,4011	0,000022	1,4099	0,1586
R^2	0,9861	F	582,48	$g.l.$	181	$\hat{\sigma}$	0,0221
R_a^2	0,9844	$\hat{\phi}_1$	0,2070	$\hat{\phi}_2$	0,1966	$\hat{\phi}_3$	0,4422
		R_c^2	0,9930	R_{ac}^2	0,9920	$\hat{\sigma}_a$	0,0148

(*) evento ocorre em novembro de 2008.

Tabela C.18: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em Itália.

variável	estimativa	erro padrão	estatística t	valor-p
(Intercept)	4,440673	0,003597	1234,6822	<0,0001
tempo	0,001609	0,000058	27,5481	<0,0001
evento1(*)	-0,051117	0,005485	-9,3197	<0,0001
evento2(*)	0,310524	0,067624	4,5919	<0,0001
evento3(*)	0,184447	0,084231	2,1898	0,0299
cos12	0,043848	0,003730	11,7555	<0,0001
sen12	-0,015064	0,003711	-4,0597	0,0001
cos6	0,076711	0,003727	20,5839	<0,0001
sen6	-0,078404	0,003701	-21,1858	<0,0001
cos4	0,032725	0,003727	8,7807	<0,0001
sen4	-0,019706	0,003699	-5,3270	<0,0001
cos3	0,075431	0,003726	20,2442	<0,0001
sen3	0,024815	0,003698	6,7102	<0,0001
cos2,4	0,083366	0,003726	22,3756	<0,0001
sen2,4	-0,014422	0,003698	-3,8999	0,0001
cos2	0,019234	0,002625	7,3279	<0,0001
t:evento2	-0,002543	0,000420	-6,0485	<0,0001
t:evento3	-0,001740	0,000445	-3,9065	0,0001
t:cos12	-0,000129	0,000032	-4,0914	0,0001
t:sen12	-0,000137	0,000032	-4,3319	<0,0001
t:cos6	-0,000128	0,000031	-4,0774	0,0001
t:sen6	0,000256	0,000031	8,1409	<0,0001
t:cos4	0,000097	0,000031	3,0796	0,0024
t:sen4	0,000049	0,000031	1,5654	0,1193
t:cos3	-0,000167	0,000031	-5,3028	<0,0001
t:sen3	0,000081	0,000031	2,5628	0,0112
t:cos2,4	-0,000149	0,000031	-4,7459	<0,0001
t:sen2,4	0,000175	0,000031	5,5808	<0,0001
t:cos2	-0,000080	0,000022	-3,5873	0,0004
<hr/>				
$F = 275,64$	$g.l. = 175$	$\hat{\sigma} = 0,0186$	$R^2 = 0,9778$	$R_a^2 = 0,9743$

Nota: o modelo revela sazonalidade não explicada e não apresenta nenhum outro padrão autorregressivo.

(*) evento1 ocorre em outubro de 2008, evento2 em abril de 2012 e evento3 em setembro de 2014.

Apêndice C. Aplicação dos métodos de previsão

Tabela C.19: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho nos Países Baixos.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,471481	0,008603	519,7544	<0,0001	0,027147	164,7122	<0,0001
tempo	0,003656	0,000485	7,5401	<0,0001	0,001135	3,2226	0,0013
evento1(*)	0,276732	0,032380	8,5463	<0,0001	0,054841	5,0461	<0,0001
evento2(*)	0,605970	0,062035	9,7682	<0,0001	0,100194	6,0480	<0,0001
cos12	-0,005766	0,002270	-2,5397	0,0119	0,001402	-4,1127	<0,0001
sen12	-0,021404	0,002274	-9,4141	<0,0001	0,001421	-15,0635	<0,0001
cos6	0,032686	0,004568	7,1551	<0,0001	0,002026	16,1305	<0,0001
sen6	-0,058139	0,004555	-12,7650	<0,0001	0,002031	-28,6271	<0,0001
cos4	0,034306	0,002263	15,1622	<0,0001	0,001269	27,0240	<0,0001
sen4	-0,009887	0,002263	-4,3694	<0,0001	0,001271	-7,7764	<0,0001
cos3	0,041617	0,002262	18,3969	<0,0001	0,003660	11,3704	<0,0001
sen3	0,010739	0,002261	4,7503	<0,0001	0,003660	2,9339	0,0033
cos2,4	0,029807	0,002261	13,1818	<0,0001	0,001839	16,2045	<0,0001
sen2,4	0,016481	0,002260	7,2914	<0,0001	0,001832	8,9978	<0,0001
cos2	0,006022	0,001599	3,7664	0,0002	0,000898	6,7046	<0,0001
t :evento1	-0,011242	0,001127	-9,9747	<0,0001	0,002030	-5,5374	<0,0001
t^2 :evento1	0,000071	0,000008	9,1196	<0,0001	0,000011	6,5447	<0,0001
t :evento2	-0,009570	0,000962	-9,9519	<0,0001	0,001743	-5,4899	<0,0001
t^2 :evento2	0,000018	0,000003	6,6512	<0,0001	0,000004	4,2076	<0,0001
t :cos6	0,000070	0,000038	1,8185	0,0706	0,000017	4,1019	<0,0001
t :sen6	0,000065	0,000039	1,6848	0,0937	0,000017	3,7892	0,0002
R^2	0,9335	F	128,54	$g.l.$	183	$\hat{\sigma}$	0,0228
R_a^2	0,9263	$\hat{\phi}_1$	-0,1141	$\hat{\phi}_2$	0,1422	$\hat{\phi}_3$	0,6031
		R_c^2	0,9583	R_{ac}^2	0,9530	$\hat{\sigma}_a$	0,0171

(*) evento1 ocorre em julho de 2002 e evento2 em abril de 2008.

Tabela C.20: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,607472	0,004242	1086,0284	<0,0001	0,012937	356,1515	<0,0001
tempo	0,001739	0,000076	23,0001	<0,0001	0,000204	8,5454	<0,0001
evento1(*)	0,502848	0,015672	32,0849	<0,0001	0,036470	13,7878	<0,0001
evento2(*)	-0,325701	0,011359	-28,6741	<0,0001	0,028562	-11,4031	<0,0001
cos12	0,042203	0,004159	10,1471	<0,0001	0,003478	12,1345	<0,0001
sen12	-0,053926	0,004141	-13,0232	<0,0001	0,003572	-15,0984	<0,0001
cos6	0,056111	0,004159	13,4909	<0,0001	0,002390	23,4814	<0,0001
sen6	-0,029016	0,004132	-7,0230	<0,0001	0,002373	-12,2271	<0,0001
cos4	0,072531	0,004159	17,4391	<0,0001	0,002585	28,0625	<0,0001
sen4	-0,028205	0,004128	-6,8326	<0,0001	0,002618	-10,7746	<0,0001
cos3	0,063109	0,004158	15,1764	<0,0001	0,003598	17,5389	<0,0001
sen3	0,002070	0,004127	0,5014	0,6167	0,003568	0,5800	0,5619
cos2,4	0,061775	0,004159	14,8542	<0,0001	0,003285	18,8044	<0,0001
sen2,4	0,013402	0,004127	3,2478	0,0014	0,003310	4,0487	0,0001
cos2	0,015813	0,001460	10,8345	<0,0001	0,000976	16,1958	<0,0001
t :evento1	-0,004794	0,000134	-35,7197	<0,0001	0,000355	-13,5107	<0,0001
t :cos12	-0,000210	0,000035	-5,9938	<0,0001	0,000029	-7,1494	<0,0001
t :sen12	-0,000077	0,000035	-2,1870	0,0300	0,000030	-2,5452	0,0109
t :cos6	-0,000041	0,000035	-1,1619	0,2468	0,000020	-2,0221	0,0432
t :sen6	0,000158	0,000035	4,4889	<0,0001	0,000020	7,7950	<0,0001
t :cos4	0,000009	0,000035	0,2472	0,8050	0,000022	0,3979	0,6907
t :sen4	0,000047	0,000035	1,3336	0,1840	0,000022	2,1139	0,0345
t :cos3	-0,000128	0,000035	-3,6559	0,0003	0,000030	-4,2705	<0,0001
t :sen3	0,000031	0,000035	0,8975	0,3707	0,000030	1,0489	0,2942
t :cos2,4	-0,000067	0,000035	-1,8991	0,0592	0,000028	-2,4136	0,0158
t :sen2,4	0,000042	0,000035	1,2100	0,2279	0,000028	1,5249	0,1273
R^2	0,9729	F	255,98	$g.l.$	178	$\hat{\sigma}$	0,0208
R_a^2	0,9691	$\hat{\phi}_1$	0,1579	$\hat{\phi}_2$	0,2704	$\hat{\phi}_3$	0,2863
		R_c^2	0,9812	R_{ac}^2	0,9781	$\hat{\sigma}_a$	0,0162

(*) evento1 ocorre em janeiro de 2008 e evento2 em abril de 2014.

Apêndice C. Aplicação dos métodos de previsão

Tabela C.21: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do comércio a retalho no Reino Unido.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,151591	0,002391	1736,0256	<0,0001	0,005560	746,7435	<0,0001
tempo	0,002865	0,000039	73,2019	<0,0001	0,000085	33,7779	<0,0001
evento(*)	0,055154	0,007155	7,7082	<0,0001	0,014393	3,8320	0,0001
cos12	0,045536	0,002421	18,8058	<0,0001	0,002485	18,3262	<0,0001
sen12	-0,020295	0,002408	-8,4273	<0,0001	0,002546	-7,9699	<0,0001
cos6	0,056706	0,002421	23,4266	<0,0001	0,001741	32,5648	<0,0001
sen6	-0,034555	0,002404	-14,3752	<0,0001	0,001727	-20,0050	<0,0001
cos4	0,047060	0,002421	19,4411	<0,0001	0,001746	26,9547	<0,0001
sen4	-0,024051	0,002403	-10,0092	<0,0001	0,001769	-13,5933	<0,0001
cos3	0,032552	0,002421	13,4483	<0,0001	0,002000	16,2733	<0,0001
sen3	-0,007287	0,002402	-3,0333	0,0028	0,001985	-3,6707	0,0002
cos2,4	0,035936	0,001201	29,9126	<0,0001	0,000886	40,5486	<0,0001
sen2,4	-0,008016	0,001201	-6,6743	<0,0001	0,000889	-9,0181	<0,0001
cos2	0,018372	0,001705	10,7764	<0,0001	0,001174	15,6503	<0,0001
t:evento	-0,000722	0,000058	-12,4401	<0,0001	0,000125	-5,7702	<0,0001
t:cos12	-0,000037	0,000020	-1,8167	0,0709	0,000021	-1,7779	0,0754
t:sen12	-0,000069	0,000020	-3,3945	0,0008	0,000021	-3,2318	0,0012
t:cos6	0,000005	0,000020	0,2321	0,8167	0,000015	0,3237	0,7461
t:sen6	-0,000035	0,000020	-1,7204	0,0871	0,000015	-2,4009	0,0164
t:cos4	-0,000045	0,000020	-2,2128	0,0282	0,000015	-3,0756	0,0021
t:sen4	-0,000054	0,000020	-2,6701	0,0083	0,000015	-3,6527	0,0003
t:cos3	0,000021	0,000020	1,0306	0,3041	0,000017	1,2575	0,2086
t:sen3	-0,000036	0,000020	-1,7638	0,0795	0,000017	-2,1539	0,0312
t:cos2	-0,000028	0,000014	-1,9534	0,0523	0,000010	-2,8512	0,0044
R^2	0,9952	F	1638,95	$g.l.$	180	$\hat{\sigma}$	0,0121
R_a^2	0,9946	$\hat{\phi}_1$	0,2546	$\hat{\phi}_2$	0,1787	$\hat{\phi}_3$	0,1732
		R_c^2	0,9962	R_{ac}^2	0,9957	$\hat{\sigma}_a$	0,0101

(*) evento ocorre em outubro de 2008.

Modelos SARIMA: *e-commerce*

Tabela C.22: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* na Alemanha.

Modelo final ^(*) : SARIMA(2, 1, 0)(1, 0, 1) ₁₂ AIC = -534,41 $\hat{\sigma} = 0,0593$				
	ϕ_1	ϕ_2	ν_1	η_1
estimativa	-0,8498	-0,5279	0,9778	-0,4668
erro padrão	0,0611	0,0609	0,0101	0,0679

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela C.23: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* em Espanha.

Modelo final ^(*) : SARIMA(2, 1, 0)(0, 1, 1) ₁₂ AIC = -369,91 $\hat{\sigma} = 0,0882$			
	ϕ_1	ϕ_2	η_1
estimativa	-0,4071	-0,3521	-0,6698
erro padrão	0,0688	0,0683	0,0736

Tabela C.24: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* em França.

Modelo final ^(*) : SARIMA(2, 1, 0)(1, 0, 2) ₁₂ AIC = -659,12 $\hat{\sigma} = 0,0423$					
	ϕ_1	ϕ_2	ν_1	η_1	η_2
estimativa	-0,6361	-0,3927	0,9895	-0,1663	-0,2344
erro padrão	0,0672	0,0665	0,0053	0,0784	0,0837

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela C.25: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* em Itália.

Modelo final ^(*) : SARIMA(1, 1, 1)(1, 0, 0) ₁₂ AIC = -329,02 $\hat{\sigma} = 0,0991$			
	ϕ_1	θ_1	ν_1
estimativa	0,2478	-0,9879	0,9382
erro padrão	0,0692	0,0082	0,0202

(*) Rejeita independência (teste de Ljung-Box) e normalidade (teste de Kolmogorov-Smirnov) dos erros ($\alpha = 5\%$).

Tabela C.26: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* nos Países Baixos.

Modelo final ^(*) : SARIMA(0, 1, 1)(1, 0, 0) ₁₂			AIC = -390,25	$\hat{\sigma} = 0,0587$
	θ_1	ν_1		
estimativa	-0,6575	0,7105		
erro padrão	0,0716	0,0658		

Tabela C.27: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* em Portugal.

Modelo final ^(*) : SARIMA(0, 1, 2)(2, 0, 2) ₆						AIC = -340,27	$\hat{\sigma} = 0,0976$
	θ_1	θ_2	ν_1	ν_2	η_1	η_2	
estimativa	-0,5826	-0,1992	-0,0881	0,8879	0,4172	-0,3389	
erro padrão	0,0758	0,0799	0,0383	0,0391	0,0948	0,0966	

(*) Rejeita independência dos erros pelo teste de Ljung-Box ($\alpha = 5\%$).

Tabela C.28: Resultados do ajustamento do modelo SARIMA à série dos logaritmos do *e-commerce* no Reino Unido.

Modelo final ^(*) : SARIMA(0, 1, 1)(1, 0, 1) ₁₂				AIC = -628,82	$\hat{\sigma} = 0,0468$
	θ_1	ν_1	η_1		
estimativa	-0,5564	0,9840	-0,4695		
erro padrão	0,0622	0,0075	0,0753		

Modelos Holt-Winters: *e-commerce*

Tabela C.29: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* na Alemanha.

Modelo multiplicativo					EQM = 20,9007
$\hat{\alpha} = 0,3605$	$\hat{\beta} = 0,0437$	$\hat{\gamma} = 0,6417$	$\hat{l}_1 = 67,6832$	$\hat{b}_1 = -0,1188$	
$\hat{s}_1 = 1,0843$	$\hat{s}_2 = 0,8946$	$\hat{s}_3 = 1,1433$	$\hat{s}_4 = 0,9629$	$\hat{s}_5 = 0,9744$	
$\hat{s}_6 = 0,7838$	$\hat{s}_7 = 0,8427$	$\hat{s}_8 = 0,8743$	$\hat{s}_9 = 1,0853$	$\hat{s}_{10} = 1,1243$	
$\hat{s}_{11} = 1,2552$	$\hat{s}_{12} = 0,9749$				

Tabela C.30: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* em Espanha.

Modelo multiplicativo				EQM = 77,9908
$\hat{\alpha} = 0,5371$	$\hat{\beta} = 0,0343$	$\hat{\gamma} = 0,9737$	$\hat{l}_1 = 73,5414$	$\hat{b}_1 = 0,5700$
$\hat{s}_1 = 1,1446$	$\hat{s}_2 = 1,3344$	$\hat{s}_3 = 1,0345$	$\hat{s}_4 = 0,8284$	$\hat{s}_5 = 1,0050$
$\hat{s}_6 = 0,8891$	$\hat{s}_7 = 0,8735$	$\hat{s}_8 = 0,6135$	$\hat{s}_9 = 1,0112$	$\hat{s}_{10} = 1,1224$
$\hat{s}_{11} = 1,0119$	$\hat{s}_{12} = 1,1316$			

Tabela C.31: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* em França.

Modelo multiplicativo				EQM = 13,1270
$\hat{\alpha} = 0,3571$	$\hat{\beta} = 0,0145$	$\hat{\gamma} = 0,9728$	$\hat{l}_1 = 58,4730$	$\hat{b}_1 = 0,0549$
$\hat{s}_1 = 0,9960$	$\hat{s}_2 = 0,9790$	$\hat{s}_3 = 0,9514$	$\hat{s}_4 = 0,9625$	$\hat{s}_5 = 0,9065$
$\hat{s}_6 = 1,0897$	$\hat{s}_7 = 0,8650$	$\hat{s}_8 = 0,7305$	$\hat{s}_9 = 0,8373$	$\hat{s}_{10} = 1,1790$
$\hat{s}_{11} = 1,1942$	$\hat{s}_{12} = 1,3089$			

Tabela C.32: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* em Itália.

Modelo aditivo				EQM = 82,6041
$\hat{\alpha} = 0,0000$	$\hat{\beta} = 0,0000$	$\hat{\gamma} = 1,0000$	$\hat{l}_1 = 76,5442$	$\hat{b}_1 = 0,0867$
$\hat{s}_1 = -18,5049$	$\hat{s}_2 = -27,5132$	$\hat{s}_3 = 9,1701$	$\hat{s}_4 = 36,5535$	$\hat{s}_5 = 35,0576$
$\hat{s}_6 = 26,1910$	$\hat{s}_7 = -0,2507$	$\hat{s}_8 = -37,0507$	$\hat{s}_9 = -29,0090$	$\hat{s}_{10} = -1,3715$
$\hat{s}_{11} = -16,1007$	$\hat{s}_{12} = 22,8285$			

Apêndice C. Aplicação dos métodos de previsão

Tabela C.33: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* nos Países Baixos.

Modelo multiplicativo					EQM = 12,5451
$\hat{\alpha} = 0,2346$	$\hat{\beta} = 0,0729$	$\hat{\gamma} = 0,7062$	$\hat{l}_1 = 44,7318$	$\hat{b}_1 = 0,2373$	
$\hat{s}_1 = 1,0070$	$\hat{s}_2 = 0,9632$	$\hat{s}_3 = 1,0820$	$\hat{s}_4 = 0,9663$	$\hat{s}_5 = 1,0488$	
$\hat{s}_6 = 1,0127$	$\hat{s}_7 = 0,7968$	$\hat{s}_8 = 0,9282$	$\hat{s}_9 = 1,1124$	$\hat{s}_{10} = 1,1009$	
$\hat{s}_{11} = 1,0023$	$\hat{s}_{12} = 0,9793$				

Tabela C.34: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* em Portugal.

Modelo multiplicativo					EQM = 420,3237
$\hat{\alpha} = 0,1823$	$\hat{\beta} = 0,0589$	$\hat{\gamma} = 0,6781$	$\hat{l}_1 = 203,2920$	$\hat{b}_1 = 2,2286$	
$\hat{s}_1 = 1,1915$	$\hat{s}_2 = 1,0632$	$\hat{s}_3 = 0,9320$	$\hat{s}_4 = 1,0118$	$\hat{s}_5 = 1,1139$	
$\hat{s}_6 = 0,9367$	$\hat{s}_7 = 0,8972$	$\hat{s}_8 = 0,7744$	$\hat{s}_9 = 0,8494$	$\hat{s}_{10} = 1,0460$	
$\hat{s}_{11} = 1,1568$	$\hat{s}_{12} = 1,0271$				

Tabela C.35: Estimativas iniciais para o nível, o declive e os fatores sazonais e estimativas das constantes de alisamento, resultantes da aplicação do método de Holt-Winters à série do *e-commerce* no Reino Unido.

Modelo multiplicativo					EQM = 10,0033
$\hat{\alpha} = 0,4942$	$\hat{\beta} = 0,0368$	$\hat{\gamma} = 1,0000$	$\hat{l}_1 = 34,6895$	$\hat{b}_1 = -0,0303$	
$\hat{s}_1 = 0,8261$	$\hat{s}_2 = 0,8977$	$\hat{s}_3 = 0,9335$	$\hat{s}_4 = 0,8804$	$\hat{s}_5 = 0,9299$	
$\hat{s}_6 = 0,9369$	$\hat{s}_7 = 0,8909$	$\hat{s}_8 = 0,9372$	$\hat{s}_9 = 1,0448$	$\hat{s}_{10} = 1,1467$	
$\hat{s}_{11} = 1,3004$	$\hat{s}_{12} = 1,2755$				

Modelos de regressão linear: *e-commerce*

Tabela C.36: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* na Alemanha.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,219740	0,008490	497,0445	<0,0001	0,012794	329,8276	<0,0001
tempo	-0,001065	0,000104	-10,2667	<0,0001	0,000150	-7,1084	<0,0001
evento1(*)	-0,992202	0,101265	-9,7981	<0,0001	0,117553	-8,4404	<0,0001
evento2(*)	-0,918353	0,387223	-2,3716	0,0188	0,387428	-2,3704	0,0178
cos12	0,130748	0,010020	13,0491	<0,0001	0,006957	18,7926	<0,0001
sen12	-0,015157	0,009952	-1,5230	0,1295	0,007117	-2,1297	0,0332
cos6	-0,058663	0,009990	-5,8723	<0,0001	0,005187	-11,3097	<0,0001
sen6	-0,109088	0,009919	-10,9974	<0,0001	0,005142	-21,2170	<0,0001
cos4	-0,016531	0,009987	-1,6552	0,0996	0,006495	-2,5454	0,0109
sen4	0,002157	0,009915	0,2175	0,8280	0,006535	0,3300	0,7414
cos2,4	-0,023451	0,009981	-2,3495	0,0199	0,009673	-2,4244	0,0153
sen2,4	0,019584	0,009911	1,9760	0,0497	0,009613	2,0372	0,0416
cos2	-0,049326	0,007035	-7,0112	<0,0001	0,004853	-10,1648	<0,0001
t:evento1	0,007512	0,000626	11,9995	<0,0001	0,000743	10,1044	<0,0001
t:evento2	0,008283	0,001993	4,1560	0,0001	0,001998	4,1459	<0,0001
t:cos12	-0,000224	0,000085	-2,6335	0,0092	0,000060	-3,7644	0,0002
t:sen12	0,000050	0,000085	0,5887	0,5568	0,000061	0,8251	0,4093
t:cos6	0,000475	0,000084	5,6345	<0,0001	0,000044	10,7876	<0,0001
t:sen6	0,000319	0,000084	3,7773	0,0002	0,000044	7,2848	<0,0001
t:cos4	0,000254	0,000084	3,0069	0,0030	0,000055	4,6176	<0,0001
t:sen4	-0,000173	0,000084	-2,0573	0,0411	0,000055	-3,1313	0,0017
t:cos3	0,000219	0,000042	5,2434	<0,0001	0,000061	3,6058	0,0003
t:sen3	0,000010	0,000042	0,2432	0,8081	0,000061	0,1671	0,8673
t:cos2,4	0,000211	0,000084	2,5062	0,0131	0,000081	2,5908	0,0096
t:sen2,4	-0,000089	0,000084	-1,0531	0,2937	0,000081	-1,0960	0,2731
t:cos2	0,000272	0,000060	4,5618	<0,0001	0,000041	6,6166	<0,0001
R^2	0,9512	F	138,85	$g.l.$	178	$\hat{\sigma}$	0,0500
R_a^2	0,9444	$\hat{\phi}_1$	-0,1582	$\hat{\phi}_2$	0,1068	$\hat{\phi}_3$	0,4737
		R_c^2	0,9630	R_{ac}^2	0,9571	$\hat{\sigma}_a$	0,0407

(*) evento1 ocorre em outubro de 2011 e evento2 em maio de 2015.

Apêndice C. Aplicação dos métodos de previsão

Tabela C.37: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* em Espanha.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,251555	0,018753	226,7193	<0,0001	0,051937	81,8601	<0,0001
tempo	0,004758	0,000326	14,6116	<0,0001	0,000828	5,7464	<0,0001
evento(*)	5,294297	0,249386	21,2293	<0,0001	0,499464	10,6000	<0,0001
cos12	0,095381	0,009180	10,3905	<0,0001	0,009143	10,4320	<0,0001
sen12	0,044438	0,009169	4,8467	<0,0001	0,009181	4,8404	<0,0001
cos6	0,012441	0,018459	0,6740	0,5012	0,013277	0,9371	0,3487
sen6	-0,073322	0,018329	-4,0003	0,0001	0,013172	-5,5667	<0,0001
cos4	-0,095366	0,018458	-5,1667	<0,0001	0,014228	-6,7026	<0,0001
sen4	0,012183	0,018326	0,6648	0,5070	0,014403	0,8458	0,3976
cos3	0,059167	0,009161	6,4584	<0,0001	0,006778	8,7292	<0,0001
sen3	0,021339	0,009159	2,3298	0,0209	0,006774	3,1501	0,0016
cos2,4	0,031477	0,018459	1,7052	0,0899	0,009911	3,1761	0,0015
sen2,4	0,005429	0,018317	0,2964	0,7673	0,009886	0,5491	0,5829
cos2	-0,019839	0,006477	-3,0632	0,0025	0,002956	-6,7122	<0,0001
t :evento	-0,077206	0,003377	-22,8621	<0,0001	0,006913	-11,1678	<0,0001
t^2 :evento	0,000246	0,000011	22,3008	<0,0001	0,000022	11,1195	<0,0001
t :cos6	0,000383	0,000156	2,4610	0,0148	0,000112	3,4347	0,0006
t :sen6	-0,000081	0,000156	-0,5228	0,6017	0,000112	-0,7303	0,4652
t :cos4	0,000445	0,000156	2,8579	0,0048	0,000119	3,7274	0,0002
t :sen4	-0,000211	0,000156	-1,3573	0,1764	0,000121	-1,7455	0,0809
t :cos2,4	0,000104	0,000156	0,6653	0,5067	0,000084	1,2390	0,2153
t :sen2,4	-0,000209	0,000156	-1,3446	0,1804	0,000083	-2,5131	0,0120
R^2	0,8672	F	56,62	$g.l.$	182	$\hat{\sigma}$	0,0924
R_a^2	0,8519	$\hat{\phi}_1$	0,4431	$\hat{\phi}_2$	-0,0024	$\hat{\phi}_3$	0,2494
		R_c^2	0,9124	R_{ac}^2	0,9007	$\hat{\sigma}_a$	0,0709

(*) evento ocorre em abril de 2008.

Tabela C.38: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* em França.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	4,016446	0,005298	758,1463	<0,0001	0,008271	485,5942	<0,0001
tempo	0,002610	0,000049	53,3737	<0,0001	0,000074	35,1045	<0,0001
cos12	0,111742	0,007194	15,5337	<0,0001	0,006285	17,7784	<0,0001
sen12	0,039760	0,007157	5,5551	<0,0001	0,006441	6,1725	<0,0001
cos6	0,093368	0,007186	12,9938	<0,0001	0,004814	19,3938	<0,0001
sen6	-0,120740	0,007136	-16,9210	<0,0001	0,004775	-25,2851	<0,0001
cos4	-0,068899	0,007188	-9,5856	<0,0001	0,005724	-12,0374	<0,0001
sen4	-0,036465	0,007133	-5,1122	<0,0001	0,005786	-6,3026	<0,0001
cos3	0,060686	0,007186	8,4452	<0,0001	0,008998	6,7444	<0,0001
sen3	0,058470	0,007133	8,1968	<0,0001	0,008925	6,5514	<0,0001
cos2,4	0,017660	0,003567	4,9511	<0,0001	0,003258	5,4204	<0,0001
sen2,4	0,009562	0,003565	2,6818	0,0080	0,003260	2,9333	0,0034
cos2	0,041266	0,005062	8,1521	<0,0001	0,003791	10,8844	<0,0001
t :evento(*)	0,000979	0,000054	18,1364	<0,0001	0,000071	13,7329	<0,0001
t :cos12	0,000168	0,000061	2,7664	0,0063	0,000053	3,1656	0,0015
t :sen12	-0,000141	0,000061	-2,3039	0,0224	0,000055	-2,5750	0,0100
t :cos6	-0,000010	0,000061	-0,1719	0,8637	0,000041	-0,2566	0,7975
t :sen6	0,000341	0,000061	5,6340	<0,0001	0,000041	8,4282	<0,0001
t :cos4	0,000484	0,000061	7,9771	<0,0001	0,000048	10,0336	<0,0001
t :sen4	0,000186	0,000061	3,0702	0,0025	0,000049	3,8097	0,0001
t :cos3	-0,000020	0,000061	-0,3246	0,7458	0,000075	-0,2633	0,7923
t :sen3	-0,000249	0,000061	-4,1079	0,0001	0,000075	-3,3326	0,0009
t :cos2	-0,000194	0,000043	-4,5388	<0,0001	0,000032	-6,0759	<0,0001
R^2	0,9801	F	405,06	$g.l.$	181	$\hat{\sigma}$	0,0360
R_a^2	0,9777	$\hat{\phi}_1$	-0,0013	$\hat{\phi}_2$	0,0935	$\hat{\phi}_3$	0,3124
		R_c^2	0,9823	R_{ac}^2	0,9798	$\hat{\sigma}_a$	0,0320

Nota: o modelo revela sazonalidade não explicada.

(*) evento ocorre em agosto de 2015.

Apêndice C. Aplicação dos métodos de previsão

Tabela C.39: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* em Itália.

variável	estimativa	erro padrão	estatística t	valor-p
(Intercept)	4,260244	0,016610	256,4792	<0,0001
tempo	0,001606	0,000141	11,4314	<0,0001
cos12	-0,152346	0,023576	-6,4620	<0,0001
sen12	0,269639	0,023405	11,5206	<0,0001
cos6	0,219011	0,023572	9,2911	<0,0001
sen6	-0,303095	0,023403	-12,9511	<0,0001
cos4	-0,040870	0,011700	-3,4932	0,0006
sen4	-0,037878	0,011700	-3,2375	0,0014
cos3	0,074089	0,023571	3,1433	0,0019
sen3	0,152269	0,023394	6,5088	<0,0001
cos2,4	0,062182	0,011699	5,3150	<0,0001
sen2,4	-0,044554	0,011687	-3,8123	0,0002
t :cos12	0,001510	0,000199	7,5997	<0,0001
t :sen12	-0,002099	0,000199	-10,5594	<0,0001
t :cos6	-0,000899	0,000199	-4,5243	<0,0001
t :sen6	0,001165	0,000199	5,8617	<0,0001
t :cos3	0,000278	0,000199	1,3966	0,1642
t :sen3	-0,001140	0,000199	-5,7357	<0,0001
<hr/>				
$F = 55,49$	$g.l. = 186$	$\hat{\sigma} = 0,1180$	$R^2 = 0,8353$	$R_a^2 = 0,8202$

Nota: o modelo revela sazonalidade não explicada e não apresenta nenhum outro padrão autorregressivo.

Tabela C.40: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* nos Países Baixos.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	3,453011	0,024878	138,8005	<0,0001	0,031295	110,3371	<0,0001
tempo	0,005367	0,000240	22,3852	<0,0001	0,000299	17,9369	<0,0001
evento(*)	-0,765359	0,065889	-11,6159	<0,0001	0,080272	-9,5346	<0,0001
cos12	0,019887	0,019873	1,0007	0,3189	0,023441	0,8484	0,3962
sen12	0,052102	0,019844	2,6256	0,0097	0,023101	2,2554	0,0241
cos6	-0,086457	0,019869	-4,3514	<0,0001	0,020056	-4,3109	<0,0001
sen6	-0,024281	0,019757	-1,2290	0,2214	0,020350	-1,1932	0,2328
cos4	-0,065732	0,019862	-3,3094	0,0012	0,017324	-3,7942	0,0001
sen4	0,006916	0,019729	0,3505	0,7265	0,017653	0,3917	0,6952
t :evento	0,004949	0,000425	11,6365	<0,0001	0,000522	9,4746	<0,0001
t :cos12	0,000142	0,000143	0,9983	0,3200	0,000167	0,8519	0,3942
t :sen12	-0,000415	0,000144	-2,8900	0,0045	0,000167	-2,4878	0,0129
t :cos6	0,000842	0,000143	5,9036	<0,0001	0,000144	5,8578	<0,0001
t :sen6	-0,000028	0,000143	-0,1989	0,8426	0,000146	-0,1942	0,8461
t :cos4	0,000553	0,000143	3,8766	0,0002	0,000124	4,4408	<0,0001
t :sen4	-0,000112	0,000143	-0,7872	0,4326	0,000127	-0,8847	0,3763
t :cos3	0,000175	0,000043	4,1123	0,0001	0,000033	5,2542	<0,0001
t :sen3	0,000058	0,000043	1,3559	0,1775	0,000033	1,7366	0,0825
R^2	0,9747	F	285,33	$g.l.$	126	$\hat{\sigma}$	0,0501
R_a^2	0,9713	$\hat{\phi}_1$	0,2652	$\hat{\phi}_2$	NA	$\hat{\phi}_3$	NA
		R_c^2	0,9765	R_{ac}^2	0,9731	$\hat{\sigma}_a$	0,0452

(*) evento ocorre a outubro de 2011.

Apêndice C. Aplicação dos métodos de previsão

Tabela C.41: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* em Portugal.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	5,307259	0,022421	236,7095	<0,0001	0,028061	189,1352	<0,0001
tempo	0,003505	0,000659	5,3202	<0,0001	0,000808	4,3374	<0,0001
evento1(*)	0,180195	0,041852	4,3056	<0,0001	0,051202	3,5193	0,0004
evento2(*)	-1,368550	0,102934	-13,2954	<0,0001	0,125539	-10,9014	<0,0001
cos12	0,146772	0,019316	7,5983	<0,0001	0,022691	6,4682	<0,0001
sen12	0,035706	0,019244	1,8554	0,0651	0,022406	1,5936	0,1110
cos6	0,033539	0,009588	3,4982	0,0006	0,009868	3,3987	0,0007
sen6	-0,132439	0,009590	-13,8095	<0,0001	0,009889	-13,3926	<0,0001
cos4	-0,040568	0,009576	-4,2366	<0,0001	0,008673	-4,6773	<0,0001
sen4	-0,041932	0,009577	-4,3786	<0,0001	0,008716	-4,8108	<0,0001
cos3	0,034884	0,009577	3,6426	0,0003	0,007856	4,4402	<0,0001
sen3	-0,022509	0,009573	-2,3513	0,0198	0,007863	-2,8626	0,0042
cos2,4	0,022013	0,009576	2,2989	0,0226	0,007396	2,9762	0,0029
sen2,4	-0,026765	0,009572	-2,7962	0,0057	0,007369	-3,6323	0,0003
t :evento1	-0,008520	0,000732	-11,6452	<0,0001	0,000901	-9,4558	<0,0001
t :cos12	-0,000565	0,000163	-3,4673	0,0007	0,000190	-2,9733	0,0029
t :sen12	-0,000308	0,000163	-1,8851	0,0610	0,000190	-1,6208	0,1051
R^2	0,9304	F	156,14	$g.l.$	187	$\hat{\sigma}$	0,0967
R_a^2	0,9244	$\hat{\phi}_1$	0,2356	$\hat{\phi}_2$	NA	$\hat{\phi}_3$	NA
		R_c^2	0,9343	R_{ac}^2	0,9282	$\hat{\sigma}_a$	0,0899

(*) evento1 ocorre em julho de 2004 e evento2 em fevereiro de 2013.

Nota: o modelo revela sazonalidade não explicada.

Tabela C.42: Resultados do ajustamento do modelo de regressão linear múltipla à série dos logaritmos do *e-commerce* no Reino Unido.

variável	estimativa	erro padrão	estatística t	valor-p	corrigido para autocorrelação		
					erro padrão	estatística z	valor-p
(Intercept)	3,547588	0,010359	342,4740	<0,0001	0,025043	141,6595	<0,0001
tempo	-0,002845	0,000233	-12,1861	<0,0001	0,000542	-5,2469	<0,0001
t^2	0,000046	0,000001	41,3291	<0,0001	0,000002	18,3809	<0,0001
cos12	0,092624	0,009696	9,5531	<0,0001	0,010952	8,4574	<0,0001
sen12	-0,078084	0,009659	-8,0839	<0,0001	0,011141	-7,0087	<0,0001
cos6	0,004982	0,009695	0,5138	0,6080	0,006842	0,7281	0,4665
sen6	-0,080906	0,009630	-8,4011	<0,0001	0,006869	-11,7781	<0,0001
cos4	-0,017136	0,009694	-1,7677	0,0788	0,005783	-2,9633	0,0030
sen4	-0,058396	0,009623	-6,0681	<0,0001	0,005741	-10,1708	<0,0001
cos3	0,018568	0,004809	3,8610	0,0002	0,002906	6,3893	<0,0001
sen3	-0,027635	0,004806	-5,7498	<0,0001	0,002919	-9,4679	<0,0001
cos2,4	0,016200	0,004807	3,3701	0,0009	0,003404	4,7589	<0,0001
sen2,4	-0,018466	0,004804	-3,8442	0,0002	0,003410	-5,4152	<0,0001
t :cos12	0,000232	0,000082	2,8435	0,0050	0,000092	2,5267	0,0115
t :sen12	0,000342	0,000082	4,1669	<0,0001	0,000094	3,6376	0,0003
t :cos6	0,000421	0,000082	5,1549	<0,0001	0,000058	7,3043	<0,0001
t :sen6	0,000050	0,000082	0,6092	0,5431	0,000058	0,8566	0,3917
t :cos4	0,000314	0,000082	3,8413	0,0002	0,000049	6,4437	<0,0001
t :sen4	0,000011	0,000082	0,1345	0,8932	0,000049	0,2257	0,8214
R^2	0,9886	F	894,99	$g.l.$	185	$\hat{\sigma}$	0,0485
R_a^2	0,9875	$\hat{\phi}_1$	0,3225	$\hat{\phi}_2$	0,3156	$\hat{\phi}_3$	NA
		R_c^2	0,9921	R_{ac}^2	0,9912	$\hat{\sigma}_a$	0,0386

Apêndice D

Funções no ambiente R

Funções implementadas

hw_est_h: devolve as estimativas a h-passos para um dado modelo de Holt-Winters

```
hw_est_h = function(modelo, p, h){
  y = modelo$x
  n = length(y)
  alpha = modelo$alpha
  beta = modelo$beta
  gamma = modelo$gamma
  l = c()
  b = c()
  s = c()
  l[1] = modelo$fitted[,2][1] # inicialização
  b[1] = modelo$fitted[,3][1]
  s[1:p] = modelo$fitted[,4][1:p]
  if (modelo$seasonal == "multiplicative"){ ## modelo multiplicativo
    for (i in 2:(n-p)){
      l[i] = alpha * (y[p+i-1]/s[i-1]) + (1 - alpha) * (l[i-1] + b[i-1])
      b[i] = beta * (l[i] - l[i-1]) + (1 - beta) * b[i-1]
      s[i+p-1] = gamma * (y[i+p-1]/l[i]) + (1 - gamma) * s[i-1]
    }
  }
  est_h = c()
  for (j in p:(n-h)){ # estimativas a h-passos
    mod.j = j - p + 1 + (h-1)%%p
```

```

        est_h[j-11] = (l[j-11] + h * b[j-11]) * s[mod.j]
    }
}
else{  ## modelo aditivo
  for (i in 2:(n-p)){
    l[i] = alpha * (y[p+i-1] - s[i-1]) + (1 - alpha) * (l[i-1] + b[i-1])
    b[i] = beta * (l[i] - l[i-1]) + (1 - beta) * b[i-1]
    s[i+p-1] = gamma * (y[i+p-1] - l[i]) + (1 - gamma) * s[i-1]
  }
  est_h = c()
  for (j in p:(n-h)){  # estimativas a h-passos
    mod.j = j - p + 1 + (h-1)%%p
    est_h[j-11] = l[j-11] + h * b[j-11] + s[mod.j]
  }
}
return(est_h[1:(n-p-h+1)])
}

```

eqm_h: devolve o erro quadrático médio a h-passos para um dado modelo de Holt-Winters

```

eqm_h = function(modelo, h, p){
  serie = modelo$x
  erro_quad = (serie[(p+h):length(serie)] - hw_est_h(modelo, p, h))^2
  med = sum(erro_quad)/length(erro_quad)
  return(med)
}

```

int_prev_h: devolve as previsões pontuais e os intervalos de previsão para cada horizonte temporal h para um dado modelo de Holt-Winters

```

int_prev_h = function(modelo, h, p, alpha){
  dados = modelo$x
  prev_h = predict(modelo, n.ahead=h)[h]
  lim_inf = prev_h - qnorm(1 - alpha/2) * sqrt(eqm_h(modelo, h, p))
  lim_sup = prev_h + qnorm(1 - alpha/2) * sqrt(eqm_h(modelo, h, p))
  return(data.frame(h=h, inf = lim_inf, prev = prev_h, sup = lim_sup))
}

```

Lista das principais funções utilizadas

Tabela D.1: Principais funções utilizadas no ambiente R.

Função	Package	Descrição
acf	stats	Calcula a função de autocorrelação (FAC)
arima	stats	Ajusta um modelo ARIMA a uma série temporal univariada
arima.sim	stats	Simula processos ARIMA
Box.test	stats	Calcula a estatística de teste Box-Pierce ou Ljung-Box para examinar a hipótese nula de independência para uma determinada série temporal
decompose	stats	Decompõe uma série temporal nas componentes tendência, sazonalidade e aleatoriedade, através de um processo de médias móveis
eqm_h	(nenhum)	Devolve o erro quadrático médio a h passos para um dado modelo de Holt-Winters (função implementada)
fitted	stats	Extrai os valores estimados por um dado modelo
forecast	forecast	Calcula previsões e intervalos de previsão para modelos de séries temporais
HoltWinters	stats	Aplica o método de Holt-Winters a uma determinada série temporal
hw_est_h	(nenhum)	Devolve as estimativas a h passos para um dado modelo de Holt-Winters (função implementada)
int_prev_h	(nenhum)	Devolve as previsões pontuais e intervalares para cada horizonte temporal h para um dado modelo de Holt-Winters (função implementada)
ks.test	stats	Aplica o teste de Kolmogorov-Smirnov a uma ou duas amostras
lm	stats	Ajusta um modelo de regressão linear aos dados
mape	Metrics	Calcula o erro percentual absoluto médio (EPAM)
mase	Metrics	Calcula o erro escalado absoluto médio (EEAM)
mse	Metrics	Calcula o erro quadrático médio (EQM)
ndiffs	forecast	Estima o número de diferenças regulares necessárias para tornar uma determinada série estacionária
nsdiffs	forecast	Estima o número de diferenças sazonais necessárias para tornar uma determinada série estacionária
pacf	stats	Calcula a função de autocorrelação parcial (FACP)
polyroot	base	Encontra zeros de um polinómio real ou complexo
predict	stats	Calcula previsões e intervalos de previsão para um dado modelo ajustado
rmse	Metrics	Calcula a raiz do erro quadrático médio (REQM)
sarima	astsa	Realiza uma avaliação de diagnóstico a modelos ARIMA
shapiro.test	stats	Aplica o teste de normalidade de Shapiro-Wilk
sim_sarima	sarima	Simula processos SARIMA
t.test	stats	Aplica o teste t-Student
ts	stats	Cria objetos em formato time-series
ur.df	urca	Aplica o teste de Dickey-Fuller Aumentado (ADF)
ur.kpss	urca	Aplica o teste KPSS
Ustat	(nenhum)	Calcula o valor da estatística U de Theil (função implementada)