

UNIVERSIDADE ABERTA



UNIVERSIDADE
AbERTA
www.uab.pt

Modelação e Previsão da Procura Turística na Ilha do Sal – Cabo Verde:

Modelo SARIMA *versus* Rede Neuronal Artificial

Gilberto Alves Neves

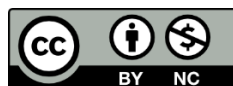
Doutoramento em Matemática Aplicada e Modelação

Orientadora (externa): Paula Odete Fernandes – Professora Coordenadora, Departamento de Ciências Empresariais e Sociais, da Escola Superior de Tecnologia e Gestão (ESTiG), do Instituto Politécnico de Bragança.

Orientadora (interna): Catarina Sofia da Costa Nunes Duarte (Catarina S. Nunes) – Professora Auxiliar, Departamento de Ciências e Tecnologia, Universidade Aberta.

Novembro de 2022

©2022. Este presente trabalho de investigação está licenciado sob uma licença CC BY-NC



AGRADECIMENTOS

Em primeiro lugar ao meu Deus e ao nosso Deus por tudo.

À Câmara Municipal do Sal pelo apoio concedido no pagamento das propinas durante todo o tempo de estudo e investigação, no quadro dos projetos de Reforço de Capacidades – Formação e Capacitação dos Recursos Humanos e Qualificação dos Jovens. A todos os docentes do programa de Doutoramento em Matemática Aplicada e Modelação, pelos conhecimentos partilhados, com profissionalismo e paciência.

Às minhas orientadoras, pela forma como me conduziram em todo o caminho usando da paciência e sabedoria para que eu permanecesse engajado e focado quando as coisas me pareciam complicadas. Sinto-me orgulhoso em as ter escolhido e grato por terem aceitado o desafio.

À Fundação para a Ciência e Tecnologia (FCT, Portugal) pelo apoio financeiro dos fundos nacionais FCT/MCTES à UNIAG, ao abrigo do Projeto n.º UIDB/04752/2020 e para o INEGI no âmbito do projeto LAETA UIDB/5022/2020.

À Organização do ICOTT'S e à equipa de revisores pela oportunidade de apresentar os dois artigos e terem sido publicados.

À Organização do OL2A e à equipa de revisores pela oportunidade de apresentar o artigo e ter sido publicado.

A todos que direta ou indiretamente me proporcionaram tempo, paciência, motivação e conhecimento para cumprir a desafiante tarefa.

E por último ao meu Deus e ao nosso Deus por tudo.

DEDICATÓRIA

Dedico este trabalho a todos os estudantes, profissionais e pais ou encarregados de educação que falam a língua portuguesa quer estejam ou sejam de Portugal Continental, dos Açores, da Madeira, na diáspora, em Cabo Verde ou nos PALOP, que sonharam um sonho possível e optaram por ingressar nesta maratona que foi o doutoramento, sacrificando horas de sono e dias de convívio com a família e amigos para que a corrida fosse chegada ao fim.



DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho.

Confirmando que eu não recorri à prática de plágio ou a qualquer forma de falsificação de resultados.

Universidade Aberta, 24 de março de 2022

Assinatura:



RESUMO

Modelação e Previsão da Procura Turística na Ilha do Sal – Cabo Verde: Modelo SARIMA versus Rede Neuronal Artificial

O objetivo principal deste trabalho é a análise da série temporal "Número de dormidas mensais nos estabelecimentos turísticos da ilha do Sal - Cabo Verde" entre 2000 e 2018 e demonstrar que, no quadro dos modelos quantitativos, estimativas e previsões fiáveis para o comportamento da procura turística são preferencialmente obtidas com modelos estatísticos de previsão adequadamente especificados e testados, e com redes neuronais artificiais que permitem prever um passo à frente. Para tal, iniciou-se a investigação com uma revisão da literatura e análise de dados que possibilitou conhecer: a dinâmica do turismo mundial; o mercado emissor europeu; o desenvolvimento turístico da ilha; as características da série temporal; modelos de previsão e tendências da procura turística. Seguidamente implementaram-se diferentes estruturas de modelos de previsão. Os resultados obtidos mostram que, a nível individual, a ilha se encontra na fase de Desenvolvimento; a nível de competitividade, a ilha está estagnada dentro da fase de Exploração; e que o seu Índice de Desenvolvimento Turístico deverá crescer em 48% para entrar na fase de Envolvimento. Quanto aos modelos de previsão, obtiveram-se: o modelo **SARIMA(2,1,0)(0,1,1)[12]**, com uma acurácia medida pelo MAPE igual a 6,77%; o modelo de redes neuronais do tipo **RNAR(12,1,7)[12]**, com um erro de 5,61% e o método de **Holt-Winters** que produziu um modelo com uma precisão de 7,94%. Todos esses modelos têm alta precisão, com destaque para a rede neuronal, apesar dos dados da série não estarem adaptados à Lei de Benford. Porém, o proposto **Algoritmo de Atribuição do Erro**, traz melhorias ao resultado do modelo SARIMA, com uma precisão de 4,98%. Esta tese pretendeu contribuir para mostrar o potencial dos modelos estatísticos de previsão e da aplicação das redes neuronais artificiais para a previsão do número de dormidas mensais na ilha do Sal. Também se avalia a precisão das previsões de cada modelo e compara os seus diferentes desempenhos.

Palavras-Chave: Turismo, Série Temporal, SARIMA, Holt-Winters, Rede Neuronal Artificial, Ilha do Sal

ABSTRACT

Modeling and Forecasting Tourist Demand in the Island of Sal – Cape Verde: SARIMA Model versus Artificial Neural Network

The main objective of this work is the analysis of the time series "Number of monthly overnight stays in tourist establishments on the Island of Sal- Cape Verde" between 2000 and 2018. In the framework of quantitative models, one can obtain reliable estimates and forecasts for the behavior of tourism demand, using statistical forecasting models adequately specified and tested, and artificial neural networks that allow one step ahead predictions. With this goal, the research began with a literature review and data analysis that allowed to understand: the dynamics of world tourism; the European outbound market; the island's tourism development; the time series characteristics; the forecasting models and tourism demand trends. Then, different forecast model structures were implemented. The results obtained show that, at the individual level, the island is in the Development stage; at the competitiveness level, the island is stagnated within the Exploration stage; and that its Tourism Development Index should grow by 48% to enter the Involvement stage. As for the prediction models, the following were obtained: the **SARIMA(2,1,0)(0,1,1)[12]** model, with an accuracy measured by MAPE equal to 6.77%; the neural network model of type **RNAR(12,1,7)[12]**, with an error of 5.61% and the **Holt Winters** method that produced a model with an accuracy of 7.94%. All models show a high accuracy, with the neural network standing out, despite the series data not being adapted to Benford's Law. However, the proposed **Error Assignment Algorithm** brings improvements to the result of the SARIMA model, with an accuracy of 4.98%. This thesis intended to contribute to show the potential of statistical forecasting models and the application of artificial neural networks to forecast the number of monthly overnight stays in the Island of Sal. The forecasting accuracy of each model was also evaluated, and their different performances were compared.

Keywords: Tourism, Time Series, SARIMA, Holt-Winters, Artificial Neural Network, Sal Island

PUBLICAÇÕES

1. Neves G.A., Nunes C.S., & Fernandes P.O. (2022). Application of Benford's Law to the Tourism Demand: The Case of the Island of Sal, Cape Verde. In: Pereira A.I. et al. (eds) Optimization, Learning Algorithms and Applications. OL2A 2021. Communications in Computer and Information Science, vol 1488 (pp 587-598). Springer, Cham. (ISBN 978-3-030-91884-2). <http://hdl.handle.net/10400.2/11589>. https://doi.org/10.1007/978-3-030-91885-9_43
2. Neves G.A., Nunes C.S., & Fernandes P.O. (2021). Seasonal Autoregressive Integrated Moving Average Time Series Model for Tourism Demand: the Case of the Island of Sal, Cape Verde. In the 3rd International Conference on Tourism Technology & Systems - ICOTTS'2021, 4 and 5 November 2021, Universidad de Cartagena, Cartagena de Indias, Colombia. https://doi.org/10.1007/978-981-16-9701-2_2
3. Neves G.A., Nunes C.S., & Fernandes P.O. (2021). Challenges and Opportunities for Island Tourist Destinations: The Case of the Island of Sal, Cape Verde. In: Abreu A., Liberato D., González E.A., Garcia Ojeda J.C. (eds) Advances in Tourism, Technology and Systems. ICOTTS 2020. Smart Innovation, Systems and Technologies, vol 209 (pp. 498-508). Springer, Singapore. (ISBN: 978-981-33-4260-6) <http://hdl.handle.net/10400.2/11413>. https://doi.org/10.1007/978-981-33-4260-6_42

ÍNDICE

Agradecimentos	iii
Dedicatória.....	iv
Resumo	vi
Publicações.....	viii
Lista de Figuras.....	xiv
Lista de Tabelas.....	xvii
Lista de Abreviaturas, Siglas e Acrónimos.....	xix
I – INTRODUÇÃO	1
I.1. Fundamentação.....	1
I.2. Objetivos e Metodologia da Tese	6
I.3. Estrutura da Tese.....	8
CAPÍTULO 1: TURISMO – O MUNDO E CABO VERDE	9
1.1 Introdução	9
1.2 Turismo – Dinâmica Mundial	9
1.3 Tendências	11
1.4 O Turismo em Cabo Verde	12
1.5 O Mercado Turístico Europeu	18
1.6 Potencialidades do Turismo Cabo-Verdiano	20
1.7 Problemas do Turismo em Cabo Verde	22
1.8 Metas e Previsões para o Turismo em Cabo Verde.....	23
1.9 Competitividade	26
1.10 Considerações Finais do Capítulo	27
CAPÍTULO 2: TURISMO – ILHA DO SAL.....	30
2.1. Introdução	30
2.2. Situação Socioeconómica da Ilha.....	31
2.3. Potencialidades Turísticas da Ilha do Sal.....	50

2.4.	Problemas do Turismo no Sal	52
2.5.	Previsão do Crescimento do Turismo	52
2.6.	Ciclo de Vida do Turismo na Ilha do Sal	54
2.7.	Índice de Desenvolvimento Turístico	57
2.8.	Benchmarking	61
2.9.	Medidas a adotar	66
2.10.	Considerações Finais do Capítulo	67
CAPÍTULO 3: MÉTODOS E MODELOS DE PREVISÃO DA PROCURA TURÍSTICA		69
3.1.	Introdução	69
3.2.	Conceito de Procura Turística	73
3.2.1.	Características e Fatores Determinantes	74
3.3.	Estado da Arte: Previsão da Procura Turística	77
3.3.1.	Métodos e modelos de previsão no período entre 1960 a 2018	77
3.3.2.	Séries Temporais	78
3.3.3.	Modelos Econométricos	82
3.3.4.	Inteligência Artificial	87
3.3.4.1.	Métodos causais baseados em IA	87
3.3.4.2.	Modelação de séries temporais baseados em IA	88
3.3.5.	Outras abordagens quantitativas	90
3.3.6.	Tendências e desafios futuros das previsões da procura turística	90
3.4.	Séries Temporais	94
3.4.1.	Definição	94
3.4.2.	História	95
3.4.3.	Importância	95
3.4.4.	Padrões	96
3.4.5.	Conceitos Gerais sobre Processos Estocásticos	96
3.4.6.	Autocorrelação	99

3.4.7. Correlograma	100
3.4.8. Sazonalidade.....	102
3.4.9. Modelos de Séries Temporais	104
3.4.9.1. Modelo SARIMA	107
3.4.10. Análise de estacionariedade	108
3.4.11. Identificação do modelo.....	109
3.4.12. Estimacão dos parâmetros.....	111
3.4.13. Avaliacão do diagnóstico.....	113
3.4.14. Escolha de modelos	118
3.4.15. Performance de previsão	119
3.5. Métodos de Suavizacão Exponencial.....	121
3.5.1. Suavizacão Exponencial Tripla	122
3.6. A Lei de Benford	125
3.7. Inteligência Artificial: Redes Neurais.....	128
3.7.1. Breve história das redes neurais	129
3.7.2. Conceito de Redes Neurais	130
3.7.3. Funcionamento.....	131
3.7.4. Vantagens e desvantagens.....	134
3.7.5. Tipo de redes neurais.....	134
3.7.6. As redes <i>Feedforward</i>	135
3.7.7. O Algoritmo Error <i>Backpropagation</i>	137
3.7.8. Redes Neurais na Previsão de Séries temporais	144
3.8. Proposta de Algoritmo de Atribuição do Erro	148
3.9. Considerações Finais do Capítulo	151
CAPÍTULO 4: RESULTADOS	154
4.1. Introduçãõ	154
4.2. Resultados	155

4.2.1. Ciclo de Vida e o Índice de Desenvolvimento Turístico	155
4.2.1.1. Índice de Desenvolvimento Turístico aplicado a Cabo Verde	158
4.2.1.2. Índice de Desenvolvimento Turístico aplicado à ilha do Sal	159
4.2.1.3. Considerações Finais	160
4.2.2. Análise da Série Temporal	161
4.2.3. Aplicação da Lei de Benford	164
4.2.3.1. Considerações Finais	167
4.2.4. Aplicação do Modelo SARIMA	169
4.2.4.1. Modelação da série	169
4.2.4.2. Ajustamento	188
4.2.4.3. Previsão	190
4.2.4.4. Considerações Finais	195
4.2.5. Aplicação do Método de Holt-Winters (Suavização Exponencial Tripla)	196
4.2.5.1. Modelo m5: Holt Winters	197
4.2.5.2. Considerações Finais	206
4.2.6. Aplicação da Rede Neuronal Artificial	206
4.2.6.1. Modelação com a função mlp()	207
4.2.6.2. Modelação com a função nnetar ()	211
4.2.6.3. Considerações Finais	216
4.2.7. Aplicação do Algoritmo de Atribuição do Erro	216
4.2.7.1. Considerações Finais	218
4.3. Comparação entre os Modelos	219
4.3.1. SARIMA <i>versus</i> Holt-Winters	220
4.3.2. SARIMA <i>versus</i> Redes Neurais	220
4.3.3. SARIMA <i>versus</i> Algoritmo de Atribuição do Erro	221
4.3.4. Comparação dos modelos estatísticos e rede neuronal com a previsão feita com a Análise Lógica Intuitiva	223

4.3.5. Considerações Finais	226
CAPÍTULO 5: CONCLUSÕES	228
5.1. Considerações finais	228
5.2. Limitações da Investigação	231
5.3. Recomendações e Novas Linhas de Pesquisa	232
5.3.1. Recomendações	232
5.3.2. Novas Linhas de Pesquisa	233
REFERÊNCIAS BIBLIOGRÁFICAS	235
Anexo I: Cenários do crescimento do fluxo turístico	250
Anexo II: Vantagens e desvantagens da aplicação de diferentes modelos de previsão da procura turística	252
Anexo III: Número de registos mensais de dormidas nos estabelecimentos turísticos na ilha do Sal no período [2000:01-2018:12]	260
Anexo IV: Resultados da aplicação do comando <i>auto.arima()</i>	262
Anexo V: Scripts	267
Anexo VI: Resultados da implementação da função <i>nnetar()</i>	273
Anexo VII: Resultados das análises dos modelos 1, 2, 3 e 4	292
Anexo VIII: Resultados da implementação da função <i>mlp()</i>	301
Anexo IX: Publicações	335

LISTA DE FIGURAS

Figura I.1: Carta de localização de Cabo Verde, ilha do Sal e o mercado europeu com os maiores países emissores de turistas para o destino ilha do Sal.	5
Figura I.2: Estrutura da Tese	8
Figura 1.1. Previsão da procura turística em diferentes destinos do globo.	11
Figura 1.2: Evolução do número de hóspedes em Cabo Verde (2000:2018).....	15
Figura 1.3 Evolução do número de dormidas em Cabo Verde (2000:2018).	15
Figura 1.4: Hóspedes e Dormidas (%) por mercado emissor dos hóspedes, 2018.	17
Figura 1.5: Estadia média (noites) por mercado emissor dos hóspedes, 2018.	17
Figura 2.1: Evolução do número da população da ilha do Sal 1970:2019 com projeção de 2020 e 2030.	32
Figura 2.2: Saldo migratório entre diferentes ilhas nos últimos 5 anos: 2008 a 2013.	36
Figura 2.3: Repartição dos emigrantes comparando Sal com outros municípios no período 2009:2014.	36
Figura 2.4: Evolução da percentagem de Agregados Familiares com ligação a rede pública de água.	37
Figura 2.5: Evolução da percentagem dos Agregados Familiares com acesso ao sistema de evacuação de águas residuais (rede de esgoto/fossa séptica).	38
Figura 2.6: Evolução da percentagem de Acesso à Eletricidade dos Agregados Familiares.....	40
Figura 2.7: Evolução da Taxa de Alfabetização da População com 15 anos ou mais.	44
Figura 2.8: Evolução do número de Hóspedes nos Estabelecimentos Turísticos.	47
Figura 2.9: Evolução do número de Dormidas nos Estabelecimentos turísticos.....	47
Figura 2.10: Evolução do Ciclo de Vida de um destino turístico.....	56
Figura 2.11: Fase do Ciclo de Vida do Destino Turístico Ilha do Sal.....	57
Figura 2.12: Construto Original do IDT.....	59
Figura 2.13: Construto do IDT (revisto).....	61
Figura 3.1: Exemplo de um correlograma.....	102
Figura 3.2: Percentual da frequência do primeiro dígito de números em um conjunto de dados, segundo a Lei de Benford.....	128
Figura 3.3: Modelo padrão de um neurónio artificial.....	131
Figura 3.4: Rede Neuronal feedforward MLP com duas camadas intermédias.	136
Figura 3.5: Exemplificação do efeito gráfico do comportamento do erro em cada iteração e sua escolha.	152
Figura 4.1: Fase do Ciclo de Vida de Cabo Verde e da ilha do Sal como destino turístico, 2010 e 2018.	157
Figura 4.2: Número mensal de dormidas nos estabelecimentos turísticos da ilha do Sal [2000:01-2018:12].	162
Figura 4.3: Decomposição multiplicativa da série de dormidas nos estabelecimentos turísticos da ilha do Sal [2000:01-2018:12].	163
Figura 4.4: Comparação entre as frequências observadas e relativas do primeiro dígito do número de dormidas mensais nos estabelecimentos turísticos na ilha do Sal [2000:01-2018:12].	168
Figura 4.5: Número de dormidas mensais nos estabelecimentos turísticos na ilha do Sal, no período [2000:01-2018:12].....	170
Figura 4.6: Correlograma simples da série.....	171
Figura 4.7: Transformação da série utilizando o logaritmo para estabilizar a variação.	172

Figura 4.8: Resultado da primeira diferença não sazonal do logaritmo da série.	173
Figura 4.9: Correlograma (ACF and PACF) da série diferenciada.	174
Figura 4.10: Resultado da primeira diferença sazonal da primeira diferença do logaritmo da série.	175
Figura 4.11: Correlograma (ACF and PACF) depois da primeira diferença sazonal do logaritmo da série.	176
Figura 4.12: Correlograma (ACF and PACF) dos resíduos padronizados.	180
Figura 4.13: Distribuição dos resíduos padronizados.	180
Figura 4.14: Análise da normalidade dos resíduos padronizados utilizando o histograma (à esquerda) e o Q-Q Plot (à direita).	181
Figura 4.15: Correlograma (ACF and PACF) dos resíduos padronizados do modelo m2 SARIMA(1,1,1) (0,1,1)[12].	186
Figura 4.16: Distribuição dos resíduos padronizados do modelo m2 SARIMA(1,1,1) (0,1,1)[12]. .	187
Figura 4.17: Análise da normalidade dos resíduos padronizados utilizando o histograma (à esquerda) e o Q-Q Plot (à direita) para o modelo m2 SARIMA(1,1,1) (0,1,1)[12].	188
Figura 4.18: Ajustamento do modelo m2 SARIMA(1,1,1) (0,1,1)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2017:12].	189
Figura 4.19: Ajustamento do modelo m2 SARIMA(1,1,1) (0,1,1)[12] comparando-o com os dados de teste, [2018:01, 2018:12], incluindo a visualização de todos os dados de treino.	190
Figura 4.20: Valores previstos para os próximos 12 meses do ano de 2019 com o modelo m2 SARIMA(1,1,1) (0,1,1)[12] e respetivos intervalos de confiança.	191
Figura 4.21: Valores previstos para os próximos 12 meses do ano de 2019 com o modelo m3 SARIMA(2,1,0) (0,1,1)[12] e respetivos intervalos de confiança.	193
Figura 4.22: Ajustamento dos dados do conjunto de treino aos dados observados utilizando o modelo m5 Holt-Winters.	198
Figura 4.23: Análise dos resíduos obtidos utilizando o modelo m5 Holt-Winters.	200
Figura 4.24: Histograma dos resíduos padronizados resultados do modelo m5 Holt-Winters.	200
Figura 4.25: Validação dos dados para o conjunto de teste, no período de janeiro de 2018 a dezembro de 2018, para o modelo m5 Holt-Winters.	201
Figura 4.26: Validação dos dados para o conjunto de teste, no período de janeiro de 2018 a dezembro de 2018, utilizando o modelo m5 Holt-Winters.	202
Figura 4.27: : Previsão a um passo à frente utilizando o modelo m5 Holt-Winters, incluindo os intervalos de confiança a 80% e 95%.	204
Figura 4.28: Comparação entre os resultados previstos pelo modelo m5 Holt-Winters com os dados observados do número de dormidas mensais em 2019.	205
Figura 4.29: Ajustamento do modelo m11 RNAR(12,1,7)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].	208
Figura 4.30: Validação do modelo m11 RNAR(12,1,7)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].	209
Figura 4.31: Resultado da previsão a um passo à frente usando a função <code>mlp()</code> do R, com o modelo m11	210
Figura 4.32: Comparação dos valores observados com os previstos pela função <code>mlp()</code> com o modelo m11 , ano de 2019.	211
Figura 4.33: Ajustamento do modelo m32 RNAR(18,1,10)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].	213

Figura 4.34: Ajustamento do modelo m32 RNAR(18,1,10)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12].	214
Figura 4.35: Resultado da previsão a um passo à frente usando a função <code>nnetar()</code> do R, com o modelo m32 .	215
Figura 4.36: Escolha da iteração $k=6$.	218
Figura 4.37: Comparação das previsões obtidas do modelo SARIMA m3 e da aplicação do Algoritmo de Atribuição do Erro com os valores observados, ano de 2019.	219
Figura 4.38: Comparação das previsões obtidas pelos modelos SARIMA (m3) versus Holt-Winters (m5).	221
Figura 4.39: Comparação das previsões obtidas pelos modelos m3 SARIMA versus m11 Rede Neuronal.	222
Figura 4.40: Comparação das previsões obtidas pelos modelos m3 SARIMA versus a aplicação do Algoritmo de Atribuição do Erro.	222

LISTA DE TABELAS

Tabela 1.1: Posicionamento de Cabo Verde face aos destinos alternativos e concorrentes.	28
Tabela 1.2: Análise comparativa de Cabo Verde com os destinos alternativos.	28
Tabela 2.1: Cenários de crescimento do número de hóspedes para o período 2019:2030.	53
Tabela 2.2: Matriz de cálculo do IDT.	59
Tabela 2.3: Quantificação do Ciclo de Vida.	60
Tabela 3.1: Variáveis que constituem cada tipo de classificação/categorização dos fatores.	76
Tabela 3.2: Critério MAPE para a Avaliação da qualidade das previsões.	121
Tabela 3.3: Lei de Benford para o primeiro dígito de números em um conjunto de dados.	127
Tabela 3.4: Glossário Redes Neurais x Estatística.	130
Tabela 3.5: Funções de ativação mais utilizadas.	134
Tabela 4.1: Determinação do IDT – Cabo Verde.	158
Tabela 4.2: Determinação do IDT - ilha do Sal.	159
Tabela 4.3: Estatísticas descritivas do número mensal de dormidas nos estabelecimentos hoteleiros da ilha do Sal [2000:01-2018:12].	164
Tabela 4.4: Cálculo do valor do χ^2 para o conjunto de dados do número de dormidas.	167
Tabela 4.5: Modelos escolhidos pelo melhor AICc da aplicação da função <code>auto.arima()</code> . ..	182
Tabela 4.6: Comparação entre os critérios de informação para cada um dos modelos propostos.	183
Tabela 4.7: Comparação entre os modelos propostos com base no seu desempenho nos dados do conjunto de treino e teste.	183
Tabela 4.8: Comparação entre os valores mensais observados com as os valores previstos para os modelos m2 e m3	195
Tabela 4.9: Comparação entre os resultados previstos pelo modelo m5 Holt-Winters com os dados observados do número de dormidas mensais em 2019.	204
Tabela 4.10: Comparação dos erros de treino e teste de cada modelo de Redes Neurais (RN), com uso da função <code>mlp()</code>	207
Tabela 4.11: Comparação dos valores observados com os previstos pela função <code>mlp()</code> para o modelo m11 , ano de 2019.	210
Tabela 4.12: Comparação dos erros de treino e teste de cada modelo de Redes Neurais (RN), com uso da função <code>nnetar()</code>	212
Tabela 4.13: Comparação dos valores observados com os previstos pela função <code>nnetar()</code> para o modelo m32 , ano de 2019.	215
Tabela 4.14: Resultados da aplicação do Algoritmo de Atribuição do Erro, na previsão do ano de 2019.	216
Tabela 4.15: Comparação das previsões obtidas do modelo SARIMA m3 e da aplicação do Algoritmo de Atribuição do Erro com os valores observados do ano de 2019 e os respetivos erros.	218
Tabela 4.16: Resultados comparativos entre os métodos de previsão.	223
Tabela 4.17: Dados mensais do número de hóspedes, dormidas e estadia média por hospedagem do ano de 2019.	224
Tabela 4.18: Previsão, por cenários, do número de dormidas e do número de hóspedes para o ano de 2019.	225
Tabela 4.19: Comparação das previsões feitas com a Análise Lógica Intuitiva com os Modelos Estatísticos e Rede Neuronal, para o ano de 2019.	225

LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

A

ASA – Aeroportos de Segurança Aérea

AR – *Autoregressive*

AIC – *Akaike Information Criterion*

AICc – AIC Corrigido

ACF – *Autocorrelation Function*

ARIMA – *Autoregressive Integrated Moving Average*

ARIMAX – *Autoregressive Integrated Moving Average Exogenous*

ARFIMA – *Fractionally Integrated Autoregressive Moving Average*

ARARMA – *ARMA autoregressive*

ARCH – *Autoregressive Conditional Heteroskedastic*

ADF – *Argumented Dickey Fuller*

ARMAX – *Autoregressive Moving Average Exogenous*

ARAR – *Autoregressive Autoregressive*

AIDS – *Almost Ideal Demand System*

ADLM – *Autoregressive Distributed Lag Model*

B

BM – Banco Mundial

BVAR – *Bayesian Vector Autoregressive*

BGVAR – *Bayesian Global Vector Autoregressive*

BIC – *Bayesian Information Criterion*

C

CRIEF – Centro de Recursos Integrado de Educação e Formação

CTCV – Câmara do Turismo de Cabo Verde

CMS – Câmara Municipal do Sal

D

DL – *Distributed Lag*

E

EUA – Estados Unidos da América

ECV – Escudos Cabo-verdianos

ES – *Exponential Smoothing*

ECM – *Error Correction Model*

EC-LAIDS – *Error Correction-Linear AIDS*

G

GOPEDS – Grandes Opções do Plano Estratégico de Desenvolvimento Sustentável do Turismo

GA – *Genetic Algorithm*

GFS – *Genetic Fuzzy System*

GARCH – *Generalised Autoregressive Conditional Heteroskedastic Model*

GETS – *General-To-Specific*

GVAR – *Global Vector Autoregressive*

GPR – *Gaussian Process Regression*

H

HQC – *Hannan-Quinn information Criterion*

I

IDT – Índice de Desenvolvimento Turístico

INE – Instituto Nacional de Estatística

IMC – Inquérito Multiobjectivo Contínuo

INPS – Instituto Nacional de Previdência Social

IA – Inteligência Artificial

ISS – *Innovations State Space Models*

K

KPSS – Kwiatkowski–Phillips–Schmidt–Shin

L

LVQ – *Learning Vector Quantization*

M

m.c. – Mil Contos

MIOTH – Ministério das Infraestruturas, Ordenamento do Território e Habitação

MF – Ministério das Finanças

MAPE – *Mean Absolute Percentage Error*

MA – *Moving Average*

MIDAS – *Mixed-Data Sampling*

MARIMA – *Multivariate Autoregressive Integrated Moving Average*

MECC – Ministério de Economia, Crescimento e Competitividade

N

NMSE – *Normalized Mean Square Error*

Naïve 1 – *No-Change Model*

Naïve 2 – *Constant Growth Rate Model*

O

OMT – Organização Mundial de Turismo

P

PIB – Produto Interno Bruto

p.p. – Pontos Percentuais

R

RN – Redes Neurais

RB – Ruído Branco

RMSPE – *Root Mean Square Percentage Error*

REG-ARIMA – *Regression with ARIMA errors*

RPROP – *Resilient backpropagation*

S

SARIMA – *Seasonal Autoregressive Integrated Moving Average*

SR – *Static Regression*

SARIMA-In – *SARIMA with Intervention*

SARIMAX – Seasonal Autoregressive Integrated Moving Average cause effect model

SSA – Singular Spectrum Analysis

SEM – Structural Equation Model

STR – Smooth Transition Regression

STSM – Structural Time Series Model

SVR – Support Vector Regression

SVM – Support Vector Machines

T

TALC – Tourism Area Life Cycle

TVP – Time Varying Parameter

TSA – Taxa de Segurança Aeroportuária

U

USD – United States Dollar

UE – União Europeia

UNWTO – United Nation World Tourism Organization

V

VIH – Vírus da Imunodeficiência Humana

VAB – Valor Acrescentado Bruto

VAR – Vector Autoregressive

v.a. – variável aleatória

VEC – Vector Error Correction

VP – Static Varying Parameters

W

WTTC – World Travel & Tourism Council

WI/FI – Sistema de comunicação de internet

WO – Webel Ollech

Z

ZEE – Zonas Económicas Exclusivas

I – INTRODUÇÃO

I.1. Fundamentação

A economia da Ilha do Sal é marcadamente turística, pois em 2017 apresentou um Produto Interno Bruto (PIB) de 21.744 mil contos (m.c.) cabo-verdianos (-2.564 m.c. em relação a 2016) com predominância do setor terciário que gera 2/3 da riqueza da ilha, representando 12,6% do PIB Nacional e resultando um PIB/per capita de 591.369 escudos, o mais elevado de Cabo Verde. Transporte e armazenagem é a principal atividade económica com 23,6% da riqueza da ilha, determinado principalmente pelo aeroporto que é o maior do país.

O *boom* no turismo de Cabo Verde iniciou-se pela Ilha do Sal, a partir do ano de 2002, que continua sendo a maior ilha turística do país, com 30 estabelecimentos turísticos, 10,6% do total existente no País (sendo 15 hotéis, 7 pensões, 4 hotéis-apartamentos, 3 aldeamentos turísticos e 1 residencial). Em 2018, a capacidade de alojamento ronda os 13.451 hóspedes e teve 379.020 hóspedes (49,5%) e 2.764.099 dormidas (56,0% de dormidas nos estabelecimentos turísticos no País) em 2018, um aumento, em dormidas, de 244.612 em relação a 2017, segundo Instituto Nacional de Estatística (INE, 2019d).

Durante o ano de 2018, a empresa Aeroportos e Segurança Aérea (ASA) registou no aeroporto da ilha cerca de 13.429 movimentos de aeronaves (+950 movimentos em relação ao ano de 2017), sendo 25,7% domésticos (- 18,5 pontos percentuais (p.p.)) e 74,3% internacionais (+21,1 p.p.). As movimentações na ilha correspondem a 39,4% das movimentações ocorridas no País. Quanto ao número de passageiros registaram-se 1.184.636 indivíduos, representando um acréscimo de 8,4% em relação a 2017.

Desses passageiros, 84,5% é de natureza internacional (uma variação de 9,9 p.p. em relação a 2017). De notar que o tráfego doméstico de passageiros foi realizado, em 2018, apenas pela companhia aérea *Binter Cabo Verde*. Uma significativa parte do tráfego

internacional esteve relacionado com o crescimento do destino/origem Portugal, Lisboa com mais de 52.716 passageiros transportados. Destaque ainda para os mercados emissores Estocolmo e Gotemburgo que registaram mais 9.439 e 15.921 passageiros, respetivamente, em relação a 2017, não esquecendo da nova rota a partir de Copenhaga que aumentou para 6.592 o número de passageiros (ASA, 2018).

Dos visitantes que chegam à ilha do Sal como destino turístico, no ano de 2018, os do Reino Unido são os mais frequentes (101.302), seguidos dos Países Baixos (41.867) e Alemanha (36.233).

Com relação aos trabalhadores, de ressaltar que do rol de pessoal empregado nos estabelecimentos turísticos em Cabo Verde, 54 em cada 100 empregados estão na ilha do Sal. Cinco mil e sessenta e oito trabalhadores estão ao serviço dos estabelecimentos turísticos, sendo que 55% são Cabo-Verdianos (+70,7 p.p. em relação a 2017) e 42% estrangeiros (+37,7 p.p. em relação a 2017), 56% são do sexo masculino e 52% feminino (em relação a 2017 o número de mulheres diminuiu em 5,4 p.p.).

A nível do número de quartos e camas, Sal é a ilha que possui maior expressão (45,7% e 44,7%) a nível nacional (INE, 2017a; 2019d). Esta ilha realiza cerca de 40,2% do Valor Acrescentado Bruto (VAB) do alojamento e restauração e conjuntamente com a ilha da Boa Vista geram cerca de $\frac{3}{4}$ do VAB desse ramo de atividade económica.

Sal é a terceira ilha em Cabo Verde com maior peso da atividade empresarial, sendo que na ilha está sediada a ASA uma das maiores empresas cabo-verdianas, e os maiores operadores turísticos do País. Das 1.230 empresas na ilha, em 2017, (+241 em relação a 2013), 1.158 estão ativas [INE, 2019h], gerando um volume de negócios de 56.744.329 mil contos (um aumento de 45,8% face a 2012) equivalente a 19,7% da faturação das empresas Nacionais.

Diante de toda essa dinâmica na maior ilha turística do país, outros produtos turísticos diferentes de 'sol e praia' vão surgindo dentro do arquipélago, nomeadamente o turismo rural e de montanha. Assim, é eminente uma exigência maior em termos de distribuição dos investimentos entre as ilhas de vocação turística diferenciada, pelo que é importante identificar as particularidades, potencialidades e necessidades da ilha do Sal e conhecer

em que fase se encontra no seu ciclo de vida para que o destino possa continuar a desenvolver em meio à potencial disputa pelos recursos escassos que o país tem para investir no setor turístico.

Ainda explicar o comportamento da procura turística na ilha bem como utilizar de ferramentas estatísticas para realizar previsões que possam ajudar na planificação e gestão do destino. Tais conhecimentos permitem agir de forma que os investimentos em serviços e estabelecimentos turísticos sejam bem alocados/justificados e o produto turístico seja melhor diversificado para que a ilha possa manter o *status* de ser considerada como um motor do desenvolvimento económico de Cabo Verde.

Neste trabalho, pretende-se com a utilização de diferentes técnicas perceber as potencialidades de cada uma aplicada às características peculiares que as séries de turismo evidenciam, tais como, a sazonalidade e a tendência. Com relação à sazonalidade, segundo Song e Witt (2000) esta não pode ser ignorada no processo de modelação quando dados mensais ou trimestrais são usados. Com este objetivo, utiliza-se o número de dormidas mensais registados nos estabelecimentos turísticos na Ilha do Sal, no horizonte temporal compreendido entre janeiro de 2000 e dezembro de 2018; estes dados são, em sua maioria, referentes aos 6 maiores países emissores turistas da ilha do Sal (Alemanha, Países Baixos, França, Reino Unido, Itália e Portugal), pertencentes ao mercado europeu. A Figura I.1 representa o mapa *mundi* a localizar Cabo Verde e a ilha do Sal diante do mercado europeu.

Optou-se por considerar ‘o número de dormidas mensais’ como a variável de interesse não apenas por ser a mais estudada para modelar a procura turística (Fernandes, Monte & Teixeira, 2009) mas também por englobar tanto as dormidas dos estrangeiros como dos Cabo-Verdianos, permitindo que a avaliação sobre os movimentos turísticos seja melhor trabalhada e mais realista (Cunha & Abrantes, 2013).

De referir que uma grande parte deste trabalho decorre durante a pandemia da Covid-19 que teve seu início, em Cabo Verde e na ilha do Sal, em março de 2020. Esta pandemia que tem abalado o mundo inteiro e especificamente o continente europeu (o maior

mercado emissor de turistas para a ilha do Sal) restringiu consideravelmente o número de dormidas nos estabelecimentos turísticos da ilha. Esta crise sanitária à escala global tem afetado e fragilizado a economia e provocado uma recessão económica com efeitos no desenvolvimento das atividades turísticas, deixando todos neste setor vulneráveis perante a imprevisibilidade da durabilidade do vírus e a versatilidade da sua universalidade. Assim a realização das previsões de dormidas para os anos de 2020 até 2030 não será contemplada e, portanto, este trabalho se limita a estudar a série de dormidas mensais no horizonte de 2000 a 2018 e a realizar as previsões para os 12 meses do ano de 2019.



Figura I.1: Carta de localização de Cabo Verde, ilha do Sal e o mercado europeu com os maiores países emissores de turistas para o destino ilha do Sal.

Fonte: Elaboração própria

I.2. Objetivos e Metodologia da Tese

Este trabalho, nos termos ora proposto, é o primeiro a ser realizado em Cabo Verde e particularmente para a Ilha do Sal, e tem como objetivo utilizar diferentes técnicas de estatística, nomeadamente o uso do método de Box-Jenkins na modelação de Séries Temporais (ver definição de séries temporais na subsecção 3.4.1) e das Redes Neuronalis Artificiais (usando a rede *feedforward* com o algoritmo de treino *backpropagation*), para modelar e prever a procura turística na ilha do Sal. Geralmente os modelos SARIMA (*Seasonal Autoregressive Integrated Moving Average*), do método Box-Jenkins, têm sido os mais frequentemente utilizados neste tipo de investigação (Cortez, 1997; Miguel, 2011; Constantino, Fernandes, & Teixeira, 2016; Song, Qiu & Park, 2018). Vários estudos têm mostrado que as Redes Neuronalis são mais eficazes na previsão da procura turística em comparação com a Regressão Linear Múltipla (Uysal & El Roubi, 1999; Law & Au, 1999; Law, 2000; Machado, Fernandes & Teixeira, 2010; Cankurt & Subasi, 2012).

Song, Qiu e Park (2018), citando Pai e Hong (2005), também dizem que em termos de previsão as Redes Neuronalis (RN) têm apresentado melhores resultados em comparação com os métodos tradicionais lineares e não lineares. Segundo Song, Qiu e Park (2019) outros modelos de séries temporais com a abordagem de Box-Jenkins e as suas variáveis exógenas, nomeadamente o ARIMAX (*Autoregressive Integrated Moving Average Cause Effect Model*) e o SARIMAX (*Seasonal Autoregressive Integrated Moving Average Cause Effect Model*), também têm vindo a demonstrar ser superiores aos demais.

A escolha em trabalhar com modelos de Séries Temporais e RN tem também como propósito a comparação entre eles permitindo verificar qual se adapta melhor aos dados da série em estudo considerando a precisão na previsão do número de dormidas mensais. No entanto, neste trabalho, admite-se por hipótese, que modelos em teoria adequados podem na prática não revelar-se como tal, caso existam fatores que influenciam negativamente o registo dos dados. Nesta linha, utiliza-se a Lei de Benford para verificar o comportamento dos dados e determinar até que ponto a adequação dos mesmos à Lei pode influenciar o desempenho dos modelos de previsão. Esta tentativa de relacionar a

Lei de Benford com o comportamento dos modelos é a primeira ser levada a cabo nos estudos de previsão da procura turística.

Em síntese, com esta tese pretende-se:

1. Determinar em que etapa do ciclo de vida se encontra o destino (ilha do Sal) e identificar a posição relativa do mesmo em termos de competitividade com mercados com o mesmo produto turístico;
2. Descrever a série temporal do número de dormidas mensais e analisar os possíveis métodos e modelos a aplicar;
3. Obter diferentes modelos de previsão, comparar e propor aquele que melhor se adapta aos dados no cálculo da previsão para o ano de 2019;
4. Compreender a influência no desempenho dos modelos de previsão da adequação dos dados de dormidas à Lei de Benford;
5. Propor um modelo final de previsão que se adequa aos dados da série de dormidas mensais em estudo.

Para atingir os objetivos propostos foi necessário primeiramente estruturar o trabalho da forma como se apresenta na Figura I.2.

I.3. ESTRUTURA DA TESE

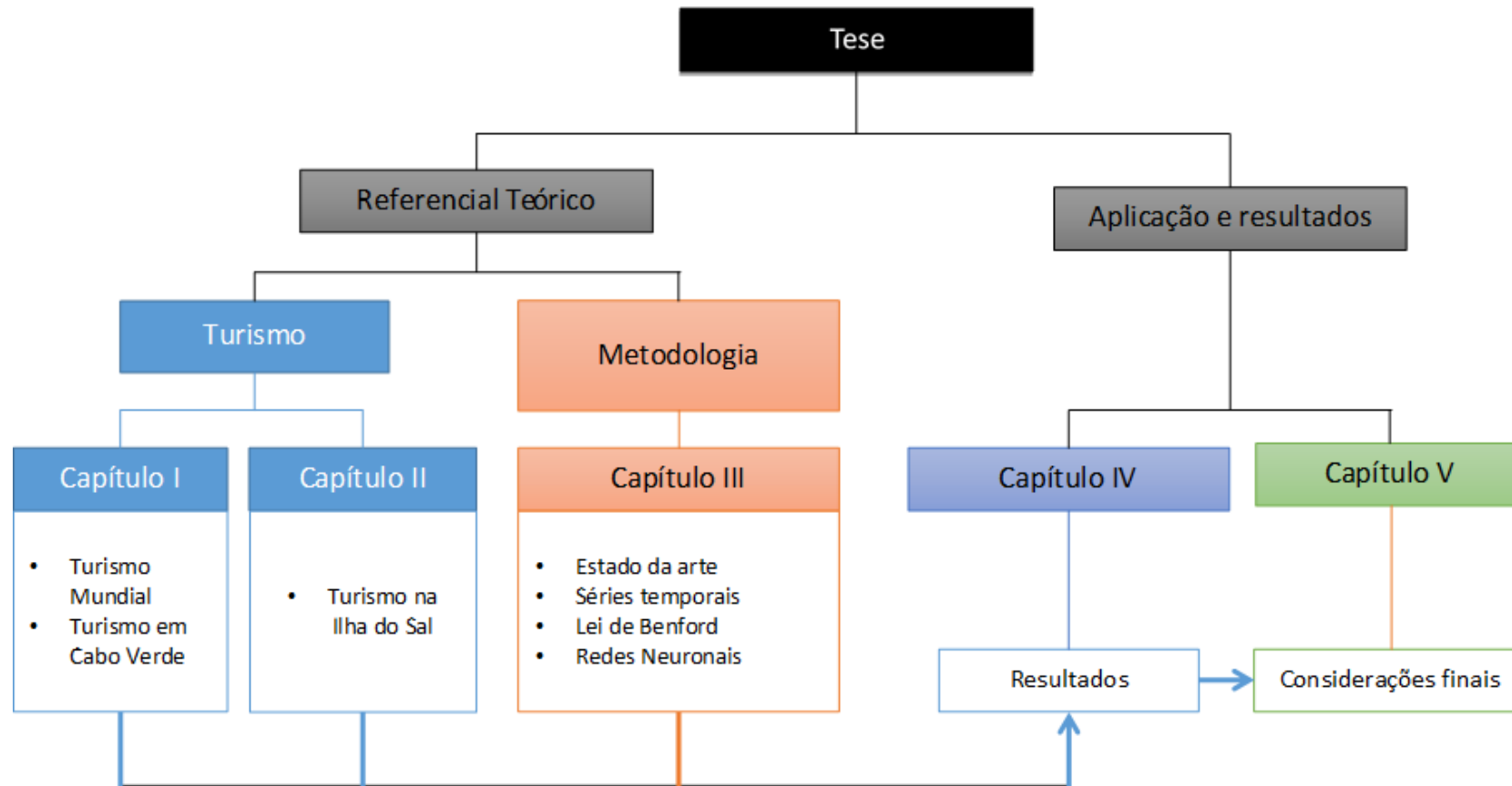


Figura I.2: Estrutura da Tese

CAPÍTULO 1: TURISMO – O MUNDO E CABO VERDE

1.1 INTRODUÇÃO

Este capítulo apresenta a dinâmica do turismo mundial e em Cabo Verde. A nível global é descrito o fluxo de turistas e alguma tendência emergente. A nível nacional é feito um balanço do crescimento do turismo, dos problemas existentes, das potencialidades; uma análise do principal mercado emissor (Europa); uma análise quanto a competitividade com relação aos destinos concorrentes, bem como das metas e previsões com cenários até 2030.

1.2 TURISMO – DINÂMICA MUNDIAL

A nível global, os fluxos de turistas têm vindo a crescer desde 2010, ano este, que se registou 952 milhões de chegadas, tendo atingido os 1.401 milhões em 2018 (uma variação de 5,4% em relação ao ano anterior), de acordo com o relatório da Organização Mundial de Turismo (OMT, 2019). Segundo a mesma fonte, o crescimento nas chegadas traduziu-se num forte aumento nas exportações geradas pelo turismo, que atingiram 1,6 triliões de dólares em 2017 e mais de 1,7 triliões em 2018, tornando o turismo num dos mais importantes setores de exportação do mundo. Para além dos 1,6 triliões de dólares em receitas que os destinos receberam, o turismo internacional gerou outros 256 bilhões de dólares de transporte internacional de passageiro feito por não residentes (OMT, 2019).

Estes resultados expressivos foram impulsionados pelo aumento da procura nos continentes Ásia/Pacífico e África, cada um tendo um aumento de 7,0% na chegada de turistas comparado com os 324 e 63 milhões no ano de 2017, respetivamente. A receita

obtida com o fluxo turístico em África é menor (38 bilhões de dólares) do que a relativa à Ásia e Pacífico (435 bilhões).

Dentre os países com maior fluxo de turistas (40,0% do fluxo mundial) estão a França, Espanha, Estados Unidos da América (EUA), China, Itália, Turquia, México, Alemanha, Tailândia e Reino Unido. Com relação ao ganho com turistas, os países como EUA, Espanha, França, Tailândia, Reino Unido, Itália, Austrália, Alemanha, Japão e China são os que receberam 50,0% da receita total com o turismo em 2018. De realçar que a Turquia e o México tiveram um aumento no número de turistas de 22,0% e 5,0%, respetivamente, em relação a 2017 e a Austrália e Japão tiveram um aumento considerável nas suas receitas de 11,0% e 19,0%, respetivamente. Neste caso, o volume de turistas na Austrália aumentou de 4,9% em 2017 para 2018 e no Japão 8,7%, o que indica que o turismo nestes países é mais qualitativo (OMT, 2019).

A OMT (2019) indica que o bom momento do turismo mundial, principalmente em regiões como Ásia e África, está relacionado com alguns fatores como: ampliação da malha aérea de conexão principalmente nas regiões onde era inexistente a conectividade aérea; estabilidade do custo do combustível; e a redução de taxas, cerca de 60,0% em relação a 1998.

Em 2018 o continente africano sobressaiu-se nos resultados, principalmente na performance do Norte da África, tendo a Tunísia e Marrocos como destaque desta região (10,0% a mais de chegadas de turistas e um aumento de 4,0% em receitas). Este resultado deve-se, principalmente, ao índice de segurança nestes países ter melhorado. O continente africano contou com a chegada de 67 milhões de turistas em 2018, o que gerou 38 bilhões de dólares americanos (USD) de receita.

Na África Subsaariana, destinos insulares como: Cabo Verde, Comores, Ilhas Maurício e Seychelles também registaram um significativo crescimento, com destaque para Comores que teve um aumento de 28,2% em relação a 2017 e Cabo Verde que teve 4,9% de aumento no volume de turistas internacionais chegando a 710 mil em 2018 contra 668 mil em 2017, gerando uma receita no país de 466 milhões de USD (OMT, 2019). O Quênia

registou resultados positivos, graças à melhoria da segurança e ampliação da conectividade aérea, assim como Togo, Uganda, Costa do Marfim e Zimbábue. O destino mais visitado, a sub-região da África do Sul relatou crescimento moderado, em parte devido a uma moeda forte e a crise de seca na Cidade do Cabo em 2018.

A Figura 1.1 mostra que até 2020 é previsto cerca de 1,4 bilhões de turistas a se deslocarem a diferentes destinos ao redor do mundo e até 2030 cerca de 1,8 bilhões.

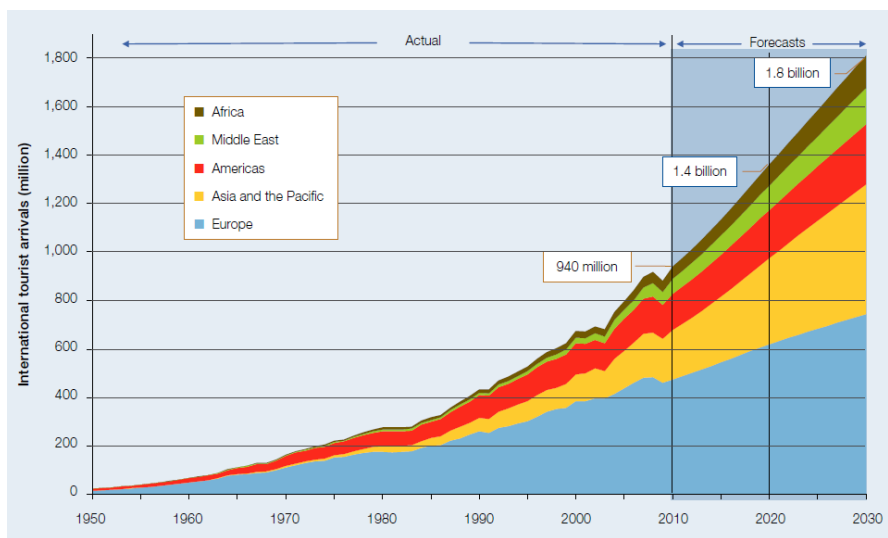


Figura 1.1. Previsão da procura turística em diferentes destinos do globo.

Fonte: UNWTO (2018, p.64)

1.3 TENDÊNCIAS

O mercado turístico mundial é competitivo e mudanças a nível das novas tecnologias já estão a ser introduzidas. A comunicação social digital e redes sociais estão a permitir o desenvolvimento de novas tendências de hábitos de consumo turístico, sendo eles, por exemplo, a busca de autenticidade e transformação, divulgação em redes sociais de momentos, experiências e destino (OMT, 2019).

As experiências relacionadas com a cultura, natureza, sol e praia destacam-se no setor turístico mundial, sendo seus protagonistas os turistas chamados de *millenials*. Estes surgiram em 1980 e são constituídos maioritariamente por mulheres com elevado nível académico e devidamente informados sobre o destino, por meio do uso da tecnologia, e

valorizam mais a autenticidade, a personalização, a sustentabilidade e um estilo de vida saudável. Os *millenials* aumentaram as suas viagens em 15,0% em 2017 sendo a maioria procedentes dos EUA, China, Reino Unido e Alemanha representando cerca de 35,0% das viagens pelo mundo. Por exemplo, na China, 60,0% dos viajantes são *millenials* e 50,0% viajam de forma independentes, ou seja, sem assistência de agências de viagens, isto derivado às novas tecnologias. Os destinos mais visitados por eles são Japão, Coréia do Sul, Índia, Austrália, Europa, Irlanda e Escandinávia. Os *millenials* são mais consumidores de viagens a cidades do que de sol e praia. Fazem parte dos outros três segmentos de turismo com perspectivas de crescimento: turismo de cidades, cruzeiros e turismo halal de muçulmanos (IPK international, 2018).

1.4 O TURISMO EM CABO VERDE

Cabo Verde é um país insular de 4 mil km² formado por um arquipélago composto por dez ilhas de origem vulcânica localizadas na região central do Oceano Atlântico, a cerca de 500 quilómetros da costa ocidental africana. A sua população em 2018 é de mais de 543.000 (INE, 2019a), espalhada em nove ilhas separadas até 300 km por mar. Cerca de 85,0% da população está concentrada em quatro das nove ilhas habitadas: Santiago (56,2%), São Vicente (15,4%), Santo Antão (10,1%) e **Sal (6,8%)**. A ilha Brava representa apenas 1,0% da população do país (INE, 2019a).

A elevada emigração e rápida queda na taxa de natalidade levou Cabo Verde a uma transição demográfica mais rápida do que no resto da África. A população emigrante é estimada entre 750.000 e 1.000.000 de indivíduos, sendo que muitas saíram antes de 1975, ano da independência do país. Nos últimos 30 anos, um quinto da população atual emigrou para o exterior. Segundo o Banco Mundial (BM), a maioria teve como destino Portugal, França e Estados Unidos (BM, 2019). Os emigrantes apresentam um papel importante no desenvolvimento de Cabo Verde por meio dos recursos financeiros que introduzem na economia local através de remessas, que têm atingido nos últimos anos um pouco mais de 10% do Produto Interno Bruto (PIB).

Dados do Ministério das Infraestruturas, Ordenamento do Território e Habitação (MIOTH) mostram que o país é relativamente urbanizado, com cerca de 68,0% da população vivendo em áreas urbanas com previsão de atingir 73,0% em 2030 (MIOTH, 2019), sendo o segundo país mais urbanizado em África. A urbanização tem sido impulsionada pelas secas frequentes e a criação de empregos nas atividades relacionadas com o turismo nos centros urbanos.

A economia de Cabo Verde depende em grande parte do setor de serviços, que responde por aproximadamente 74,0% das atividades económicas e 65,0% da participação da força de trabalho, sendo dominado pela indústria do turismo, com 45,0% do PIB e 39,0% do emprego (BM, 2019). O comércio, hotéis e restaurantes (suportado pelo setor do turismo) são responsáveis por cerca de 23,0% dos postos de trabalho, seguido pela administração pública e serviços públicos (19,0%).

Em termos do PIB, as ilhas que apresentaram um maior peso na estrutura económica do país em 2016 foram Santiago, São Vicente e **Sal**, com 52,1%, 14,8% e 14,7%, respetivamente. As ilhas que apresentaram o maior PIB per capita em 2016, foram as ilhas do **Sal (689.219 escudos cabo-verdianos)** e da Boa Vista (485.219 escudos cabo-verdianos), destacando-se significativamente das outras. As ilhas do Fogo, Maio e Brava são as que apresentaram o menor PIB per capita de Cabo Verde (INE, 2019b).

O país tem-se destacado no turismo mundial por conta de suas excelentes praias brancas, pelo mar transparente, clima ameno, pela possibilidade de praticar desportos náuticos, pelas belas paisagens naturais, pelo ecoturismo, pelos marcos históricos (por ex. a Cidade Velha na ilha de Santiago e as Salinas de Pedra de Lume na ilha do **Sal**) entre outras diversas características que atraíram os 765.696 turistas no ano de 2018.

O setor do turismo tem sido bastante impulsionado no país tanto por políticas públicas que visam fortalecer o turismo de maneira sustentável, como pela criação de incentivos fiscais e aduaneiros para incentivar novos investimentos. Com quatro aeroportos internacionais situados nas ilhas de Santiago, São Vicente, Boa Vista e **Sal**, constata-se que a atividade turística vem registrando crescimento a uma taxa média superior a 8,0%

ao ano, com tendência crescente a uma taxa média anual de mais de 10,0% durante os últimos dois anos, mais que o dobro da taxa mundial (GOPEDS-Turismo, 2019).

Os principais destinos turísticos de Cabo Verde são as ilhas do **Sal** e Boa Vista, que recebem 76,7% do total de turistas estrangeiros do país. Essas ilhas oferecem hotéis de classe média a alta e instalações de *resort* do tipo “*all inclusive*”. Nas demais 7 ilhas há um número limitado de turistas. Com exceção das ilhas de Santiago, **Sal** e Boa Vista, as demais ilhas ainda não desenvolveram muitos empreendimentos hoteleiros, instalações turísticas e infraestrutura de transporte devido às limitações nos investimentos públicos e privados para o desenvolvimento do setor de turismo.

Nos últimos dezoito anos (2000-2018) verificou-se um aumento de entradas e dormidas (Figura 1.2 e 1.3), passando de 145.076 para 765.696 (uma taxa de entrada por ano em torno dos 9,9%) e de 684.733 para 4.935.891 (uma taxa de dormidas por ano em torno dos 11,9%) respetivamente, chegando a atingir em 2012, pela primeira vez na história do país, 533.877 mil hóspedes, um número maior do que a população Cabo-verdiana residente nesta data, o que se traduz numa contribuição económica bastante importante. Apresenta ainda ligeiras quedas nos anos 2002, 2004, 2009 e 2014. De notar que a partir do ano 2010, já se pode constatar um aumento maior de entradas e dormidas dos turistas, muito embora, os principais países emissores tenham sido bastantes fustigados pela crise económico-financeira (INE, 2020).

Cabo Verde tem cerca de 284 estabelecimentos turísticos, um número que aumentou em 3,3% comparado ao ano de 2017. Esses estabelecimentos proporcionam uma capacidade de alojamento de 13.187 quartos (+5,8%), 21.046 camas (+3,1%) e 27.860 lugares sentados (+3,2%). Desse número a ilha de Santo Antão possui a maioria (71, correspondente a 25,0%) e a ilha do **Sal** a minoria (30, correspondente a 10,6%). Considerando a tipologia dos estabelecimentos turísticos, as Pensões passaram a ter maior peso (31,0%), seguindo-se as residenciais (28,9%) e os hotéis (25,7%). Do total de quartos e de camas, a ilha do Sal continua a ter maior expressão, 45,7% e 44,7%

respetivamente. Em termos de distribuição de camas, os hotéis apresentam uma maior capacidade, ou seja, mais de três quartos da capacidade total (INE, 2019c).

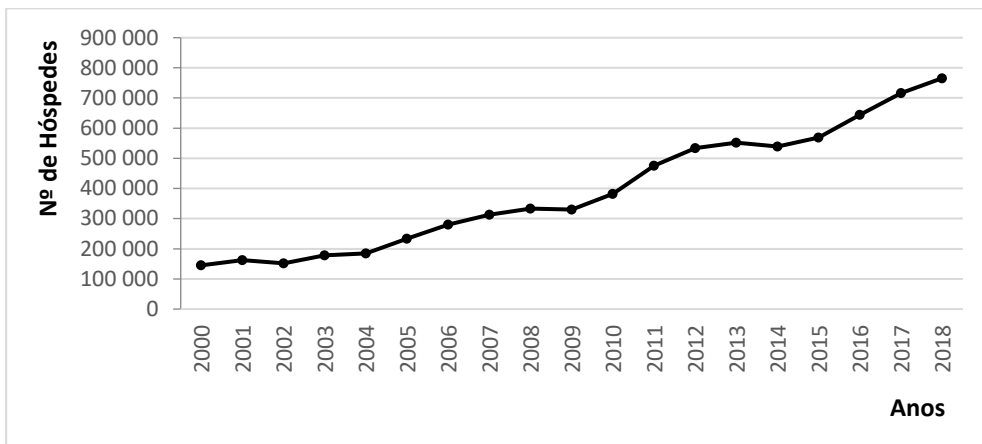


Figura 1.2: Evolução do número de hóspedes em Cabo Verde (2000:2018).

Fonte: Elaboração própria, com base em dados do INE (2001-2019)

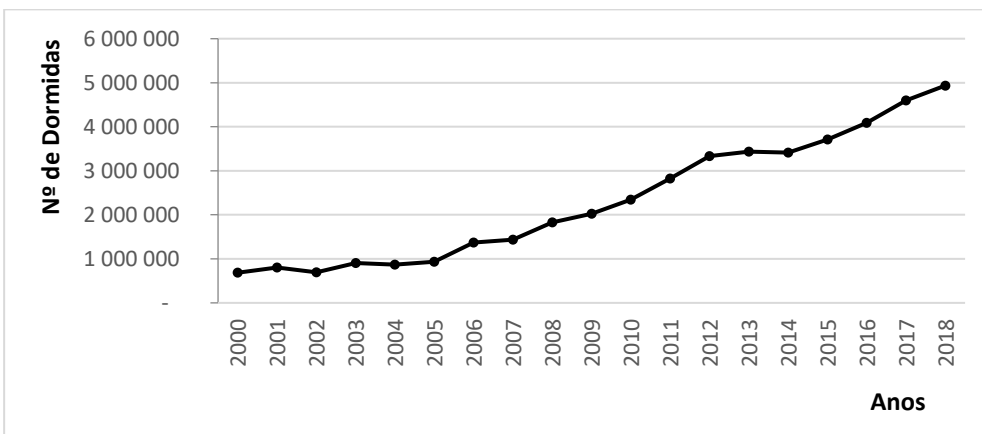


Figura 1.3 Evolução do número de dormidas em Cabo Verde (2000:2018).

Fonte: Elaboração própria, com base em dados do INE (2001-2019)

A nível do emprego, cerca de 9.417 indivíduos estão empregados nos estabelecimentos turísticos do País, um número que tem vindo a crescer com um aumento de 6,7% em relação ao ano de 2017. Dessa quantidade, 84,3% estão nos hotéis e cerca de 54 em cada

100 empregados estão na ilha do **Sal**. Quanto ao sexo, 59,3% dos empregados nos estabelecimentos turísticos são mulheres e 40,7% homens (INE, 2019c).

Quanto a serviços prestados, 97,2% dos quartos estão equipados com água quente e fria para banhos privados; 91,3% com televisão; 88,4% com ar condicionado e 79,0% com telefone. Para além dos quartos, 94,7% dos estabelecimentos possuem telefone; 31,7% tem fax; 74,6% tem computador; 78,2% tem internet; 53,5% oferecem bar; 54,9% tem restaurante; 66,2% tem WI-FI e 46,1% tem televisão por cabo (INE, 2019c).

Quanto ao preço praticado, nas épocas consideradas altas e baixas, os preços médios por quarto nas ilhas do **Sal** e da Boa Vista são os mais elevados e mais baixos nas ilhas de São Nicolau e Brava (INE, 2019c).

Em termos de mercados emissores, os europeus representam aproximadamente 71,0% dos fluxos, sendo os principais países o Reino Unido com 22,7%, a Alemanha 11,8% e a França 10,1%; seguido dos Países Baixos 10,0%, Portugal 9,3%, Itália 4,0% e outras origens (não precisados nos relatórios oficiais do INE) cerca de 20,2%, e os nacionais 7,1% (Figura 1.4, p. 15). É importante destacar que o número de turistas ingleses em Cabo Verde tem vindo a aumentar, grande parte devido ao início de funcionamento de grandes hotéis associados a grandes operadores turísticos ingleses, a TUI Thomson e Thomas Cook. Este tipo de turista, segundo o INE (2019c) preferiram, em sua maioria, como destino a Ilha do **Sal**, constituindo 56,9% das dormidas, como estabelecimento de acolhimento os hotéis com 93,2% das dormidas e tiveram a maior permanência média (8,3 noites) comparando com os outros mercados emissores (Figura 1.5). A ilha do **Sal** é o destino que continua a ter maior acolhimento, com 49,5% do total das entradas e 56,0% das dormidas no País, seguida da ilha da Boa Vista (26,9%, 33,9% respetivamente) e Santiago (11,2% e 4,5% respetivamente). Quanto à taxa de ocupação-cama a ilha do **Sal** e Boa Vista lideram, sendo 80,0% e 61,0% do total nacional (INE, 2019c).

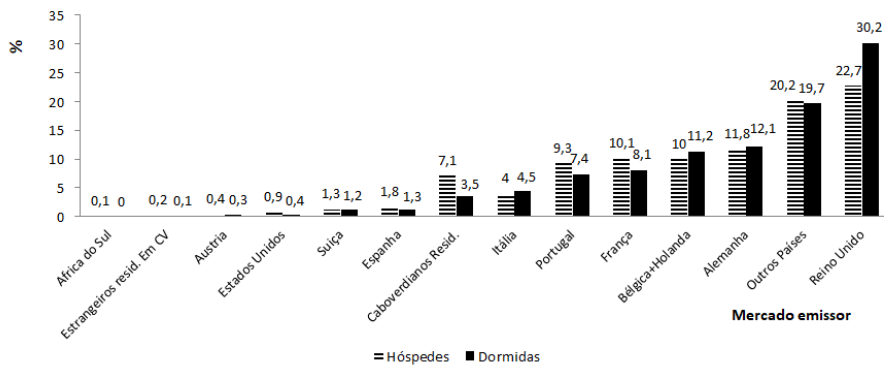


Figura 1.4: Hóspedes e Dormidas (%) por mercado emissor dos hóspedes, 2018.

Fonte: Adaptado do INE (2019d)

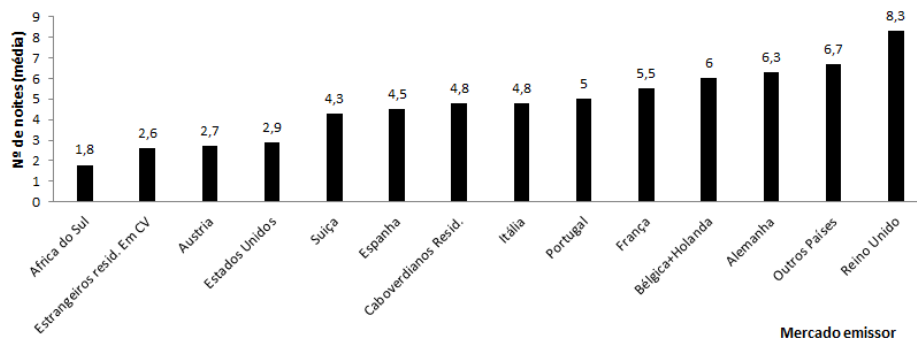


Figura 1.5: Estadia média (noites) por mercado emissor dos hóspedes, 2018.

Fonte: Adaptado do INE (2019d)

Do total dos turistas que visitaram Cabo Verde em 2018, a maioria foi do sexo feminino 53,5% (+ 7,8 p.p. com relação a 2017) e 46,2% masculino (-7,3 p.p. com relação a 2018), com idade média de 45 anos. Para 94,2% (+2,7 p.p. comparando com 2017) dos turistas o motivo principal da viagem era férias. As viagens foram feitas de duas formas: por conta própria (17,9%) ou organizada por agências e turoperadores (82,1%), segundo o Inquérito aos Gastos e Satisfação dos Turistas, realizado pelo INE (2019e). Ainda, da mesma fonte sabe-se que a estadia em Cabo Verde foi considerada excelente (48,0%), boa (45,5%), razoável (5,6%), má (0,6%) e muito má (0,2%). Cerca de 97,8% recomenda o destino Cabo Verde e 95,0% tem intenção de regressar mais vezes.

A maioria dos turistas (58,3% em 2017 e 55,6% em 2018) viajou para o País na chamada época alta (referente aos meses de outubro a março) e 41,7% (em 2017) e 44,4% (em 2018) na época baixa (referente aos meses de abril a setembro), sendo que 81,0% do total viajaram em pacote turístico e destes 75,8% no sistema *all inclusive* (+1,0p.p. com relação a 2017). De realçar que a receita mínima anual dos agregados familiares dos turistas que visitam Cabo Verde é de pelo menos 36 mil euros. Cerca de 83,7% dos turistas tinham uma ocupação profissional no país de origem, seguido de 12,2% de reformados e 3,7% não exerciam qualquer atividade económica. Com relação às tendências, 47,0% dos turistas obtiveram informação do destino Cabo Verde através da internet; 29,8% através de agências de viagem e operadores turísticos e 19,0% por meio de amigos (INE, 2017a, 2018a, 2019e).

1.5 O MERCADO TURÍSTICO EUROPEU

Levando em consideração o mercado Europeu como principal mercado emissor para Cabo Verde, pode dizer-se que em geral os residentes da Europa estão a viajar mais dentro da própria região. Segundo dados da Eurostat (2018a), em 2017 os europeus fizeram cerca de 1.2 bilhões de viagens turísticas, sendo 94,0% para destinos dentro da Europa e geralmente dentro do próprio país, com permanência em média de 5,1 noites. Comparado com 2016, este resultado teve um aumento de 4,0%. A Polónia foi o destino europeu com a maior proporção de dormidas (94,0%). Malta destaca-se em relação a turistas vindo de outros países da União Europeia (84,0%) e Chipre destaca-se em relação a turistas de países fora da União Europeia (36,0%).

As regiões costeiras têm-se tornado muito populares com relação ao número de dormidas. Segundo a Eurostat (2018b), em 2016, com relação às regiões costeiras, cerca de 103 milhões de dormidas foram nas Canárias (sendo 91 milhões não residentes); na Catalunha (80 milhões); na Jadranska Hrvatskana Croácia (74 milhões) e nas ilhas Baleares (70 milhões). Com relação aos gastos, em 2016, os europeus gastaram em média 336 euros por viagem de férias (bilhetes de avião), dentro da Europa, sendo os

Luxemburgueses quem mais gastaram (768 euros) seguido de Malta (646 euros) e da Áustria em cerca de 607 euros (Eurostat, 2018c).

De notar que os turistas chineses têm vindo a aumentar as dormidas na Europa desde 2006 (356 mil dormidas), atingindo em 2016 cerca de 25 milhões de dormidas. O número de dormidas dos turistas chineses representa cerca de 6,0% do total de noites passadas na União Europeia por turistas fora desta região. Geralmente os turistas chineses preferem pernoitar em quatro países: Reino Unido (31,0%), Itália (18,0%), França (13,0%) e Alemanha (10,0%). A China está no 8.º lugar entre os dez principais destinos fora da Europa para onde os europeus mais viajam, constituindo 3,0% dessas viagens em 2016, ou seja, 2,1 milhões de viagens feitas por europeus teve como destino a China, perfazendo um total de 32,5 milhões de noites e um gasto de 5,1 bilhões de euros. Olhando para o mercado asiático, a China é o terceiro destino mais procurado pelos europeus, com 13,7%. Outros países mais procurados pelos europeus, entre os dez principais, estão os Estados Unidos (14,0%), Turquia (10,0%), Suíça (9,0%), Noruega (6,0%), Rússia (6,0%) e Tailândia, Emirados Árabes Unidos, Marrocos e Egito com 3,0% respetivamente (Eurostat, 2018d).

Em 2017 o número de dormidas na Europa atingiu os 3,2 bilhões, um aumento de 5,1% com relação a 2016. Nesse ano, a Espanha liderou o número de dormidas (471 milhões, +3,6% p.p.) seguido da França (431 milhões, + 6,6% p.p), Itália (425 milhões, +5,4% p.p.) e Alemanha (400 milhões, +2,7% p.p.). Existindo um aumento de dormidas em todos os países da União Europeia (UE), exceto o Luxemburgo (-28 mil dormidas). De notar que o número de dormidas por não residentes tem vindo a crescer mais do que o dos residentes entre os anos 2016 e 2017 em todos os estados membros da UE, com a exceção do Luxemburgo. A Espanha lidera também no número de dormidas em estabelecimentos turísticos de não residentes (307 milhões) (Eurostat, 2018e).

Relativamente ao número de camas, a França foi em 2015, o destino com o maior número de camas nos estabelecimentos turísticos (5,1 milhões, 16,5% do total de 31,2 milhões de camas disponíveis nos estabelecimentos turísticos da União Europeia),

seguida pela Itália (15,6%) e Espanha (11,3%) (Eurostat, 2017a). Quanto aos gastos, em média os turistas europeus residentes, gastaram cerca de 320 euros em viagens de férias, no ano de 2015 (Eurostat, 2017b).

Cerca de 58,0% das viagens feitas pelos turistas europeus dentro da Europa, em 2015, foram de no máximo três noites e três quartos dessas viagens foram domésticas. Tal se verificou principalmente na Roménia, Espanha, Portugal, França e Grécia, onde mais de 85,0% das viagens foram domésticas. Os principais meios de transportes utilizados pelos turistas europeus, dentro da Europa, foram as caravanas (64,0%), seguidos dos aviões (16,0%) e comboios (11,0%) (Eurostat, 2017c).

Os residentes dos países que mais viajaram de avião em 2017 foram de Malta e Chipre, seguidos pela Irlanda e Luxemburgo. Nenhum destes países é a origem dos turistas que mais visitam Cabo Verde. Por outro lado, os franceses e alemães viajaram mais internamente e o meio de transporte mais utilizado foi o comboio em 2017. O motivo destas viagens foi: Lazer e férias (49,0%); visita a parentes e amigos (35,0%); negócios (12,0%) e outros (4,0%).

A percentagem de belgas que viajaram para o exterior alcançou 80,0%, uma das mais altas entre os residentes, com uma média de pernoites de 6,5 noites, juntamente com os holandeses. Ambos figuram na lista de principais mercados emissores para Cabo Verde.

1.6 POTENCIALIDADES DO TURISMO CABO-VERDIANO

Segundo GOPEDS-Turismo (2019), sendo Cabo Verde um arquipélago, a diversidade do turismo deve ser vista a nível da peculiaridade de cada ilha como destino turístico, tendo sido identificados seis segmentos principais do turismo no País:

- (i) Turismo de sol e praia (principalmente nas ilhas do **Sal**, Boa Vista e Maio);
- (ii) Turismo rural e de natureza (principalmente nas ilhas de Santo Antão, São Nicolau, Santiago, Fogo e Brava);
- (iii) Turismo urbano, cultural e de eventos;

- (iv) Turismo de cruzeiros. As ilhas de São Vicente, **Sal**, Boa Vista e Santiago são as que estão mais preparadas para desenvolver este segmento, sendo que outras ilhas vizinhas beneficiarão. Com este tipo de segmento existe a possibilidade de explorar não apenas o circuito da macaronésia como também o circuito do eixo Europa América do Sul e o circuito África Ocidental;
- (v) Turismo de circuito. É um segmento que tem vindo a ganhar interesse, pois os turistas procurem experiências novas, ricas, variadas e surpreendentes. Para desenvolver este segmento é necessário a existência de miradouros, passadiços, centros interpretativos, sinalética, postos de observação de pássaros, unidades hoteleiras de pequena dimensão, restauração, etc.;
- (vi) Turismo náutico e desportivo. Sendo Cabo Verde um arquipélago, o potencial de desenvolver esse tipo de turismo é enorme. Por exemplo mergulho, pesca desportiva, *surfing*, *kite surfing* e jogos diversos de praia já é uma realidade no país e principalmente na ilha do **Sal**.

Entretanto outros segmentos têm vindo a emergir, nomeadamente de saúde e bem-estar, científico e religioso.

Sem dúvida que o turismo é um dos principais motores da economia cabo-verdiana, com efeitos macroeconómicos sobretudo no PIB onde atinge mais de 20% (GOPEDS-Turismo, 2019). O turismo é um eixo de desenvolvimento bastante recente (anos noventa), impulsionado por diversos fatores onde podemos destacar: a crescente visibilidade conferida por fenómenos associados à cultura, do qual Cesária Évora (A Diva dos pés descalço) é exemplo eminente; peculiaridade das ilhas em termos populacionais e de idiossincrasia; a localização geográfica; e os recursos naturais, desde logo as extensas praias brancas, as montanhas, a agricultura, a floresta, a pecuária, as pescas, a gastronomia, a observação da fauna e da flora, ambiente ecológico, parques e reservas naturais e centros interpretativos.

1.7 PROBLEMAS DO TURISMO EM CABO VERDE

Ainda que o turismo seja o principal impulsionador da economia cabo-verdiana, existem ainda muitos constrangimentos neste setor.

O Plano Estratégico de Desenvolvimento do Turismo 2010-2013 elaborado pelo Governo (MECC, 2014) estruturou alguns problemas com o turismo em Cabo Verde, nomeadamente: (1) acessos, existindo assim, uma insuficiência de ligação entre as ilhas e países emissores; (2) problemas de infraestruturas, como por exemplo nos setores da saúde, água e energia, saneamento, educação, infraestruturas turísticas, etc.; (3) problema da sustentabilidade, com uma política desenhada para a preservação ambiental; e (4) problemas da monotorização do turismo.

Já o GOPEDS-Turismo (2019) e o Ministério das Finanças (MF, 2019) apontam outros problemas como, por exemplo: a falta de diversificação do produto turístico; aumento da criminalidade e delinquência juvenil; turismo interno reduzido e relativamente caro; a fraca participação dos produtores locais na cadeia de valor do turismo; a instabilidade em alguns destinos competitivos do Norte de África; a descaraterização da identidade cultural; os preços elevados comparando com os destinos concorrentes; e o aumento da carga tributária do setor. Estes problemas não só prejudicam o destino como não reduzem a sazonalidade, sendo esta uma característica do turismo em Cabo Verde, já que a maior parte dos países emissores é da Europa viajando nos períodos mensais de outubro a março.

Segundo os economistas Swinkels e Longmore (2019), uma das limitações no atual modelo do turismo está relacionada com o baixo crescimento económico ocorrido entre os anos 2008 e 2016, onde se constatou que 85,0% a 90,0% dos turistas permaneceram em *resorts* no modelo *all inclusive* e com ligações limitadas com o resto da economia local. De recordar que Cabo Verde ficou em 127º entre 190 países no relatório do *Doing Business* (2018), isto é, o ambiente de negócios no país é relativamente fraco.

Ademais, para Swinkels e Longmore (2019), este baixo crescimento económico juntamente com os altos investimentos públicos contribuem para uma dívida pública

igual a 128,0% do PIB, sendo uma parte desta detida pelas empresas estatais ligadas a habitação social, eletricidade e companhias aéreas, representando, assim, uma ameaça para a sustentabilidade das conquistas de Cabo Verde.

Falando de sustentabilidade¹, o GOPEDS-Turismo (2019) aponta que o crescimento acelerado do turismo (abordado anteriormente), principalmente nas ilhas do **Sal** e Boa Vista, tem causado um impacto negativo a nível social e ambiental, levando a que o turismo siga a um rumo de insustentabilidade, necessitando, portanto, de medidas de mitigação para não se comprometer o património material e imaterial que suportam o destino.

1.8 METAS E PREVISÕES PARA O TURISMO EM CABO VERDE

A curto e a médio prazo, o Governo de Cabo Verde, numa visão em que se vislumbra os pilares de competitividade, sustentabilidade, desconcentração e maximização do impacto líquido positivo, pretende (GOPEDS-Turismo, 2019):

- Até 2021 chegar a 1 milhão de turistas por ano e entrar para o top 30 dos países mais competitivos e para o top 5 em África;
- Aumentar a média de receitas por turista acima dos principais destinos concorrentes de Cabo Verde;
- Fazer o alinhamento da fiscalidade com os principais destinos concorrentes e estabelecer um ambiente de negócios favorável;
- Inserir os setores da agricultura, pesca, cultura e desporto na promoção do turismo;
- Reforçar e consolidar o turismo de sol e praia, sendo este o modelo que impera no país;

¹ Segundo as Nações Unidas e a Organização Mundial do Turismo, a sustentabilidade do turismo assenta em cinco pilares, a saber: económico, social, ambiental, cultural, paz e segurança.

- Desenvolver novos segmentos como o turismo de montanha/ecológico, cruzeiro e de eventos/negócios;
- Eliminar as principais fraquezas do turismo cabo-verdiano;
- Reforçar a articulação entre o turismo, o ambiente e a segurança;
- Conquistar novos países emissores, nomeadamente a Rússia;
- Inverter a tendência demográfica da pobreza em menos de três anos.

Em paralelo o Governo traça quatro cenários para o crescimento do turismo em Cabo Verde para os próximos tempos: muito otimista, otimista, médio e péssimo. Veja-se cada um dos cenários (GOPEDS-Turismo, 2019):

Para o cenário muito otimista (Ver Tabela A1.1 do Anexo I)

- O fluxo turístico irá crescer em média 15,0% ao ano.
- Conquista de novos mercados emissores.
- Redução da sazonalidade.
- Alcançar mais de 1,25 milhões de turistas até 2021 e chegar aos 2,2 milhões em 2025.
- Superar os 5 milhões de turistas em 2030.
- Ultrapassar os 32,5 milhões de dormidas em 2030.
- Arrecadar em taxa turística mais de 65 milhões de euros.
- Ter uma contribuição do turismo para o PIB além dos 30,0%.
- Chegar até 2030 com cerca de 62,4 mil empregos diretos e 190 mil indiretos.
- Aumentar o número de postos de trabalho relacionados com o turismo em mais de 2,0% a partir de 2021 e em mais de 20,0% em 2030.

No cenário otimista (Ver Tabela A1.2 do Anexo I)

- O fluxo turístico poderá crescer em média 12,0% ao ano.
- Alcançar mais de 1,1 milhões de turistas até 2021 e chegar aos 1,77 milhões em 2025.
- Superar os 3,5 milhões de turistas em 2030.
- Chegar aos 22,5 milhões de dormidas em 2030.
- Arrecadar em taxa turística mais de 45 milhões de euros.

- Ter uma contribuição do turismo para o PIB perto dos 30,0%.
- Chegar até 2030 com cerca de 43 mil empregos diretos e 129 mil indiretos.
- Atingir o pleno emprego² em 2025.
- Aumentar o número de postos de trabalho relacionado com o turismo em mais de 12,0% em 2030.

No cenário Médio (Ver Tabela A1.3 do Anexo I)

- O fluxo turístico poderá crescer em média 9,0% ao ano.
- Alcançar um pouco mais de 1 milhão de turistas até 2021 e atingir mais de 1,4 milhões em 2025.
- Chegar perto de 2,4 milhões de turistas em 2030.
- Aproximar dos 15 milhões de dormidas em 2030.
- Arrecadar em taxa turística mais de 30 milhões de euros.
- Ter uma contribuição do turismo para o PIB perto dos 22,0%.
- Chegar até 2030 com cerca de 29.500 mil empregos diretos e 88.500 mil indiretos.
- Atingir o pleno emprego em 2025.
- Aumentar o número de postos de trabalho relacionados com o turismo em mais de 3,0% em 2030.

No cenário Péssimo (Ver Tabela A1.4 do Anexo I)

- O fluxo turístico poderá crescer em média 5,0% ao ano.
- Chegar a menos de 1 milhão de turistas até 2021 e atingir um pouco mais de 1 milhão em 2025.
- Chegar em 1,4 milhões de turistas em 2030.
- Ultrapassar os 9 milhões de dormidas em 2030.
- Arrecadar em taxa turística mais de 18,2 milhões de euros.
- Ter uma contribuição do turismo para o PIB inferior a 22,0%.
- Chegar até 2030 com cerca de 17.500 mil empregos diretos e 52.500 mil indiretos.

² No sector turístico, quando chega o período da época baixa muitos trabalhadores são dispensados (fim do contrato) dos serviços hoteleiros e os que prestam serviços diretos (por.ex. guias turísticos) ficam em dificuldades financeiras, devido a fraca movimentação de turistas. Assim, esta meta do pleno emprego é agir no sentido de que todos os que trabalham nos estabelecimentos turísticos e serviços diretos relacionados mantenham os seus empregos independentemente das épocas baixas ou altas.

- Atingir o pleno emprego em 2030.

Quanto ao Grupo de Trabalho da Câmara do Turismo de Cabo Verde, as metas e previsões estendem-se até 2030, em que se pretende (CTCV, 2019):

- Aumentar o número de entrada de turistas de 716 mil para 2,6 milhões;
- Aumentar as dormidas de turistas de 4,5 milhões para 18,4 milhões;
- Aumentar da taxa de ocupação dos empreendimentos turísticos de 58,0% para 68,0%;
- Diminuir a taxa de sazonalidade de 29,0% para 25,0%;
- Criar mais 74.000 postos de trabalho, direto e indireto, em todo país;
- Aumentar a despesa média por pax/dia de 59 para 109 euros;
- Garantir um nível de satisfação do turista em 85,0%;
- Aumentar as receitas globais de 311 milhões de euros para 2,3 bilhões;
- Aumentar a quota do sistema de transporte aéreo nacional de 3 milhões para 470 milhões de euros;
- Aumentar de 19 milhões para 79 milhões de euros as receitas do sistema aeroportuário;
- Aumentar de 7 milhões para 135 milhões de euros as receitas do sector produtivo nacional;
- Aumentar de 69 milhões para 370 milhões de euros o contributo fiscal e parafiscal.

1.9 COMPETITIVIDADE

Cabo Verde, em 2017, segundo critérios de avaliação da *Travel & Tourism Competitiveness Index* ficou atrás dos principais destinos concorrentes, nomeadamente as Ilhas Maurícias, Marrocos, Egito e República Dominicana, derivado de fraco desempenho nas questões referentes a disponibilidade de recursos naturais e culturais, saúde e saneamento, nível de segurança, tecnologias de informação e comunicação, diversificação das ofertas e serviços, e qualidade (GOPEDS-Turismo, 2019). Entretanto, pode analisar-se outros indicadores de 2016/2017 sobre o posicionamento de Cabo Verde face aos destinos

alternativos e concorrentes (Tabela 1.1) elaborados pela Câmara de Turismo de Cabo Verde (CTCV).

Muito se questiona quanto ao papel dos transportes na formação do preço do pacote de viagem. Mas não existem variações consideráveis por companhia aérea, pelo menos em 90% dos componentes de custo, em termos relativos. Como se pode verificar, o transporte aéreo apresenta um papel passivo e sua procura é essencialmente derivada e condicionada pela existência de infraestruturas e oferta turística, não assumindo uma utilidade em si. Mesmo com relação às companhias aéreas de *Low Cost*, abstraindo das circunstâncias específicas de determinadas rotas e destinos, por exemplo Cabo Verde, comparativamente às companhias aéreas convencionais, em termos de custos operacionais, a diferença é cada vez menor (CTCV, 2019).

Centrando na análise comparativa de Cabo Verde com os destinos alternativos, uma simulação é feita na Tabela 1.2, tendo como base de referência de distância a Cidade de Paris (Europa central). Resulta da simulação feita, aplicando o preço de 0,09 cêntimos de euro por milhas percorridas e 45 Euros de taxas aeroportuárias (indiscriminadamente e sem a taxa de segurança aérea (TSA)), tem-se, em média, comparativamente a Cabo Verde, as Canárias a -32,0%, Portugal (Lisboa) -60,0%, Tunísia -60,0%; Maurícias +111,0%, Seychelles +78,0% e Maldivas +87,0%. Assim, Cabo Verde se apresenta como sendo o destino mais caro.

1.10 CONSIDERAÇÕES FINAIS DO CAPÍTULO

É necessário descentralizar e diversificar o setor do turismo em Cabo Verde para reduzir o impacto nas Ilhas do **Sal** e Boa Vista, devido à concentração de turistas e garantir uma maior sustentabilidade, mantendo um maior equilíbrio entre o desenvolvimento do turismo e o meio ambiente.

Diante de outros destinos emergentes no país, é necessário consolidar a ilha do Sal como destino de ‘sol e praia’ e preparar as outras ilhas para receber as novas tendências do turismo mundial.

Tabela 1.1: Posicionamento de Cabo Verde face aos destinos alternativos e concorrentes.

Parâmetros	Cabo Verde	Canárias	Portugal	Tunísia	Ilhas Maurícias	Seychelles	Maldivas
Despesa Média Diária	59 €	139 €	102 €	53 €	110 €	145 €	166 €
Capacidade (camas)	11 397	653 000	567 772	241 000	29 139	10 350	41 534
Estadia Média (dias)	6,4	7,5	2,7	4,1	10,4	9,9	6,1
Visitantes (total)	716 775	15 700 000	23 953 800	4 500 000	1 425 000	349 861	1 390 000
Visitantes (EU)	575 986	13 400 000	17 246 736	1 500 000	420 000	223 911	646 343
Visitantes (EU)%	80%	85%	72%	33%	29%	64%	46%
Taxa Ocupação	58%	67%	47%	35%	63%	64%	62%
Crescimento	12,3%	7,9%	12,7%	32,0%	6,0%	15,0%	10,0%
Serviço Tipo: <i>All inclusive</i>	80,0%	55,0%	38,0%	45,0%	61,0%	75,0%	71,0%
Serviço Tipo: <i>Non All inclusive</i>	20,0%	45,0%	62,0%	55,0%	39,0%	25,0%	29,0%
Peso do Turismo no PIB (WTTC, 2017)	21,0%	29,6%	9,0%	6,6%	8,4%	22,0%	39,6%
Receitas do turismo (BM): 1995	25 €	3 750 €	4 826 €	1 571 €	526 €	191 €	180 €
Receitas do turismo (BM): 2016	344 €	14 530 €	23 776 €	1 458 €	1 559 €	432 €	2 991 €
Receitas do turismo (BM): Tendência (95/2016)	1286,0%	288,0%	393,0%	-7,0%	196,0%	125,0%	1559,0%

Fonte: CTCV (2019, p.11)

Tabela 1.2: Análise comparativa de Cabo Verde com os destinos alternativos.

Transporte aéreo	Cabo Verde	Canárias	Portugal	Tunísia	Ilhas Maurícias	Seychelles	Maldivas
Distância média a Paris (milhas)	5 286	3 462	1 806	1 842	11 696	9 758	10 300
Preço médio por milha (€)	0,09 €	0,09 €	0,09 €	0,09 €	0,09 €	0,09 €	0,09 €
Preço médio de impostos fixos (€)	45 €	45 €	45 €	45 €	45 €	45 €	45 €
Preço médio de bilhete por rota (€)	542 €	370 €	215 €	218 €	1 145 €	962 €	1 013 €
Comparando com Cabo Verde	1,0	0,7	0,4	0,4	2,1	1,8	1,9
		-32%	-60%	-60%	111%	78%	87%

Fonte: CTCV (2019, p.12)

A consolidação do turismo em Cabo Verde passa por diversificar os produtos turísticos, proteção do património cultural e natural, e fortalecimento da governança³ e instrumentos de gestão.

O setor privado local e as populações de cada ilha devem beneficiar dos lucros turísticos de forma a melhorar as suas condições de vida.

É preciso pensar em transformar uma procura potencial numa procura real, porque como identificado, as viagens dos europeus em 2018 se concentraram dentro da sua região, e provavelmente, por isso, a permanência média dos europeus em Cabo Verde vem diminuindo.

A ambição de Cabo Verde é diversificar o turismo e torná-lo sustentável para que as gerações de hoje e de amanhã, em todas as ilhas, sejam beneficiadas e que os turistas tenham experiências positivas no país.

Os resultados da receita obtidos com o fluxo turístico em África indicam que o turismo neste continente precisa de incentivar o aumento do gasto médio por turista/dia de permanência.

Atualmente, a Internet tem-se destacado como um dos principais meios de comunicação do turismo no mundo. Uma grande parte dos turistas tem descoberto o destino Cabo Verde por esse meio. Outros por meio de agências de viagens e operadores turísticos, e outros através de amigos. Esta última constatação demonstra que o país tem sido bem visto pelos turistas, que acabam transmitindo as suas experiências a outros indivíduos.

Dentre todos os cenários de previsão traçados pelo Governo, o mais provável tende a oscilar entre o cenário Otimista e Médio, ou seja, prevê-se um crescimento médio anual do turismo entre 9% e 12%.

³ É a forma como o Governo lida com a questão do turismo, podendo ser por leis, incentivos, etc.

CAPÍTULO 2: TURISMO – ILHA DO SAL

2.1. INTRODUÇÃO

A ilha do Sal, que possui apenas um único município (Município do Sal) e um único concelho (Concelho do Sal), é conhecida como essencialmente turística povoada de gentes, em sua maioria de outras ilhas e municípios, sendo a quinta região em termos populacionais, com cerca de 6,8% da população residente em Cabo Verde e cresce a uma taxa média anual de 4,5%. O passado da ilha remonta à sua descoberta em 1460 e desenvolvimento a partir do comércio do sal na localidade de Pedra de Lume e na Cidade de Santa Maria, passando pelo aeroporto na Cidade de Espargos e pelo porto da Vila da Palmeira.

Hoje, a ilha parece mais atrativa aos seus habitantes e visitantes do que outrora devido a um conjunto de indicadores de desenvolvimento vivenciados e documentados. Com duas cidades, uma vila e uma localidade, um número maior de agregados familiares tem vindo a beneficiar de um ambiente mais aprazível e de mais qualidade de vida que passa pela eletricidade, água, saneamento, escolaridade, emprego e habitação.

Neste capítulo é descrita a situação socioeconómica da ilha, realçadas as potencialidades da mesma como destino turístico, as previsões de crescimento, algumas medidas que poderão ser adotadas partindo de uma análise feita aos destinos com características semelhantes, o Ciclo de Vida e o Índice de Desenvolvimento Turístico da ilha do Sal como um destino turístico.

2.2. SITUAÇÃO SOCIOECONÓMICA DA ILHA

2.2.1. Caracterização Populacional

A ilha tem vindo a crescer a nível demográfico (Figura 2.1) a uma taxa de crescimento natural⁴ de 2,1% e a uma taxa de crescimento médio anual de 4,5% e atualmente conta com uma população residente de pelo menos 39.693 habitantes, sendo 53,1% do sexo masculino e 46,9% feminino (INE, 2019a), equivalente a uma densidade de 177 habitantes por km² e é a quinta região em termos populacionais, com cerca de 6,8% da população residente em Cabo Verde. Esta ilha tinha 5.505 habitantes em 1970 e, segundo INE (2013), terá 53.732 até 2030 e é assim uma das ilhas mais atrativas de Cabo Verde, que recebe imigrantes estrangeiros, mas sobretudo das outras ilhas, em especial de Santo Antão, São Vicente, São Nicolau e Santiago. A população é muito jovem, com idade média e mediana em torno de 28,4 anos e 28,0 anos, respetivamente. Cerca de 3,1% da população é idosa, sendo 2,8% do sexo masculino e 3,3% do sexo feminino. De notar que o índice de masculinidade na ilha tem vindo a aumentar e se encontra em 113,3%, ou seja, existe cerca de cento e treze homens para cada cem mulheres (INE, 2019a). Cerca de 28.619 residentes no Sal têm entre 15 e 64 anos de idade.

Segundo o inquérito Multiobjetivo Contínuo (IMC) de 2019, a sociedade salense é constituída por cerca de 12.895 agregados familiares (uma diminuição em relação a 2018) com um tamanho médio de 3,1 membros, sendo que existe em média 0,9 crianças menores de 18 anos, e 0,2 idosos nesses agregados, que na sua maioria são chefiados por homens (58,3%) com idade média de 43 anos. De realçar que a maioria dos representantes dos agregados (56,7%) encontram-se na faixa etária dos 35-64 anos de idade. O estado civil “Solteiro” é o que mais prevalece na população (34,8%) seguido do estado “União de Fato” (23,0%). Quanto à distribuição dos agregados segundo tipologia, temos que “Conjugais Nucleares” é a tipologia com maior número de agregados (25,7%) seguido de “Unipessoal” com 21,5%.

⁴ É uma palavra técnica utilizada pelo Instituto Nacional de Estatística (INE) de Cabo Verde. A taxa de crescimento natural é a diferença entre a taxa de natalidade e a taxa de mortalidade.

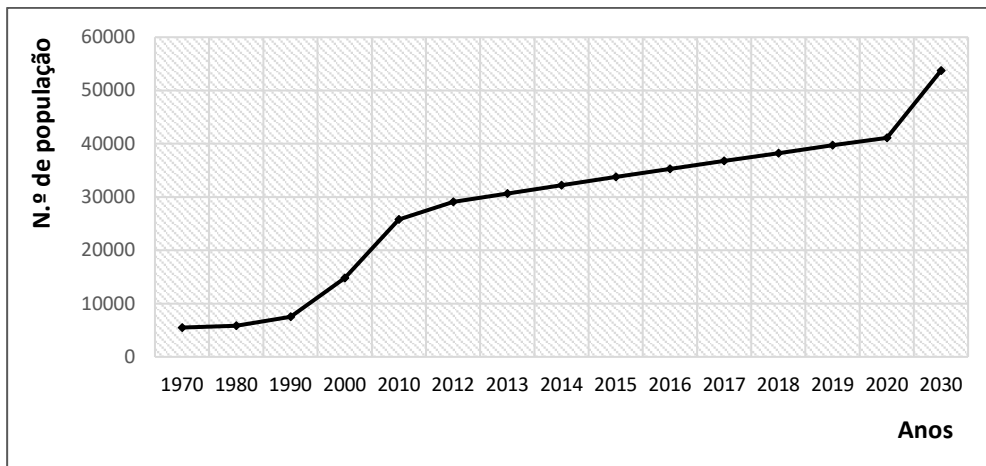


Figura 2. 1: Evolução do número da população da ilha do Sal 1970:2019 com projeção de 2020 e 2030.

Fonte: Elaboração própria, com dados do INE (2019a)

2.2.2. Nível de Pobreza

Sal é a segunda ilha de Cabo Verde que apresenta uma pobreza absoluta global (monetária) mais baixa com uma incidência de 19,1% (correspondente a 6.445 indivíduos pobres e 1.024 agregados pobres); com uma profundidade⁵ de 5,0% (1687 indivíduos e 269 agregados) e uma severidade de 1,7% (574 indivíduo e 92 agregados). Ainda relacionado com o nível de pobreza, existe uma pobreza extrema (monetária) cuja incidência ronda os 2,2% da população (correspondente a 742 indivíduos e 119 famílias muito pobres), sendo que a maioria da população muito pobre diz respeito as mulheres (58,0%), sendo que 6,0% delas tem menos de 25 anos e 73,0% delas com o nível básico de escolaridade. De realçar que o Sal concentra cerca de 3,6% (6.477 indivíduos) e 1,4% (762 indivíduos) da população pobre do país que vive em meio urbano e rural respetivamente (INE, 2017b).

Convém atentar para algumas conclusões verificadas no Perfil da Pobreza em Cabo Verde (INE, 2017b):

⁵ É uma das 3 medidas de Foster-Greer-Thorbecke. É um termo técnico que indica o montante mínimo para se erradicar a pobreza.

- Foi feita uma comparação do Sal com a Boa Vista (outra ilha mais turística do país) e verificou-se que os pobres da ilha da Boa Vista estavam ligeiramente menos afastados da linha da pobreza do que no Sal, existindo nesta muita variabilidade;
- No quadro geral do país, “quando a dimensão dos agregados passa de 3 para 4 elementos, a proporção de pobres aumenta de 14,0% para 22,0%, um salto bastante significativo”, ou seja, a pobreza e a dimensão do agregado familiar tem uma relação positiva;
- Constatou-se com alto grau de fiabilidade que viver na ilha do Sal reduz os riscos de pobreza monetária absoluta⁶. Isto por que apresenta um “*gap*”⁷ da pobreza muito baixo, igual a 5,0%;
- A distribuição dos rendimentos é assimétrica positiva, ou seja “(...) um maior número de indivíduos com rendimentos baixos do que com rendimentos altos (...)”.
- O índice de Gini com valor igual a 0,56 indica uma elevada desigualdade e excessiva concentração da riqueza na ilha. Dados de 2015 mostram uma diminuição do índice que em 2002 era de 0,56, ou seja, passados treze anos, essa desigualdade é espelhada no valor de 0,32;
- O indicador de Entropia Generalizada, de valor 0,27, mostra que o Sal é a quarta ilha onde existe maior discrepância entre os ricos e os pobres;
- Para erradicar a pobreza no município seria necessário distribuir aos pobres, um determinado valor monetário, no mínimo 5,0% do montante da linha da pobreza monetária cujo valor ronda os 95.461\$00 e 81.710\$00 no meio urbano e rural respetivamente.

⁶ Pobreza absoluta é quando os recursos de um indivíduo são insuficientes para satisfazer as necessidades básicas alimentares e não alimentares (INE, 2017b).

⁷ Esta medida usualmente é designada por “*gap*”, pois mede a insuficiência do rendimento do indivíduo em relação à linha de pobreza (INE, 2017b).

2.2.3. Migrações

Os fluxos migratórios direcionados para a ilha do Sal entraram nos últimos anos numa fase de crescimento, tanto de estrangeiros como de nacionais (de outras ilhas do arquipélago). Este fenómeno ocorre num contexto de crescimento do turismo que a ilha tem vindo a beneficiar.

Segundo dados do INE (2019a), a população imigrante em Cabo Verde é estimada em 14.347, o que corresponde a 2,6% da população residente no país. Desses, 11,2% (1.612) residiam na ilha do Sal, isto é 4,2% da população residente no município com idade média de 36,8 anos; sendo 61,8% masculino e 38,2% feminino. De salientar que o Sal é o quarto município com mais imigrantes residentes e o segundo em que o peso dos imigrantes na população é maior.

Quanto à distribuição da população imigrante segundo os principais países de proveniência, uma boa parte dos imigrantes são oriundos da Guiné Bissau (51,7%), S. Tomé e Príncipe (5,0%), Angola (4,6%), Senegal (11,7%), Portugal (6,8%) e Brasil (3,2%).

Segundo o nível de instrução, 95,9% tem algum tipo de instrução, sendo que a maioria (77,6%) possui o nível secundário. Dos 14.010 agregados familiares residentes no município em 2018, 11,5% eram agregados com pelo menos um imigrante. Destes a maioria reside em “Apartamentos” (76,1%) e 23,9% em “Moradias independentes” cujo número médio de divisões por imóvel e para dormir é de 2,1 divisões e 1,4 divisões, respetivamente, e com um número médio de membros por imóvel de cerca de 2,9.

Os agregados estão distribuídos em “Casais isolados” (23,8%), “Conjugais nucleares” (23,8%), “Monoparentais compósitos” (23,8%), “Unipessoal” (19,0%) e “Monoparentais nucleares” (9,5%).

Os que vivem sem ligação à rede pública de água rondam os 14,2% (-27,2 % com relação a 2014), acima do nível nacional. Todos têm acesso a eletricidade e instalações sanitárias, o que não ocorria em 2014 onde 9,7% residiam em alojamentos sem eletricidade e 29,6% partilhavam instalações sanitárias com outros agregados (INE, 2019f).

Em relação ao emprego, cerca de 81,7% dos imigrantes com pelo menos 15 anos de idade se encontravam empregadas em 2014 (acima do valor nacional) sendo 41,3% com inscrição no Instituto Nacional de Previdência Social (INPS), 7,6% em desemprego e 10,6% inativos. Dados de Estatísticas da Imigração mais recentes mostram que os imigrantes que trabalham no município permanecem em média 64 meses empregados, trabalhando em média 49,3 horas semanais, sendo a maioria (66,8%) inscritos no INPS (INE, 2019f). Relativamente ao nível de pobreza, a população imigrante tem uma incidência de pobreza monetária de 10,8% e dentro da população pobre do município são 8,5% (INE, 2017b).

A ilha tem um saldo migratório positivo, ou seja, a quantidade de indivíduos que entram na ilha para viverem é superior ao número dos que saem da ilha. A população salense é composta de estrangeiros (4,2%), de cabo-verdianos (94,2%) e de indivíduos com dupla nacionalidade (1,5%). Cerca de 36,5% dos residentes é natural de outro município sendo a segunda ilha com mais fluxo migratório (Figura 2.2). Dados do INE (2017b) mostraram que pouco mais de um quarto dos agregados é representado por naturais do Sal, sendo que cerca de 65,8% são nascidos em outros concelhos e 7,7% imigrantes.

De notar que dos trabalhadores que estão ao serviço dos estabelecimentos turísticos 42,0% são estrangeiros (+37,7 p.p. em relação a 2017) (INE, 2017a, 2018a).

Na questão do acesso às tecnologias de informação e comunicação no alojamento, 89,7% dos agregados possuem um televisor, 100% tem internet, 21,7% um rádio, 50,2% uma televisão por assinatura, 44,0% um computador portátil, 38,5% um Tablet e 5,8% um Desktop. A proporção de acesso a essas tecnologias é acima do valor nacional com exceção ao acesso a nível da rádio e Desktop. Todos os imigrantes possuem um telemóvel.

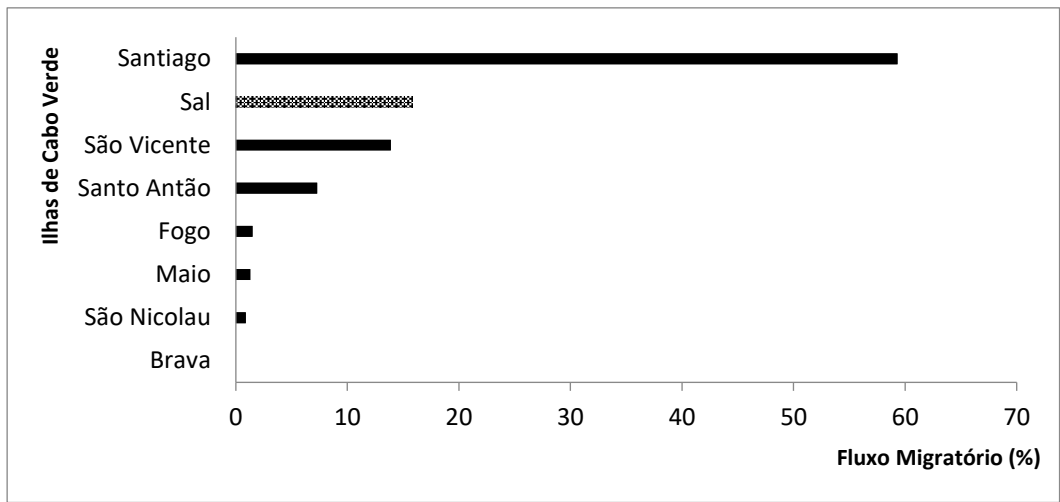


Figura 2.2: Saldo migratório entre diferentes ilhas nos últimos 5 anos: 2008 a 2013.

Fonte: Dados do INE (2015a)

Com relação à emigração internacional, cerca de 3,7% dos indivíduos que residiam na ilha do Sal saíram para o exterior (Figura 2.3). Os motivos da emigração, em sua grande parte, dizem respeito a estudos, seguido de reunião familiar; sendo Portugal o país de destino na maioria das vezes.

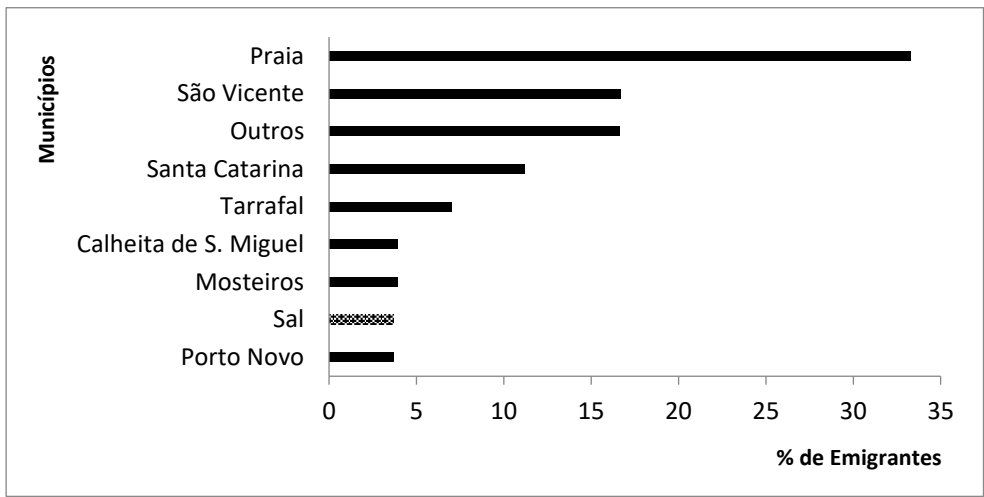


Figura 2.3: Repartição dos emigrantes comparando Sal com outros municípios no período 2009:2014.

Fonte: Dados do INE (2015a)

2.2.4. Água e Saneamento

Quanto à condição de vida da população e dos agregados em relação à água, tendo por base os dados publicados pelo Inquérito Multiobjectivo Contínuo (IMC), cerca de 85,4% (+11,9 p.p. comparando com o ano de 2017) dos agregados e 80,0% (+5,8 p.p.) da população tem acesso a água no alojamento através da rede pública de distribuição (Figura 2.4). Das famílias com água canalizada, 81,9% (+13,3 p.p.) tem a ligação de água feita no interior do alojamento e 3,6% no exterior. Quanto à fonte de abastecimento 12,5% (-8,0 p.p.) dos agregados e 17,4% (3,4 p.p.) da população ainda usam os chafarizes públicos.

Cerca de 98,2% da população (+1,2 p.p. em relação a 2017) tem acesso a água através da ligação direta da água canalizada, de chafariz ou acesso através da casa de vizinhos. Quanto ao tratamento da água, 47,0% (+13,5 p.p.) dos agregados não tem hábito de tratar a água, 2,1% trata raras vezes e 11,7% regularmente. Os que têm o hábito de tratar a água, fazem-no utilizando lixívia (59,0%) ou então fervendo (33,3%) ou filtrando (7,7%). De notar que 39,1% (-2,0 p.p.) dos agregados têm o hábito de consumir água engarrafada (INE, 2019a).

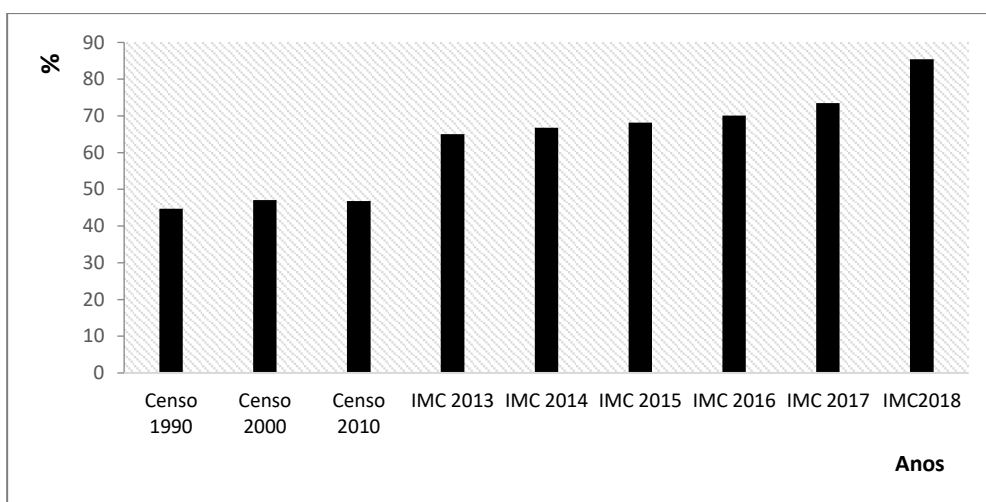


Figura 2.4: Evolução da percentagem de Agregados Familiares com ligação a rede pública de água.

Fonte: Elaboração própria, com dados do INE (2019a)

Quanto ao sistema de evacuação de águas residuais, 97,2% (+5,3 p.p.) dos agregados (92,9% da população) tem acesso a um tipo de sistema de evacuação, podendo ser a rede pública de esgoto, a fossa séptica ou outro não identificado (Figura 2.5). Cerca de 58,4% (-29,4 p.p.) dos agregados (91,9% da população) tem a fossa séptica como meio de evacuar as águas residuais e apenas 4,4% (e 1,0% da população) faz a evacuação pela rede pública. Mas 7,8% dos agregados e da população na ilha não tem nenhum tipo de sistema de evacuação das águas residuais. Com relação a instalações sanitárias, apenas 2,8% dos agregados estão sem instalações sanitárias (uma diminuição de 4,0p.p em relação a 2017). De sublinhar que 12,5% dos agregados partilham entre si as instalações sanitárias (INE, 2019a).

Quanto aos resíduos sólidos (lixos domésticos), cerca de 96,1% (+2,4 p.p.) dos agregados e 95,3% da população (um aumento de 2,5 p.p. em relação a 2017) depositam o lixo em contentores, 3,2% (e 3,5% da população) em carros de lixo; 0,4% dos agregados (0,6% da população) os jogam na natureza (INE, 2019a).

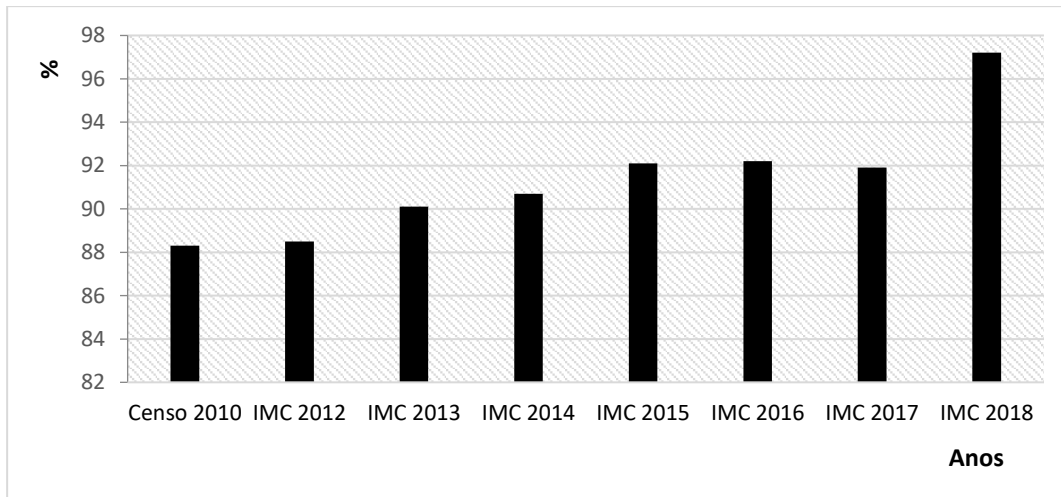


Figura 2.5: Evolução da percentagem dos Agregados Familiares com acesso ao sistema de evacuação de águas residuais (rede de esgoto/fossa séptica).

Fonte: Elaboração própria com dados do INE (2019a)

Quanto à energia utilizada para cozinhar, 96,8% (+2,0 p.p.) dos agregados usam o gás para cozerem seus alimentos e 1,1% ainda cozinham em Lenha. A maioria dos agregados (85,4%) tem cozinhas dentro do alojamento e 3,6% não possuem nenhum lugar digno para cozinhar. De salientar que 2,1% dos agregados (uma redução de 2,3 % em relação ao ano 2017) e 1,0% da população não utilizam nenhuma fonte de energia para prepararem seus alimentos (INE, 2019a).

Dados de 2018 mostram que a quantidade de água dessalinizada produzida na ilha foi de 1.566.955 m³ (+9,2% em relação a 2017), sendo que 6.181 m³ foram utilizados para consumo interno. De notar que a quantidade de água entregue à rede foi de 1.560.774 m³. O preço médio praticado na venda de água ronda os 337.27 escudos/m³. A ilha já possui duas unidades dessalinizadoras com uma capacidade de 10.000 m³/dia (Electra, 2017, 2018).

2.2.5. Energia e Eletricidade

Quanto ao acesso à energia elétrica a ilha está acima da média nacional com 94,1% da população e 95,0% dos agregados com acesso (Figura 2.6). A maioria do acesso da parte da população, é oriunda da rede pública (97,6%), sendo as restantes oriundas de gerador/motor (1,8%) e via painéis solares (0,6%). Dos 14.010 agregados, 98,5% tem acesso a energia elétrica através da rede pública, 1,1% de gerador/motor e apenas 0,4% através de painéis solares (INE, 2019a).

A principal forma de iluminação que os agregados utilizam nos lares é a eletricidade (94,1%) oriunda da rede pública e 5,4% utilizam velas. A produção térmica de energia elétrica ronda os 46,8KWh, enquanto que a produção renovável está em 21,0KWh, num total de produção de 67,8 GWh e uma taxa de penetração de energia renovável de 30,9%, este acima da taxa nacional. A ilha possui 1 parque solar, 1 central térmica e ainda adquire energia elétrica a 1 produtor privado local (Cabeólica).

Dados de 2017 mostram que a potência instalada no Sal é de 19.106 KW, um aumento de 3.456 KW em relação a 2016. A eletricidade produzida na Ilha ronda os 67.821.192 MWh (um aumento de 10,5% em relação a 2017), a segunda maior produção da empresa pública Electra Norte, sendo 73,0% foi produzida nas duas centrais da Electra e 27,0%

adquirida ao produtor Cabeólica. O preço médio praticado na venda de eletricidade ronda os 25,36 escudos cabo-verdianos (ECV)/KWh.

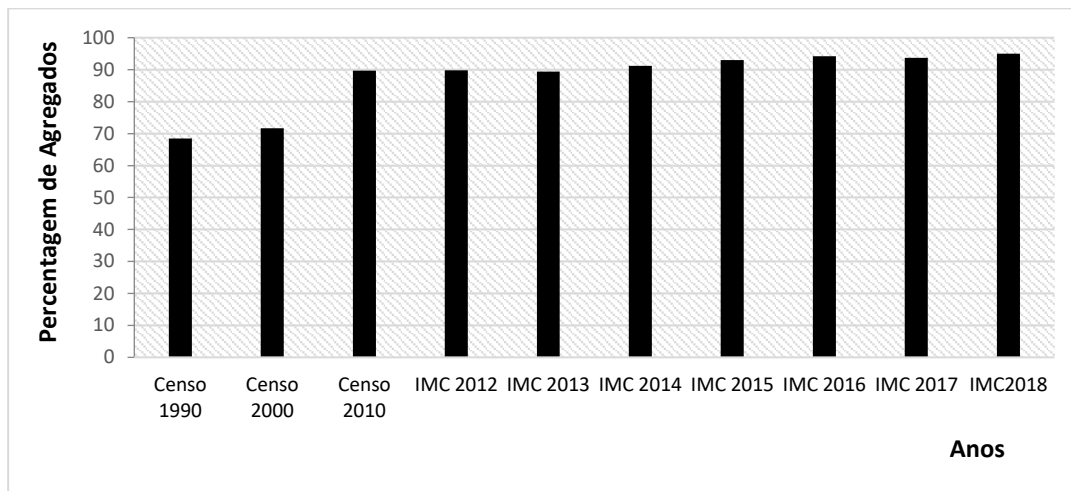


Figura 2.6: Evolução da percentagem de Acesso à Eletricidade dos Agregados Familiares.

Fonte: Elaboração própria com dados do INE (2019a)

A ponta máxima, por unidade de produção, é de 12,2KW, um aumento de 4,0% em relação a 2017. O número de apagões gerais (*Black-Outs*) amentou de 11 para 26 em 2018 com duração mínima de 644 minutos. O consumo de combustível na produção de eletricidade foi de 231 gr/KWh (uma diminuição de 1,9% em relação a 2017) acima da média do consumo total da Electra Norte. Na central térmica foi consumido cerca de 46.908 litros de óleo lubrificante na produção de eletricidade (-0,4 gr/KWh em relação a 2017). Tendo sido fornecidos à rede de distribuição cerca de 150.854.854 MWh. Quanto à rede de distribuição de energia existem 77 postos de transformação do tipo alvenaria, 145 km de rede subterrânea de média tensão e 151 km de rede aérea de baixa tensão (INE, 2016; 2018b; Electra, 2017; 2018).

2.2.6. Habitação

A habitação é um dos maiores problemas sociais nesta ilha, onde a população mais do que se duplicou desde 2000, sendo a maioria migrante (56,9%). O défice habitacional no

Sal (quantidade de habitações inexistentes ou inadequadas sem condições de habitabilidade) é crítico e ronda os 20,2%, correspondendo a 1.666 agregados familiares e a 5.689 indivíduos, sendo a terceira ilha com maior déficit. A oferta não está à altura da procura, gerando bairros clandestinos na maior ilha turística do país. Até 2030 há necessidade de haver um acréscimo de 11 mil novas habitações na ilha. De notar que os preços praticados na venda dos terrenos é de 3 mil ECV/m² (em Santa Maria), 1.9 a 2.8 mil ECV/m² (em Espargos); e com relação a materiais básicos temos, por exemplo: 1 saco de cimento (50kg) por 811ECV, 1 bloco de cimento por 65 ECV, areia (1 m³) por 4.500 ECV, cascalho (1 m³) por 1.500 ECV, uma porta por 34.000 ECV, uma janela por 24.000 ECV.

Segundo o levantamento socioeconómico realizado em 2017, pelo MIOth e com apoio da Câmara Municipal do Sal (CMS), cerca de 655 agregados familiares vivem em casas de latas e barracas em assentamentos não legalizadas (os chamados de Bairros Informais), sendo que 35,0% desses com tamanho médio de 3 e 4 membros. Em termos da população, 2.030 indivíduos (55,0% do sexo masculino e 45,0% do sexo feminino) vivem nos bairros informais (Alto São João, Alto Santa Cruz, Terra Boa, Lixeira de Morrinho Branco, arredores de Murdeira e arredores de Santa Maria). Desses 2.030 indivíduos, 36,6% são crianças menores de 16 anos de idade e 0,9% são idosos. Cerca de 2,0% dos agregados que vivem nestes bairros informais ganham menos de 5.000 ECV ao mês (CMS, 2017a) (MIOth, 2019).

Das 14.010 famílias salenses, contabilizadas no IMC de 2018, a maioria habitava em alojamentos clássicos (99,3%), principalmente do tipo Moradia Independente (54,4%). Com relação aos alojamentos não clássicos (que podem ser barracas, contentores e improvisados em edifícios não destinados a habitação), cerca de 0,7% dos agregados familiares viviam nesses alojamentos (+ 0,2 p.p.), valor, semelhante ao nível nacional (INE, 2019a).

Um indicador que merece ser destacado é a proporção da população urbana que vive em casas degradadas ou precárias. Os dados indicam que houve uma melhoria neste indicador e que em 2014, apenas 29,3% (INE, 2015b) da população vivia em casas

degradadas ou precárias (uma redução de 29,4% desde 2010). Entretanto estima-se que até 2020 sejam requalificadas cerca de 229 casas de famílias carenciadas (MIOTH, 2019).

2.2.7. Saúde

O contexto sanitário do Sal caracteriza-se por uma população em crescimento acelerado, o maior e mais concorrido aeroporto do país, chegadas e partidas, em 2018, de vários voos nacionais (3.452) e internacionais (9.978), em média 10.000 turistas/diários, dificuldades de evacuação (n.º de doentes evacuados = 19 em 2018, correspondendo a 4% das evacuações para fora do país), com 1 hospital regional, 1 centro de saúde, 1 delegacia e 1 unidade sanitária de base, num total de 41 camas. Para além das estruturas de saúde pública, a ilha do Sal conta ainda com várias **estruturas de saúde privadas**, designadamente 9 clínicas (saúde à família, laboratórios, fisioterapia), 3 clínicas dentárias e 5 ópticas, sendo que a razão de médicos e de enfermeiros por 10 mil habitantes é de 8.72 e 13.35 respetivamente.

Com relação à venda de medicamentos existem 4 farmácias: 1 pública e 3 privadas. Desde o ano de 2014 e até 2018 o número de partos assistidos por profissionais de saúde vem aumentando, de 617 para 747 partos por ano. Das 842 grávidas registadas na ilha, em 2018, todas receberam a cobertura vacinal antitetânica.

A taxa de mortalidade da população é de 3,1% (um aumento de 0,2 p.p. em relação a 2017), sendo que a maioria dos óbitos (54,7%) ocorrem na camada masculina. De salientar que em relação ao ano de 2017 houve um aumento de 14 óbitos em 2018. O número de óbitos infantis é de 14, sendo 8 deles de neonatal precoce, 1 de neonatal tardio e os restantes pós-neonatal, que ocorreram no Hospital Regional do Sal. Das causas de óbitos ocorridos no Sal, a mais frequente é doenças do aparelho circulatório (22,2% dos óbitos) seguido de doenças do aparelho respiratório, tumores ou neoplasias, e doenças infecciosas e parasitárias. Convém salientar o fato de existir na ilha o serviço de teleconsultas, sendo as especialidades mais procuradas a Neurologia, Cardiologia e Ortopedia.

Dados de 2018 mostram que existe um total de 9 casos registados de tuberculose na ilha (-6 casos em relação a 2017), cuja taxa de incidência e de prevalência por cem mil

habitantes é de 24%. Quanto à infeção por vírus da imunodeficiência humana (VIH), dados referentes a 2018 mostram 36 casos novos notificados (correspondendo a 7,2% do total de novos casos no país que é de 503), sendo 19 do sexo masculino e 17 do sexo feminino. Entretanto o número de indivíduos que vivem com o VIH e que fazem o seguimento junto dos serviços de saúde é de 104. Lembrando que dados do estudo sobre seroprevalência do VIH nos usuários de drogas em 2016 mostraram que 6,8% dos 2.320 usuários de drogas no país estão concentrados no Sal.

No que toca à prevalência contraceptiva, a taxa é de 38,8% (um aumento de 7,7% em relação a 2016). Sendo o Sal um dos três principais centros urbanos do país, cerca de 10,0% dos 1.655 trabalhadores do sexo a nível nacional estão concentrados na ilha. Considerando trabalhadores e profissionais do sexo que usam drogas em Cabo Verde, 75,0% são do Sal (INE, 2016; MF, 2017; MSSS, 2018a; MSSS, 2018b).

2.2.8. Escolaridade

Sal é uma das raras ilhas onde a procura escolar cresce por causa da imigração, com 24 estabelecimentos escolares (13 do pré-escolar, 8 do básico e 3 do secundário) num universo de 302 professores, e um Centro de Recursos Integrado de Educação e Formação (CRIEF) que alberga Pólos Universitários (Uni-CV, Uni-Mindelo e Lusófona) e realiza formações profissionais. A taxa líquida⁸ de escolarização no ensino básico e secundário ronda os 90,6% e 68,9% e uma taxa de abandono e reprovação no ensino secundário de 6,5% e 16,2%, respetivamente, com maior incidência no sexo masculino.

A educação desenvolve-se em todos os níveis: do total (1.579) de crianças de 0-3 anos, apenas 24,2% estão na creche; no pré-escolar com 1.005 crianças, em que 13,0% se encontra fora dos jardins (CMS, 2018a); na educação e formação de adultos com cerca de 123 formandos; no ensino básico com 3.677 estudantes; no ensino secundário com 1.374 estudantes na via geral e 1.807 na via técnica, estando em curso a reorganização da rede escolar; 242 na formação profissional e 136 no ensino superior que se iniciou em 2016.

⁸ É a razão entre o número de matrículas de alunas que estão na faixa etária para o ensino básico e secundário e a população total na mesma faixa etária.

Nesta ilha a taxa de alfabetização na população de 15 ou mais anos tem vindo a aumentar (Figura 2.7) e ronda os 96,8% em 2018 (+ 0,8 p.p. em relação a 2017), sendo que no sexo masculino é de 98,9%, no sexo feminino de 94,3% e na camada juvenil atinge os 100% (+ 1,4 p.p. em relação a 2017), em ambos os sexos. A taxa de alfabetização juvenil (15 a 24 anos) é de 99,2%. Da população que é detentor com mais de quatro anos de estudo, a maioria 47,1% tem o nível secundário e 5,5% o superior. Em geral o número médio de estudos no município aproxima-se dos oito anos. Da população masculina 97,7% está alfabetizada e da população feminina 92,2% está alfabetizada.

Apenas 3,0% da população nunca frequentou o ensino escolar, 51,5% dos salenses possuem o nível secundário e apenas 9,0% têm nível superior. A nível de anos de estudo os salenses têm em média 9,5 anos de estudos. Como é sabido no processo da educação o hábito de leitura tem o seu papel, pois permite desenvolver, entre vários aspetos, o vocabulário e a escrita. Na ilha cerca de 64,6% da população tem hábito de leitura, acima da média nacional (INE, 2016; 2019a).

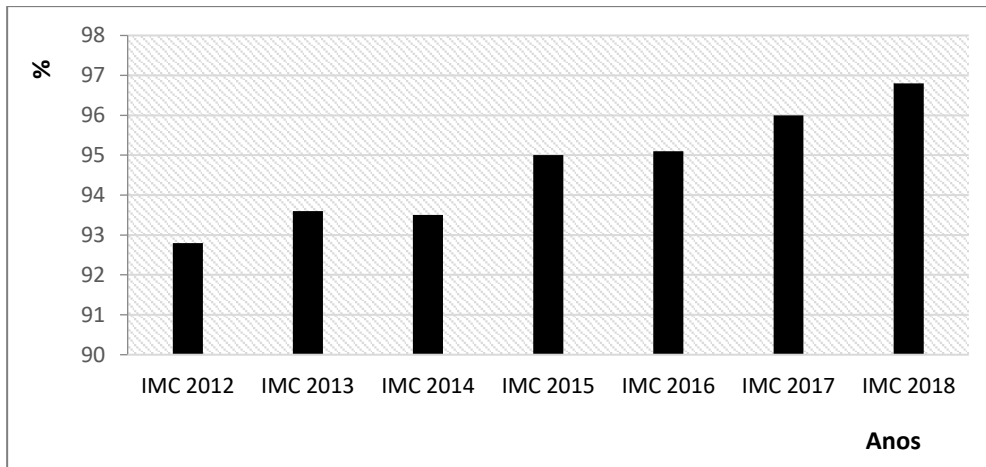


Figura 2.7: Evolução da Taxa de Alfabetização da População com 15 anos ou mais.

Fonte: Elaboração própria com dados do INE (2019a)

2.2.9. Emprego

A força de trabalho do Sal é constituída por cerca de 26.142 ativos, com a segunda maior taxa de atividade a nível nacional, de 65,3% (-5,8 p.p. em relação a 2017), sendo 73,6% do sexo masculino e 68,1% do sexo feminino, e a taxa de desemprego é de apenas 7,4% (-0,9 p.p. em relação a 2016), sendo 8,8% do sexo masculino e 5,5% do sexo feminino (INE, 2016). Quanto à taxa de emprego, a ilha conta com 62,6% e uma taxa de subemprego de 2,8%, é assim uma ilha de oportunidades. Entretanto na camada juvenil (15-24 anos) a taxa de desemprego ronda os 19,8% (uma redução de 3,0 p.p. em relação a 2016) (INE, 2018c; INE, 2019g).

Os trabalhadores que estão ao serviço dos estabelecimentos turísticos são 5.068, em que 55,0% são Cabo-Verdianos (+70,7 p.p. em relação a 2017) e 42,0% estrangeiros (+37,7 p.p. em relação a 2017), 56,0% são do sexo masculino e 52,0% do sexo feminino (em relação a 2017 o número de mulheres diminuíram em 5,4 p.p.). Do pessoal ao serviço 77,0% estão em aldeamentos turísticos, 58,0% estão em hotéis, 43,0% em hotéis-apartamentos e 12,0% em pensões, sendo que 45,0% possui contrato permanente (um aumento de 25,4 p.p em relação a 2017), 59,0% com contrato a termo (uma diminuição de 20,5 p.p. em relação a 2017) e 7,0% sem contrato (+6,1 p.p. em relação a 2017).

Dos trabalhadores com contrato a termo, 58,0% tem contrato de 3 meses, 57,0% tem contrato de 6 meses e 62,0% tem contrato de pelo menos 1 ano. De ressaltar que do rol de pessoal empregado nos estabelecimentos turísticos em Cabo Verde, 54 em cada 100 empregados estão na ilha do Sal. A nível do número de quartos e camas, Sal é a que possui maior expressão (45,7% e 44,7%, respetivamente) a nível nacional (INE, 2019c).

2.2.10. Situação Financeira

Em 2015 a ilha do Sal era a terceira (8,8%) no total das despesas do consumo realizado a nível nacional, cuja despesa média e mediana anual por pessoa (216.853 ECV e 165.509 ECV, respetivamente) estava acima da média do país; sendo a maior parte das despesas realizadas pelos homens. A maior parte das despesas estão concentradas em dois grupos, ou seja, despesas com a habitação-água-eletricidade, seguido das despesas com a

alimentação e bebidas não alcoólicas, representando o maior peso orçamental (31,8% e 25,5%, respetivamente).

A ilha do Sal é a terceira ilha onde as despesas com o primeiro grupo (que inclui a habitação) são maiores. Em conjunto com as ilhas de Boa Vista, Brava e São Nicolau, o município viu o peso relativo deste no total das despesas aumentar em 4,4 p.p. (12,2% em 2001 para 16,5% em 2015) (INE, 2017b). A ilha gera um PIB/per capita anual de 591.369 ECV (689.219 em 2016), o mais elevado de Cabo Verde, onde em média cada indivíduo gasta anualmente 216.853 ECV para o consumo de bens alimentícios e não alimentícios e cada agregado familiar gasta 826.767 ECV (INE, 2017b), acima da média nacional.

Com uma representatividade de 31,6%, a maior parte dos agregados familiares apresenta um rendimento médio mensal no escalão 20 000 a 30 000 ECV, e que 60,0% (524) dos agregados encontram-se entre 20 000 e 50 000 ECV. No entanto, o custo de vida na Ilha do Sal é sobejamente sabido ser um dos mais caros do país com elevados preços dos arrendamentos, serviços e todos os tipos de produtos.

2.2.11. Economia e Turismo

A economia do Sal é marcadamente turística com um Produto Interno Bruto de 21.744 mil contos (-2.564 em relação a 2016) com predominância do setor terciário que gera 2/3 da riqueza da ilha, representando 12,6% do PIB Nacional e resultando um PIB/per capita de 591.369 ECV, o mais elevado de Cabo Verde. Transporte e armazenagem é a principal atividade económica com 23,6% da riqueza da ilha, determinado principalmente pelo aeroporto que é o maior do país. Segundo o MF (2019), o setor de alojamento e restauração participou, em 2018, com 16% do PIB nacional; e com 44% do PIB desse setor.

O *boom* no turismo de Cabo Verde iniciou-se pelo Sal, a partir do ano de 2002, que continua sendo a maior ilha turística do país, com 30 estabelecimentos turísticos, 10,6% do total existente no país (sendo 15 hotéis, 7 pensões, 4 hotéis-apartamentos, 3 aldeamentos turísticos e 1 residencial) e uma capacidade de alojamento de 13.451 hóspedes e teve 379.020 hóspedes (49,5%) e 2.764.099 dormidas (56,0% de dormidas

nos estabelecimentos turísticos no país) em 2018, (Figura 2.8 e 2.9), um aumento, em dormidas, de 244.612 em relação a 2017. Dos visitantes que chegam à ilha do Sal como destino turístico, os do Reino Unido são os mais frequentes (101.302) seguidos dos Países Baixos (41.867) e Alemanha (36.233) (INE, 2017a; 2019c; 2019d;2019g).

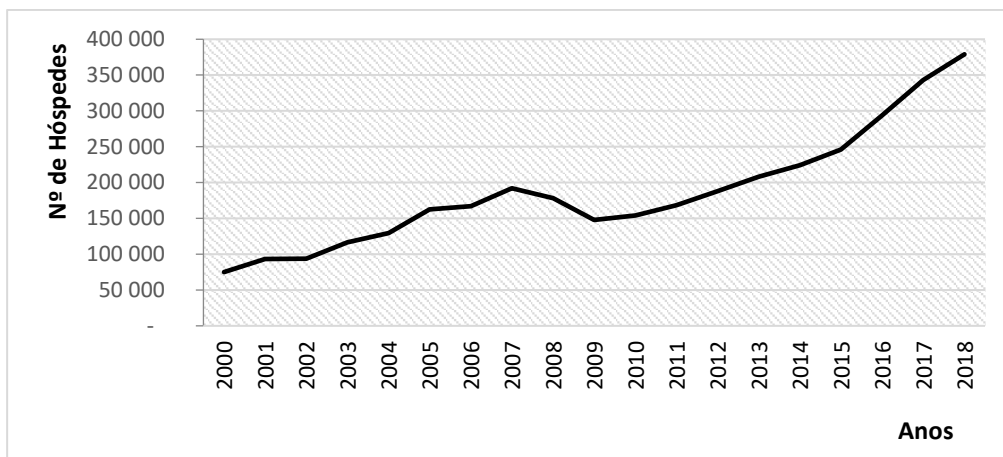


Figura 2.8: Evolução do número de Hóspedes nos Estabelecimentos Turísticos.

Fonte: Elaboração própria, com dados do INE (2019d)

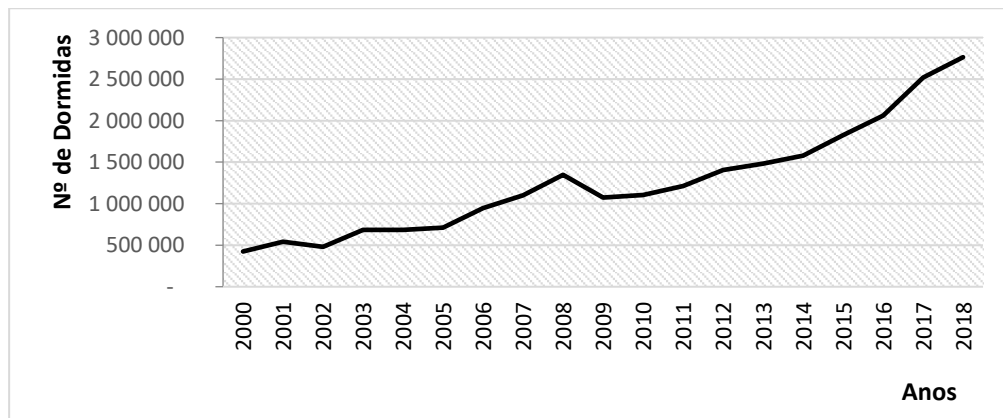


Figura 2.9: Evolução do número de Dormidas nos Estabelecimentos turísticos.

Fonte: Elaboração própria, com dados do INE (2019d)

No período entre 2000 e 2018, os investimentos em infraestruturas desportivas, aliadas às infraestruturas hoteleiras e às excelentes condições naturais da Ilha do Sal, contribuíram para a realização de vários eventos importantes a nível internacional.

Alguns dos eventos que mobilizaram visitantes nacionais e estrangeiros foram: *Kite-Surf World Tour* (2018: fevereiro); finais dos campeonatos nacionais em 2018 (Andebol - maio/junho; Basquetebol - julho; Karaté - setembro; Atletismo - junho/julho; Boxe - agosto; Xadrez - março).

Além disso, a abertura de novos hotéis de renome internacional também levou a um aumento do número de turistas, como o Hilton Hotel (2017: outubro; 2018: janeiro), Riu Funaná (2006: outubro) renovado para Club-Hotel Riu Funaná (2016), Riu Garopa (2011) renovado para Riu Palace Cabo Verde (2016: março), Meliá Lana (2016: dezembro), Meliá Tortuga (2011: maio), Meliá Dunas (2014: novembro), Tui Sensimar (2016: dezembro). Paralelamente, foram realizados alguns eventos culturais, como o Mundo da Literatura (2017 e 2018) e o Festival de Santa Maria (2000 a 2018), que também dinamizaram o movimento e aumentaram o número de visitantes na ilha (Neves, Nunes & Fernandes, 2021; CMS, 2017b, 2018b).

Esta ilha realiza cerca de 40,2% do Valor Acrescentado Bruto (VAB) do alojamento e restauração, e conjuntamente com a ilha da Boa Vista geram cerca de $\frac{3}{4}$ do VAB desse ramo de atividade económica. Sal é a terceira ilha com maior peso da atividade empresarial, sendo que na ilha está sediada a ASA, uma das maiores empresas cabo-verdianas, e os maiores operadores turísticos do país. Das 1.231 empresas na ilha (+241 em relação a 2013), 1.158 estão ativas (INE, 2019h), sendo que a quantidade de indivíduos ao serviço é em número de 12.360 (+64,6% em relação a 2012) gerando um volume de negócios de 56.744.329 mil contos cabo-verdianos (+45,8% face a 2012) equivalente a 19,7% da faturação das empresas nacionais.

Das empresas ativas na ilha, 574 têm contabilidade organizada e 585 não têm contabilidade organizada. As empresas com contabilidade organizada empregam 10.785 indivíduos e têm um volume de negócios de 55.617.839 mil contos cabo-verdianos (INE, 2019h). No ramo da atividade piscatória, a ilha conta com 170 botes num total de 482

pescadores, sendo que o número médio de pescadores por bote é de 2,8, e 7 embarcações num total de 63 pescadores.

2.2.12. Transporte e Passageiros

Durante o ano de 2018, registou-se no aeroporto da ilha cerca de 13.429 movimentos de aeronaves (+950 movimentos em relação a 2017), sendo 25,7% domésticos (- 18,5 p.p.) e 74,3% internacionais (+21,1 p.p.). As movimentações na ilha correspondem a 39,4% das movimentações ocorridas no país. Quanto ao número de passageiros registaram-se 1.184.636, representando um acréscimo de 8,4% em relação a 2017. Desses passageiros, 84,5% é de natureza internacional (uma variação de 9,9% em relação a 2017).

De notar que o tráfego doméstico de passageiros foi comercializado, em 2018, apenas pela companhia aérea *Binter Cabo Verde*. Uma boa parte do tráfego internacional esteve relacionado com o crescimento do destino Lisboa com mais de 52.716 passageiros transportados. Destaque ainda para os destinos Estocolmo e Gotemburgo que registaram mais 9.439 e 15.921 passageiros, respetivamente, em relação a 2017. A nova rota a partir de Copenhaga-Dinamarca que aumentou para 6.592 o número de passageiros. Quanto a cargas movimentadas no aeroporto do Sal houve cerca de 429.339 (- 30,5 p.p. em relação a 2017). E a movimentação de correios ronda os 70.747 (-30,9 p.p. em relação a 2017).

De destacar que os principais operadores internacionais têm sido: a Thomson Fly; a TAP Air Portugal; a Cabo Verde AirLines; a Tui Fly Netherlands e a Tui Fly GMBH, sendo as suas quotas de mercado de 24,1%, 10,2%, 9,5%, 8,7% e 8,3%, respetivamente. Entretanto a Thomson Fly é a operadora que movimenta um maior número de turistas para a ilha. A principal origem/mercado emissor que chega ao aeroporto do Sal é Lisboa-Portugal com um total de 161.777 passageiros, seguido de Gatwick-Inglaterra com 82.168 (ASA, 2018).

2.3. POTENCIALIDADES TURÍSTICAS DA ILHA DO SAL

As extensas praias de areia branca, associadas às condições climáticas favoráveis (sol, vento, correntes, marés e forte ondulações) são recursos turísticos naturais que poderão constituir-se num autêntico potencial para à prática do turismo balnear e dos desportos náuticos. O seu clima tropical, com temperatura amena durante o ano, permite desenvolver o turismo praticamente ao longo de todo o ano.

Com o potencial turístico que possui, a ilha do Sal pode desenvolver diferentes tipos de turismo, desde o turismo de ‘sol e praia’ (mergulho, windsurf, surfe, kitesurfe, kite beach, shark bay, pesca e outros; turismo de natureza (caminhadas, passeios de bicicleta, moto-quatro, passeio à volta a ilha), turismo cultural (Museu do Sal, artesanato, vila da Palmeira, localidade de Pedra de Lume, gastronomia) e turismo de saúde (massagem/spa, banho de lama, banho de sal, serviços de saúde especializados).

A grande extensão da costa, dotada de excelentes praias de areia branca, constitui o seu principal potencial. A ilha dispõe de uma das maiores extensões de praia e encontra-se entre as mais belas do país. A praia de Santa Maria com 3km de areias brancas foi nomeada pela TripAdvisor® como uma das 25 praias mais bonitas do mundo e uma das sete maravilhas de Cabo Verde. As salinas de Pedra de Lume, as salinas de Santa Maria, Buracona (piscina natural conhecida como Olho d’água), Regona, Baía da Murdeira, Baía da Parda (habitat de tubarões de até 2 metros de comprimento), Costa da Fragata (espaço natural protegido), Monte Leão (rocha que assemelha uma face de leão), Monte Grande (ponto mais alto da ilha) são outros atrativos turísticos naturais existentes na ilha (MF, 2019).

De destacar que a ilha tem sido palco de realizações de Jogos Africanos de Praia e Etapas do Circuito Mundial de Kitesurf. O Festival Internacional de Santa Maria e o Festival de Literatura Mundo têm atraído turistas de várias partes do mundo; e as romarias religiosas impulsionam o fluxo de indivíduos entre ilhas tornando a ilha a mais turística do arquipélago.

As potencialidades turísticas desta ilha, umas existentes e outras a serem exploradas, não se resumem aos recursos naturais e culturais. As infraestruturas hoteleiras e conexas (por

exemplo, infraestruturas rodoviárias e aeroportuárias) também fazem parte das suas potencialidades. A ilha dispõe atualmente de um número considerável de infraestruturas de alojamento de padrão internacional.

Estando mais perto da Europa, um dos principais mercados do turismo emissor, e dispondo de um aeroporto de nível internacional, aliada às potencialidades turísticas naturais, dispõe de condições que à partida determinaram o desenvolvimento do turismo nesta ilha.

Sendo uma ilha de pequena dimensão, com cerca de 216km² de superfície, plana e árida, está entre as ilhas do arquipélago com maiores potencialidades para a prática do turismo balnear. Trata-se de uma ilha vulcânica, com alguns cones vulcânicos, onde sobressai a cratera das salinas de Pedra de Lume. O Monte Grande representa a sua maior elevação, com cerca de 406m de altitude e pequenas bacias hidrográficas na parte central. A cratera das salinas de ‘Pedra de Lume’ onde se fazia a exploração do sal para exportação poderá ser explorada para fins turísticos. A água desta localidade é muito utilizada para fins terapêuticos e estima-se que diariamente passam pelas salinas mais de 500 turistas (CMS, 2021). Há ainda, nas salinas, uma capela do século passado, velhos barracões, as instalações de transporte por cabo, um sistema de moagem e ensacamento do sal que se apresenta sem qualquer alteração há mais de cinquenta anos e que serve ainda de atração turística. Exemplo espetacular de Arqueologia Industrial, Pedra de Lume é uma memória viva, património cultural que se tornou uma das sete maravilhas de Cabo Verde.

A ilha está sob a influência dos ventos alísios do leste provenientes do deserto do Sahara. A força desses ventos poderá ser utilizada para à produção da energia eólica, para além de constituírem um grande potencial para a prática de desportos náuticos como, o Windsurf. Em suma, o turismo de sol e praia durante o ano, os desportos náuticos, pesca, natação, festival de música, são entre as principais potencialidades turísticas da ilha. Pearce (1987) já tinha considerado a predominância do turismo de ‘Sol e Praia’, especialmente em ilhas tropicais e subtropicais, uma consequência direta da situação insular.

No que se refere à acessibilidade, os principais aglomerados populacionais da ilha estão ligados por uma rede de estradas em bom estado de conservação. A estrada de Santa Maria-Aeroporto é considerada a via principal de ligação ao maior centro turístico da ilha. O aeroporto internacional Amílcar Cabral é uma outra potencialidade da ilha, pois impulsiona o turismo internacional e permite a distribuição do fluxo turístico para o território regional. Uma outra importante infraestrutura de apoio ao turismo é a construção de um hospital na Cidade de Santa Maria.

2.4. PROBLEMAS DO TURISMO NO SAL

Todas essas potencialidades turísticas da ilha do Sal, umas existentes e outras em exploração, e outras a serem exploradas no futuro, têm levado a que, em termos de carga turística, exista uma grande pressão sobre o território com uma média de 3 turistas para cada 1 habitante e a uma densidade populacional, somado a população residente com a turística, de 200 hab./Km² (segundo o Estudo de Avaliação da Dinâmica Social e Ambiental do Sector Turístico de Cabo Verde elaborado pelo MF (2018)) Como consequência, existe uma enorme carga sobre alguns serviços, nomeadamente eletricidade, gás, vapor e ar condicionado; captação, tratamento e distribuição de água.

A identidade cultural de Cabo Verde não é devidamente representada nos produtos de artesanato e experiências comercializados na ilha do Sal. Em relação ao ambiente existe um certo descuido quanto à proteção dos atrativos e recursos naturais, bem como o desordenamento urbano nas proximidades da orla da praia de Santa Maria (MF, 2019).

2.5. PREVISÃO DO CRESCIMENTO DO TURISMO

Em relação à previsão do crescimento do turismo para os próximos anos, apenas foi contemplado o nível de hospedagem. Constata-se, no *Master Plan* do Turismo 2020-2030 para a Ilha do Sal, encomendado pelo MF (2019), cenários com ganhos otimistas, com ganhos moderados e com poucos ganhos em competitividade e sustentabilidade. Pode ver-se cada um dos cenários na Tabela 2.1. Estes cenários foram construídos com base

numa técnica denominada de **Análise Lógica Intuitiva** cuja ponte forte é a sua facilidade em desenvolver cenários flexíveis e consistentes sob uma perspetiva intuitiva, assim como referido no próprio *Master Plan*, MF (2019):

O método escolhido para a definição de cenários foi o da Análise Lógica Intuitiva. Esta metodologia não utiliza algoritmo matemático. Ela considera primordialmente que as decisões sejam fundamentadas em um conjunto de inter-relações e interdependências adimensionais envolvendo variáveis sociais, políticas, tecnológicas, ambientais e econômicas. Além disso, é um método intuitivo porque estimula a intuição dos gestores a respeito das incertezas e possibilidades futuras. O cenário do turismo na ilha para os próximos anos depende diretamente de um conjunto de variáveis condicionantes que indicam as direções prováveis que este setor poderá seguir. Assim, foram definidas 8 variáveis de análise, dentro de duas condicionantes: condicionantes externas: (i) Desempenho da Economia Mundial; (ii) Desempenho da Economia Nacional; condicionantes internos: (iii) Desempenho do Turismo Mundial; (iv) Desempenho do Turismo Nacional; (v) Acesso e Logística; (vi) Governança e Investimentos Públicos no Turismo; (vii) Investimentos Privados no Setor de Turismo; (viii) Estrutura de Mercado.

(MF, 2019, pp.118,119)

De sublinhar que as variáveis analisadas e os cenários construídos, com base nas condicionantes externas e internas, foram feitos antes da pandemia da Covid 19. A técnica de Análise Lógica Intuitiva não será descrita neste trabalho por não fazer parte dos objetivos do mesmo.

Tabela 2.1: Cenários de crescimento do número de hóspedes para o período 2019:2030.

Ano	Cenários		
	Ganhos otimistas (12% ao ano)	Ganhos moderados (9% ao ano)	Poucos ganhos (5% ao ano)
2019	424.576	413.204	398.040
2020	475.525	450.392	417.942
2021	532.589	490.927	438.839
2022	596.499	535.111	460.781
2023	668.079	583.271	483.820
2024	748.249	635.765	508.011
2025	838.038	692.984	533.412
2026	938.603	755.353	560.083
2027	1.051.235	823.334	588.087
2028	1.177.384	897.434	617.491
2029	1.318.670	978.204	648.366
2030	1.476.910	1.066.242	680.784

Fonte: MF (2019)

2.6. CICLO DE VIDA DO TURISMO NA ILHA DO SAL

Assim como qualquer ser vivo que passa pelas fases de nascimento, desenvolvimento e morte, também as empresas têm o seu ciclo de vida que segundo Greiner (1998) são a criatividade, direção, delegação, coordenação e colaboração; e identicamente o turismo tem o seu próprio ciclo de vida, com relação ao destino turístico, sendo que as fases do ciclo do destino vão desde a sua exploração até o seu declínio ou rejuvenescimento.

2.6.1. O Modelo

O modelo de ciclo de vida de um destino turístico mais utilizado para descrever o desenvolvimento turístico ao longo do tempo, segundo Hovinen (2002) é o de Butler (1980) cujas fases do ciclo são (Figura 2.10):

1. **Exploração** – nesta fase surgem raras vezes nos destinos, grupos pequenos de turistas que acomodam em instalações destinadas à própria população. Os impactos sociais, económicos e ambientais não sofrem qualquer alteração.
2. **Envolvimento** – esta fase se caracteriza pelo aumento regular de visitantes ao destino e pela perceção da comunidade como oportunidade de negócio. A comunidade começa a perceber os benefícios e a criar acomodações exclusivas para os visitantes. A aproximação da população e dos visitantes torna-se mais formal e alguns sectores da comunidade começam a ser melhorados. Inicia-se um processo de publicidade e marketing sobre o destino, bem como alguma organização.
3. **Desenvolvimento** – sendo esta a fase de pré-consolidação, a promoção do destino é mais intensificada, o número de turistas tende a ser igual ou superior ao da população, e a comunidade começa a perceber alguma mudança e impactos sociais, económicos e ambientais, tanto positivos como negativos. Mais infraestruturas turísticas são criadas, bem como a exploração de atrativos culturais e naturais. Aumenta o investimento de estrangeiros na comunidade.

4. **Consolidação** – o turismo passa a ser o maior responsável pelo desenvolvimento económico do destino. A promoção do destino ganha outras dimensões e objetiva-se aumentar a estadia média dos turistas, diversificar e conquistar novos mercados.
5. **Estagnação** – nesta fase o destino já tem atingido ou excedido sua capacidade máxima de turistas. Depara-se com problemas ambientais, sociais e económicos. As promoções do destino empreendidas em outras fases, tornaram-no bem-sucedido e as peculiaridades que o caracterizavam deixam de ter o atrativo de antes, e começa-se a reinventar atratividades não naturais e até exportadas.
6. Chegando ao final da fase de estagnação duas coisas pode acontecer:
 - 6.1. **Declínio** – nesta fase de estagnação o destino começa a seguir rumo ao ponto de partida, a fase exploratória, pois já não tem capacidade de competir com outros destinos, o número de turistas diminuirá drasticamente e as visitas tornam-se cada vez mais raras. As infraestruturas turísticas perdem sua utilidade e a comunidade tende a adquiri-las a baixos custos.
 - 6.2. **Rejuvenescência** – nesta fase de estagnação, o destino aposta na exploração de outros atrativos naturais e culturais até então pouco ou nada aproveitados. É uma fase de se reinventar com o que existe de original no destino

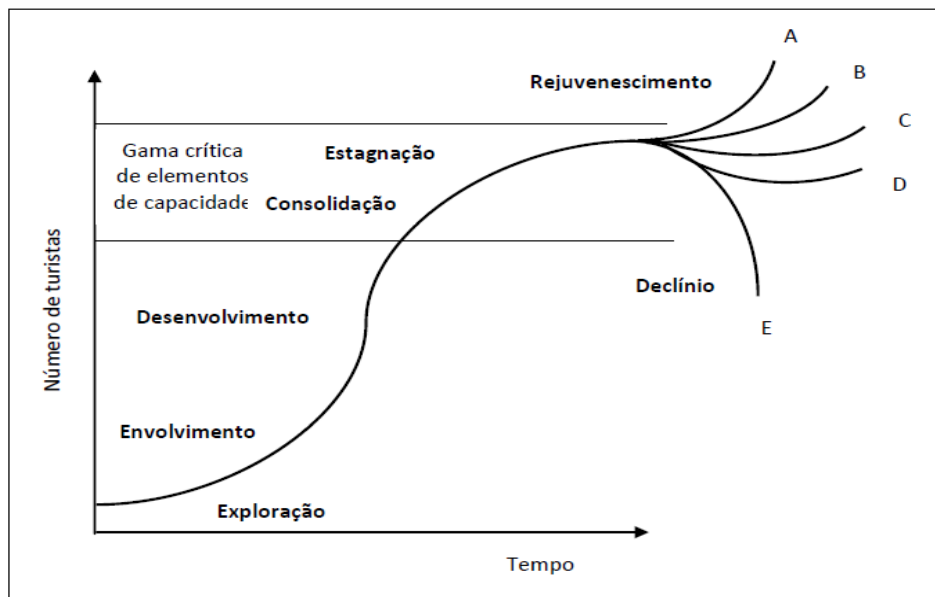


Figura 2.10: Evolução do Ciclo de Vida de um destino turístico.

Fonte: Adaptado de Butler (1980, p.7)

Segundo o MF (2019) a ilha do Sal como destino turístico, e enquadrado no segmento de 'sol e praia', encontra-se na terceira fase do seu ciclo de vida, ou seja numa fase de desenvolvimento (Figura 2.11), pelas seguintes razões características dessa fase: em épocas altas, o número de turistas é superior ao número da população; já se verifica uma pressão sobre o meio ambiente e em infraestruturas básicas e urbanas; investidores estrangeiros estão presentes na ilha, atuando em empreendimentos hoteleiros e em serviços de operadores turísticos; a quantidade de ofertas está a aumentar e o produto começa a ser diversificado; há um elevado número de empregos derivados e relacionados com o turismo na ilha e melhores condições de vida da população.



Figura 2.11: Fase do Ciclo de Vida do Destino Turístico Ilha do Sal.

Fonte: MF (2019, p. 112)

2.7. ÍNDICE DE DESENVOLVIMENTO TURÍSTICO

Com o modelo do Ciclo de Vida Turístico os decisores dependem do conhecimento da fase do ciclo de vida do destino para determinar suas opções estratégicas (Butler, 1980). Entretanto este modelo, originalmente denominado por modelo TALC – *Tourism Area Life Cycle*, traz algumas limitações identificadas por La Torre, Naranjo e Cárdenas (2012), por exemplo com relação aos indicadores que determinam quando a mudança de fase deve ocorrer, já que muitas vezes se verificam elementos diversos em diferentes fases do ciclo, ou seja não restritivos a apenas uma fase.

Para tentar resolver essas limitações, Coelho e Butler (2012) propuseram uma ferramenta denominada IDT – Índice de Desenvolvimento Turístico, que permitisse de forma determinística quantificar as fases do Ciclo de Vida de um destino e cuja preocupação central evidenciada no índice é fazer sobressair um certo nível de bem-estar da população local como resultado da atividade turística. Este índice é calculado em função do número da população local (P), do número de turistas que visitam o destino (T) e do número de condições turísticas criadas no destino (D), e é dado por:

$$IDT_{jt} = \sum_{i=1}^n \frac{1}{n} \left(\frac{D_{ijt}}{P_{jt}} \right) \times \left(\frac{T_{jt}}{T_{Tt}} \right) \quad (1)$$

ou

$$IDT_{jt} = \sum_{i=1}^n \frac{1}{n} \left(\frac{D_{ijt}}{D_{iTt}} \right) \times \left(\frac{P_{Tt}}{P_{jt}} \right) \times \left(\frac{T_{jt}}{T_{Tt}} \right) \quad (2)$$

onde, D_{ijt} – são os valores das variáveis (i) relativamente às condições turísticas de atrações e fixação do destino (j) em um determinado período de tempo (t), P_{jt} – é população residente do destino (j) em um determinado período de tempo (t), T_{jt} – número de turistas do destino (j) em um determinado período de tempo (t), D_{iTt} – é a soma dos valores das variáveis (i) relativamente às condições turísticas de atrações e fixação do destino (j) em um determinado período de tempo (t), P_{Tt} – soma da população de todos os destinos (j) em um determinado período de tempo (t), T_{Tt} – soma do número de turistas em todos os destinos (j) em um determinado período de tempo (t), n – número de variáveis ambientais, económicas e socioculturais utilizadas no destino para o cálculo do IDT.

Utilizando como exemplo o destino 1 ($j = 1$) em um determinado período de tempo ($t = 0$) ele demonstra a expressão (1) e consequentemente a expressão (2) da seguinte forma:

$$IDT_{1t=0} = \frac{1}{n} \left(\frac{D_{11}/P_1}{\sum D_{1j0}/\sum P_{j0}} + \frac{D_{21}/P_1}{\sum D_{2j0}/\sum P_{j0}} + \dots + \frac{D_{n1}/P_1}{\sum D_{nj0}/\sum P_{j0}} \right) \times \left(\frac{T_1}{\sum T_{j0}} \right) \quad (3)$$

reescrevendo a expressão (3),

$$IDT_{1t=0} = \frac{1}{n} \left(\frac{D_{11} \times \sum P_{j0}}{\sum D_{1j0} \times P_1} + \frac{D_{21} \times \sum P_{j0}}{\sum D_{2j0} \times P_1} + \dots + \frac{D_{n1} \times \sum P_{j0}}{\sum D_{nj0} \times P_1} \right) \times \left(\frac{T_1}{\sum T_{j0}} \right) \quad (4a)$$

$$IDT_{1t=0} = \frac{1}{n} \left(\frac{D_{11}}{\sum D_{1j0}} + \frac{D_{21}}{\sum D_{2j0}} + \dots + \frac{D_{n1}}{\sum D_{nj0}} \right) \times \left(\frac{\sum P_{j0}}{P_1} \right) \times \left(\frac{T_1}{\sum T_{j0}} \right) \quad (4b)$$

deste modo resulta na expressão (2).

As variáveis relacionadas às condições turísticas de atração e fixação são apresentadas na Figura 2.12. Para um destino de ‘sol e praia’ (como é o caso da Ilha do Sal), podem ser

adicionadas variáveis como quantidade/qualidade das praias e outras, desde que possuam uma relação direta com o bem-estar (Coelho & Lourenço, 2015).

Essa proposta levou a que fosse quantificado o Ciclo de Vida para um contexto competitivo, como mostrado na Tabela 2.3.

Tabela 2.2: Matriz de cálculo do IDT.

	Condições Turísticas				População Residente	Turistas
	Variavel 1	Variavel 2	...	Variavel n		
Destino 1	D_{11}	D_{21}	...	D_{n1}	P_1	T_1
Destino 2	D_{12}	D_{22}	...	D_{n2}	P_2	T_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Destino w	D_{1w}	D_{2w}	...	D_{nw}	P_w	T_w
	$\sum D_{1jt}$	$\sum D_{2jt}$...	$\sum D_{njt}$	$\sum P_{jt}$	$\sum T_{jt}$

Nota: w, destinos; n, variáveis do destino.

Fonte: Adaptado de Coelho (2010, p. 259)

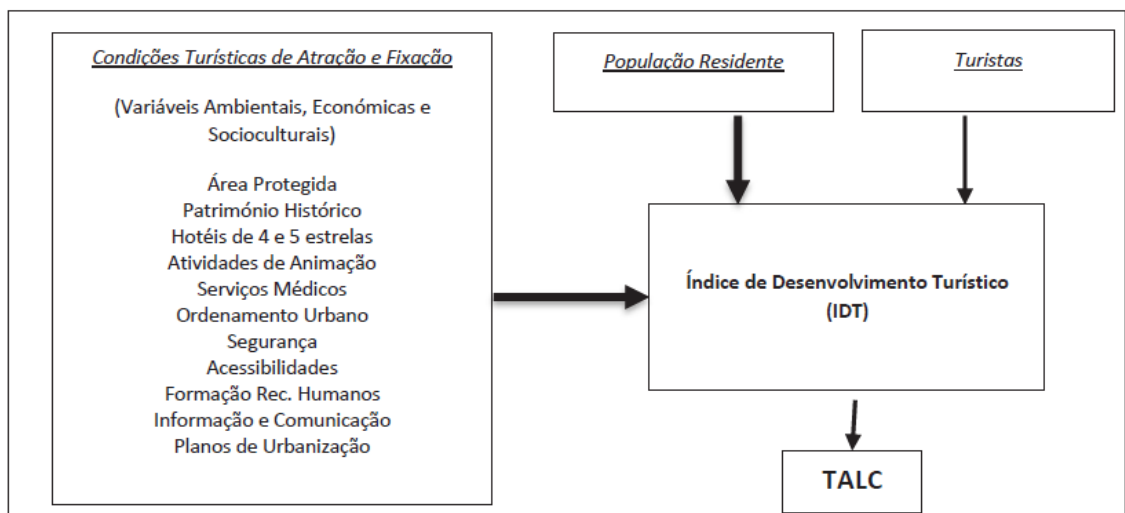


Figura 2.12: Construto Original do IDT.

Fonte: Coelho e Lourenço (2015, p. 24)

Tabela 2.3: Quantificação do Ciclo de Vida.

Fases do Ciclo de Vida	Valores do IDT	Ângulo da reta tangente no ponto de mudança de fases
Exploração	$0 < IDT < 0,194$	$0^\circ < \alpha < 11^\circ$
Envolvimento	$0,194 \leq IDT < 0,577$	$11^\circ \leq \alpha < 30^\circ$
Desenvolvimento	$0,577 \leq IDT < 1,401$	$30^\circ \leq \alpha < 90^\circ$
Consolidação	$1,401 < IDT < 3,019$	$11^\circ < \alpha < 45^\circ$
Estagnação	$IDT > 3,019$	$0^\circ \leq \alpha < 11^\circ$
Declínio	$IDT_{(t)} > IDT_{(t+1)}$	$-90^\circ < \alpha < 0^\circ$

Fonte: Coelho e Butler (2012, p.22)

Devido à diversidade das variáveis que constituem o fator *Condições Turísticas de Atração e Fixação*, à dificuldade na obtenção dos dados em tempo útil e por ser entendido como bens e serviços disponíveis em uma determinada região, Coelho e Lourenço (2015) propuseram a substituição do conjunto de variáveis relativo a esse fator pela variável nominal PIB. O PIB representa um conjunto de bens e serviços produzidos e ou disponível em uma determinada região num determinado período de tempo, não considerando as transferências recebidas e nem expedidas para o exterior.

Para a realização dessa substituição Coelho e Lourenço (2015) definiram alguns pressupostos (A) e hipóteses (H):

A₁ - quanto maior é o valor do PIB, mais desenvolvido tende a ser a região;

A₂ - quanto maior a distribuição *per capita* da receita turística melhor é o nível de bem-estar da população residente;

A₃ - quanto maior a distribuição *per capita* do PIB melhor é o nível de bem-estar da população residente;

H₁ - o nível do PIB influencia o nível de receitas turísticas;

H₂ - não depende do número da população residente;

H₃ - não depende do número de turistas, quando depurados da influência do turismo respetivamente;

H₄ - o número de população residente não depende do número de turistas.

Se todas as hipóteses forem verificadas e diante dessa nova abordagem, o Índice de Desenvolvimento Turístico é dado pela expressão (5), sendo que a variável PIB agora é considerada depurada das receitas turísticas (PIB^*) e é adicionada uma nova variável – a Receita Turística (RT):

$$IDT_{jt} = \frac{1}{2} \left[\left(\frac{PIB_{jt}^*}{PIB_{Tt}^*} \right) + \left(\frac{RT_{jt}}{RT_{Tt}} \right) \right] \times \left(\frac{P_{Tt}}{P_{jt}} \right) \times \left(\frac{T_{jt}}{T_{Tt}} \right) \quad (5)$$

E o novo construto do Índice de Desenvolvimento Turístico que permite melhorar o modelo de Ciclo de Vida de um destino turístico (TALC) é apresentado na Figura 2.13.

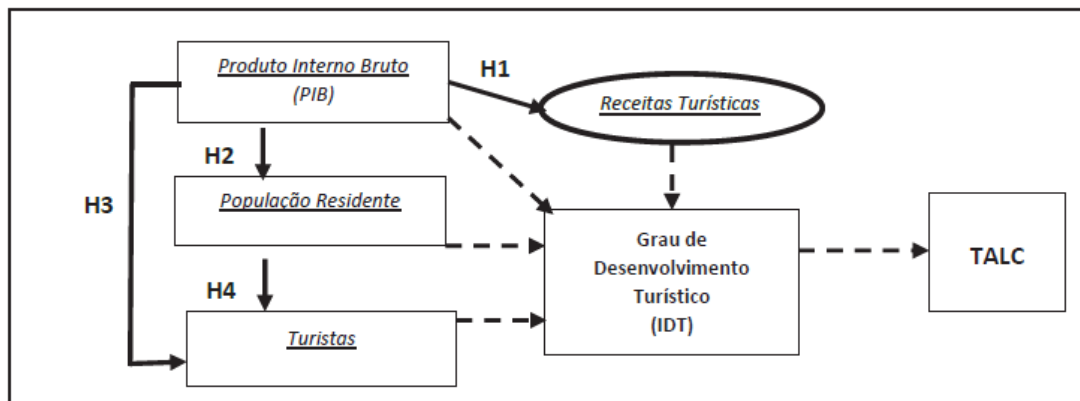


Figura 2.13: Construto do IDT (revisado).

Fonte: Coelho e Lourenço (2015, p. 30)

2.8. BENCHMARKING

Existem alguns destinos turísticos cujas características se assemelham ao destino ilha do Sal. Aprender com esses destinos poderá ser uma mais-valia para que se possa consolidar a ilha do Sal como destino relativamente ao segmento ‘sol e praia’. O *Master Plan* de Turismo para a Ilha do Sal (MF, 2019) analisa quatro destinos: Canárias, Punta Cana, Costa Rica e Fernando de Noronha, buscando identificar, por exemplo, estratégias de promoção, produtos e atividades turísticas. As ilhas Maurícias e Seychelles também são

apresentadas por De Brito (2007) como destinos a qual a ilha do Sal poderá aprender com suas experiências.

O que se pode aprender com as ilhas Canárias?

As ilhas Canárias possuem uma relativa semelhança com Cabo Verde (por serem ilhas de origem vulcânica, com especificidades únicas), porém é um destino consolidado.

Diversificam o seu produto turístico e proporcionam diferentes experiências e atividades relacionadas aos segmentos de ‘sol e praia’, turismo náutico, ecoturismo, geoturismo, turismo gastronómico e cultural.

O mercado nessa região gera um fluxo turístico que atinge em média 12,5 milhões de turistas ao ano, sendo que a maior procura acontece quando chega o inverno na Europa. O crescimento acelerado tende a caracterizar-se como destino de massa em algumas ilhas e os impactos negativos advenientes têm a ver com a perda de biodiversidade, degradação dos habitats costeiros e marinhos, destruição de sítios arqueológicos e grande procura por água e energia.

O governo das Canárias tem vindo a incentivar os turistas a saírem dos hotéis e contribuir mais no desenvolvimento da economia local.

As ilhas têm vindo a manifestar preocupações com o ambiente buscando modelos mais sustentáveis. Quanto a isto, estão em busca por uma economia de baixo carbono, pois possuem excelentes condições para a produção de energia eólica, solar e térmica. Os privados estão a ter iniciativas nesse sentido, promovendo o uso de energia solar nos seus estabelecimentos, sensibilizando os hóspedes quanto ao consumo de energia, tudo isto tendo em vista um turismo mais inteligente, sustentável e inclusivo.

O que se pode aprender com Punta Cana?

Punta Cana é um dos destinos turísticos que mais tem crescido no Caribe e dentre suas ofertas e atividades turísticas estão o *snorkeling*, passeio de barco, vela, *windsurf*, caiaque, tirolesa, trilhos e passeios a cavalo. O turismo de massa é o tipo mais praticado com grande oferta de *resorts all inclusive*. Dado ao rápido desenvolvimento, fatores

negativos, por exemplo, como a destruição de habitats naturais, poluição, contaminação por poluentes *offshore* e *onshore*, pesca excessiva causando danos nos recifes de corais, praias, flora e fauna marinha, têm tido impactos significativos. Para mitigar esses efeitos negativos, Punta Cana aposta na capacitação, voluntarismo, educação ambiental e diversificação da oferta turística. Os privados são também chamados a encontrar um modelo de turismo sustentável. Desenvolveu-se em Punta Cana um centro de sustentabilidade e biodiversidade para buscar soluções sustentáveis de forma a equilibrar a relação entre o turismo, o meio ambiente e a inclusão da comunidade local. Projetos relacionados com a pesca, a agricultura sustentável, a conservação de tartaruga e de reservas de corais, iniciativas de apoio a educação, capacitação, cultura e desporto têm sido desenvolvidos.

O que se pode aprender com Costa Rica?

Costa Rica é um destino baseado na natureza e é um dos países mais ricos em biodiversidade do mundo. Em algumas regiões o turismo é caracterizado como sendo de 'sol e praia'. Sua estratégia no turismo é enfatizada na sua sustentabilidade. Nesta linha dotou a medida de remunerar, por meio de incentivos governamentais e renda turística, os produtores rurais para manterem-se a vegetação natural e recuperar as áreas degradadas.

Também criou um programa de certificação para a sustentabilidade turística onde são avaliadas e diferenciadas diversas empresas (entre sustentáveis e convencionais) com relação ao consumo de água e energia, utilização de produtos verdes, manejo de resíduos, nível de emissão de poluentes, manutenção de áreas verdes, proteção de fauna e flora, benefícios económicos para a comunidade, valorização da cultura local e garantia de saúde e emprego para a população.

Essas medidas trouxeram resultados satisfatórios e contribuíram para o aumento de 79,0% na cobertura vegetal do território, fazendo do país um exemplo no tocante ao turismo ecológico.

Costa Rica oferece aos turistas aventura, sol e praia, ecoturismo, cultura, turismo rural e bem-estar. Ainda estão a desenvolver outros nichos como famílias, lua-de-mel, reuniões e convenções, observação de aves e cruzeiros. Entre as atividades que podem ser praticadas estão o ciclismo, caminhadas, *rafting*, desportos náuticos, pesca desportiva, acampamento, visita a sítios arqueológicos e históricos, escaladas, golfe, quadriciclo, observação de aves e animais selvagens.

O que se pode aprender com Fernando de Noronha?

Fernando de Noronha é um arquipélago com 21 ilhas de origem vulcânica e nele se refugiam diversas espécies ameaçadas de extinção. Albergando um dos santuários ecológicos mais importantes do planeta, medidas são tomadas para balancear os impactos do turismo com a preservação biológica local. Para tal, é feito um controle das visitas e pagamento de taxas ambientais. Estes ajudam na preservação do património ambiental e histórico, contenção dos impactos ao ecossistema, manutenção dos estudos científicos e biológicos, educação e sensibilização sobre o turismo ecológico.

O arquipélago é um destino que carrega um alto valor ao pacote da viagem como um todo, mas oferece diversos atrativos e atividades, nomeadamente: banho em piscinas naturais, banho de praia, trilhos, observação de golfinhos, aves, desova de tartarugas, mergulhos, passeios de barco, visitas a sítios em ruínas e históricos. Tudo devidamente agendado com antecedência.

O que se pode aprender com as ilhas Maurícias?

A República das Maurícias é um pequeno arquipélago subsaariano, sem recursos naturais que saiu de uma economia de baixa renda voltada para a cultura da cana do açúcar para uma economia diversificada orientada para a exportação. Para combater o desemprego, o governo desenvolveu sua estratégia orientada para as exportações. Com isso abriu-se as portas à economia, encorajando o assentamento de empresas de manufatura e conectando ao mercado europeu. Além das indústrias têxteis que fizeram das Maurícias

uma ponte entre África e Ásia e entre Europa e América, também emerge o turismo como fator de desenvolvimento económico.

Com os complementos dos serviços financeiros e tecnologias de informação e comunicação, a economia do arquipélago crescia em média 4,9% ao ano entre 1991 e 2010, possuindo um PIB per capita de 13.000 USD. Outras apostas foram feitas a nível da indústria biomédica, biotecnologia, energias renováveis, economia do mar, indústria cultural e do conhecimento e turismo de saúde. O turismo representa apenas 7,0% do PIB, sendo que 70,0% do PIB tem a ver com a indústria de manufatura. Isto porque o governo tem como objetivo criar oportunidades para o povo, os empresários e investidores, sendo o sector público/privado um dos pilares da governança das Ilhas Maurícias.

Para o desenvolvimento das Maurícias, o governo teve que aprender com experiências de, por exemplo, Taiwan e Hong Kong no tocante a exploração de uma economia exportadora (com isso foram estabelecidas zonas económicas exclusivas - ZEE); com Dubai e Irlanda no tocante a criação de um *hub* de comércio (com isso criou-se os *freeports* dentro das ZEE); com Singapura no tocante a registro de empresas em 24h.

Os Mauricianos são um povo descendente de chineses, europeus e indianos e o governo promove a cultura das Ilhas Maurícias como um todo, abre a economia ao mundo, cria oportunidades para todos, aprende com outros e garante a proteção do ambiente e da população (Expressos das Ilhas, 2018).

O que se pode aprender com as Seycheles?

A República de Seychelles, está localizada no Oceano Índico ocidental e é constituída por várias ilhas a norte e nordeste de Madagáscar, com cerca de 87 mil habitantes e um PIB per capita de 25.900 USD. As suas belas paisagens, ambientes naturais intatos e garantida por lei a preservação da fauna e da flora, fazem dela um líder mundial no desenvolvimento de um turismo sustentável. O turismo detém cerca de 30,0% da empregabilidade e contribui em mais de 70,0% das receitas no país. Considerado como

um destino de qualidade, devido a políticas de conservação da biodiversidade e do ambiente que trouxe ganhos tanto para o turismo como para a população das ilhas.

Um dos grandes interesses do governo é que a população tome parte no desenvolvimento do turismo, encorajando-a para a abertura de pequenos negócios no sector do turismo e pequenos hotéis. O Estado criou algumas taxas como a taxa de marketing do turismo e a taxa corporativa de responsabilidade social, cuja receitas são investidas, por exemplo, no turismo, na educação e na saúde. Com condições de segurança criadas em todas as ilhas, os turistas são incentivados a saírem dos hotéis e a consumirem os serviços prestados pela população local. O turismo de massa e de *charters* não é bem visto pelos empresários por não trazer benefícios à população. De salientar que o sector privado emprega cerca de 70,0% da população (Expresso das Ilhas, 2014).

2.9. MEDIDAS A ADOTAR

Do *Benchmarking* analisado, a ilha do Sal pode recorrer a algumas medidas adotadas por esses destinos similares com vista a melhorar a qualidade do turismo, torná-lo mais sustentável e mais competitivo. São elas:

- ✓ Iniciativas e incentivos do Governo que priorizam parâmetros de sustentabilidade, incluindo projetos de certificação.
- ✓ Apostar no ecoturismo como meio de financiar as áreas protegidas.
- ✓ Incentivar a criação de projetos visando a mitigação dos impactos negativos gerados pelo desenvolvimento turístico.
- ✓ Integrar a população local nos projetos de desenvolvimento turístico.
- ✓ Apostar na regulamentação e ordenamento das atividades turísticas.
- ✓ Estabelecer planos de gestão e de redução de resíduos.
- ✓ Desenvolver sistemas e plataformas *online* para a promoção e gestão do destino turístico.
- ✓ Valorizar o património natural.
- ✓ Limitar a quantidade de visitas aos atrativos turísticos da ilha.

- ✓ Aplicar cobranças de taxas ambientais a serem revertidas a preservação, conservação, pesquisa e manutenção do quesito ambiental.
- ✓ Criar planos específicos que norteiam o desenvolvimento e sustentabilidade do destino.

2.10. CONSIDERAÇÕES FINAIS DO CAPÍTULO

O crescimento previsto para o turismo na ilha do Sal levanta a questão de se atentar para os diversos impactos negativos decorrentes deste tipo de fluxo, procurando aprimorar a qualidade da experiência do turista e a valorização/proteção dos ativos culturais e naturais locais.

O fenómeno imigração tem impacto tanto na distribuição espacial como na estrutura da população da ilha, pelo que reflexos negativos tanto a nível individual como social (questão da legalização, precaridade de postos de trabalho, da habitação, da segurança social, da educação, etc.) podem surgir caso medidas de políticas de integração harmoniosas com a realidade local deixarem de ser equacionadas ou forem insuficientes.

A ilha do Sal é um destino turístico atrativo, mas com desafios significativos para aumentar a sua competitividade e garantir a sustentabilidade, como melhorias na estruturação de seus recursos e produtos turísticos, melhorias da infraestrutura turística, efetiva conservação das áreas protegidas, valorização da cultura local, entre outros. Alguma mobilização deve ser feita pelo governo central, mas outras devem ser lideradas pelo governo local envolvendo o setor privado e a população local. Entretanto, várias medidas têm vindo a ser tomadas para melhorar o destino, como por exemplo a reestruturação dos serviços de saneamento, criação de espaços verdes, a aderência ao projeto Bandeira Azul, entre outros, sendo que tais medidas carecem de eficiência.

O desenvolvimento do turismo na ilha tem tido impactos positivos e negativos na perceção da população, e esta sente a necessidade de ser envolvida no processo e poder contribuir junto dos hoteleiros para um turismo sustentável e com benefícios económicos, sociais e culturais para todos.

Os destinos com características semelhantes às da ilha do Sal apresentam o mesmo problema de um desenvolvimento não sustentável derivado do rápido crescimento do turismo. A ilha do Sal tende a ir no mesmo caminho, pelo que se pode aprender com esses destinos e reverter a situação enquanto decorre o processo de consolidação do destino.

Considerando o segmento turístico chave – ‘sol e praia’, e os produtos oferecidos no destino, entende-se que a ilha do Sal se encontra no estágio de “desenvolvimento/crescimento”, de acordo com a metodologia de análise de ciclo de vida de destinos/produtos turísticos definida por Butler (1980).

CAPÍTULO 3: MÉTODOS E MODELOS DE PREVISÃO DA PROCURA TURÍSTICA

3.1. INTRODUÇÃO

Oliveira (2001) classificou visitante como toda a pessoa que se desloca a um local situado fora do seu ambiente habitual, durante um período inferior a 12 meses consecutivos e cujo motivo principal da visita é outro que não seja o de exercer uma atividade remunerada no local visitado. O mesmo autor refere que os visitantes são classificados em dois tipos: os turistas e os excursionistas.

Turistas são todos os visitantes temporários que permanecem pelo menos uma noite num estabelecimento de alojamento coletivo, ou num alojamento privado no local visitado. Por outro lado, excursionistas são todos os visitantes do dia que não passam a noite no local visitado (Cunha, 2001).

Esta descrição se alinha com a definição do turismo, do ponto de vista da procura, dada pela OMT (1995), na Conferência de Ottawa de 1991, com a inclusão de uma adenda em 1999, pela Comissão de Estatísticas da ONU, que diz:

O turismo compreende as atividades realizadas pelas pessoas durante suas viagens e estadias em lugares diferentes de seu entorno habitual, por um período de tempo consecutivo inferior a um ano, tendo em vista lazer, negócios ou outros motivos não relacionados ao exercício de uma actividade remunerada no lugar visitado.

(OMT, 1995, p. 1; ONU/OMT, 1999, p.5)

Assim, essas pessoas chamadas de “Turistas” são quem procuram um “ambiente fora de seu ambiente” e são elas quem gera a procura turística. Esta, possui diferentes fatores que a influenciam e determinam, nomeadamente fatores económicos, sociais,

tecnológicos e políticos, e acontecem em um determinado lugar denominado por “Destino Turístico”.

Essa procura tem vindo a ser pesquisada e uma avalanche de estudos sobre a previsão da procura turística tem acompanhado o crescimento deste setor. Para se ter uma ideia, segundo Song e Li (2008) 420 estudos foram publicados no período de 1960 a 2002, sendo muitos deles sobre a modelação e previsão da procura turística e utilizando na maioria das vezes a variável “Chegadas de turistas” como endógena.

Vários são os métodos e técnicas, e variadíssimos os modelos e metodologias empregues no estudo da previsão da procura turística que iniciaram na década de 60 e que têm vindo a ser melhorados e inovados. De abordagens simples aos mais complexos, os métodos de séries temporais, métodos econométricos e inteligência artificial têm marcado este campo de investigação. Entretanto métodos híbridos têm vindo a destacar-se e mais timidamente métodos de julgamento.

Este capítulo vai iniciar-se com a apresentação do conceito da procura turística, das suas características, fatores e determinantes, bem como a descrição dos métodos, técnicas e modelos que têm vindo a ser utilizados na previsão da procura turística. Esta descrição foi baseada na análise feita em 5 estados da arte elaborados por Song, Qiu e Park (2019) no período entre 1960 a 2019; Wu, Song e Shen (2017) no período entre 2007 a 2015; Goh e Law (2011) no período entre 1995 a 2009; Song e Li (2008) no período entre 2000 a 2008 e Witt e Witt (1995) no período entre 1965 a 1995, abrangendo um período entre 1960 a 2018. Apresenta-se, ainda, as tendências futuras sob a análise de diversos estudos realizados por diferentes investigadores.

Da análise cronológica dos métodos de previsão utilizados, feita por Song, Qiu e Park (2019) observa-se que o modelo SR (*Static Regression*) foi o mais utilizado na década de 60 e 70. Na década de 80 já se começara a aplicar as séries temporais no estudo de previsão e produzindo melhorias em alguns modelos como a AR (*Autoregressive*), ES (*Exponential Smoothing*), Naïve e Análise de Tendências. Na década de 90 os modelos de séries temporais continuavam a surgir e a melhorar. Também começaram a tornar-se

populares modelos baseados em regressões dinâmicas como os da metodologia de Box-Jenkins com variáveis exógenas e ADLM (*Autoregressive Distributed Lag Model*).

Novos modelos econométricos como VAR (*Vector Autoregressive*) e AIDS (*Almost Ideal Demand System*) surgiram, bem como modelos baseados em Inteligência Artificial (IA). Uma avalanche de métodos que espoletou na década de 90 continuaria com o aparecimento dos métodos híbridos indo até os anos de 2009. A partir de então os estudos de previsão da procura turística têm feito em grande maioria uso de modelos avançados de séries temporais e econométricos. Entretanto novas abordagens metodológicas continuam a ser investigadas e testadas como por exemplo os modelos baseados em IA e combinações com outros modelos de métodos diferentes (modelos híbridos).

Com este leque de possibilidades em usar diferentes modelos e por vezes combiná-los, Song e Li (2008) afirma que não existe nenhum modelo capaz de superar todos os demais no tocante à precisão de previsão. Mesmo não tendo ainda um modelo universal, Song, Qiu e Park (2019) observou um certo crescimento no número de modelos utilizados em cada estudo, passando em média de 1,1 (antes dos anos 80) para 1,9 (antes da década de 90) tendo a partir desta aumentado para 2,5 modelos.

Redes Neurais Artificiais (RNA) ou simplesmente Redes Neurais (RN) são uma metodologia muito abrangente cuja definição mais simples é a do inventor de um dos primeiros neurocomputadores, Dr. Robert Hecht-Nielsen. Ele define uma rede neuronal como: “[...] um sistema computacional composto por um número de elementos de processamento simples, altamente interligados, que processam informação pelo seu estado dinâmico de resposta à entradas (*inputs*) externos.” (Caudill, 1987, pp.46-52) e que tem por base e como objetivo tentar "reproduzir" o funcionamento da rede de neurónios do cérebro humano. Vista por Matich (2001), é mais uma forma de imitar certas características do ser humano, como a capacidade de memorizar e de associar coisas.

As redes neuronais estão marcadas por dois grandes momentos históricos: a primeira relacionada com o abandono da investigação das RN, que aconteceu no período de 1969

a 1981, devido a limitações na rede *Perceptron* (a sua limitação prende-se com o facto de não poder categorizar elementos não lineares) e a segunda com destaque ao ressurgimento das RN devido, entre outros avanços, ao ressurgimento do algoritmo *backpropagation* no período de 1981 a 1987 (Tépan, 2013; Moreno, 2002). A partir desta data os interesses na aplicação das RN em diversas áreas tem vindo a aumentar e em particular no campo da previsão de séries temporais.

As RN têm vindo a ser uma alternativa aos modelos convencionais na previsão das séries temporais devido a sua capacidade em extrapolar padrões a partir de dados existentes e de adaptar o seu comportamento (não altera sua estrutura) à medida que novos dados são introduzidos (Cortez, 1997). De entre vários tipos de RN destaca-se, neste trabalho, a rede *feedforward* multicamada associada ao algoritmo de aprendizagem *backpropagation*. Também são apontadas algumas vantagens e aplicações e uma breve descrição do funcionamento das RN, bem como o seu uso na previsão de séries.

Em síntese, neste capítulo pretende-se tratar o conceito da procura turística, as suas características e fatores; modelos e métodos mais utilizados, nomeadamente o modelo SARIMA do método de Box-Jenkins e as redes neuronais artificiais (RN); bem como suas vantagens e desvantagens quanto ao desempenho de previsão; desafios e tendências futuras que permitam melhorar a previsão da procura turística.

Ainda vai ser abordado o método de Suavização Exponencial que, apesar de não ter sido considerado, inicialmente, no plano de tese, se encontrou por bem estudá-lo e aplicá-lo para comparação com outros métodos, dado a sua variante (Suavização Exponencial Tripla, conhecida como Holt-Winters) que contempla a componente de sazonalidade na sua formulação. Também será trazido para o campo dos modelos de previsões a Lei de Benford, se bem que não se achou evidências, na literatura, da sua aplicação ao contexto da modelação da procura turística. Para terminar, vai ser apresentada uma proposta para a minimização do erro de precisão da previsão feita pelo modelo SARIMA, denominada de **Algoritmo de Atribuição do Erro (AAE)**.

3.2. CONCEITO DE PROCURA TURÍSTICA

A procura turística pode ser entendida ou expressa de diferentes maneiras:

- Número de turistas que se deslocam a um determinado destino, e que consomem uma determinada quantidade de bens e serviços (Cunha, 2001);
- Número de pessoas que viajam ou têm intenções de viajar para lugares distantes do local de residência e de trabalho visando usufruírem das ofertas turísticas⁹ (Cooper et al., 2001);
- Um conjunto de bens e serviços consumidos a um preço específico por pessoas que se deslocam para um determinado destino num determinado período de tempo (Pearce, 1989);
- Uma determinada quantidade monetária paga, pelos turistas, por um conjunto de bens e serviços durante a sua deslocação e estadia no local de destino (Matias, 2007);
- Uma quantidade de produtos turísticos que as pessoas estão dispostas a pagar sob determinadas circunstâncias em um certo período de tempo (Song & Witt, 2000).
- Um conjunto de produtos e serviços que o turista está disposto a consumir em uma determinada quantidade a um preço específico escolhido dentro demais preços alternativos, durante um certo período de tempo (Cooper, 2004);
- Número de noites que os turistas pernoitam nos estabelecimentos de acomodação turísticos no local de destino, também faz parte da medição da procura turística, segundo Lim (1997).

Em suma, pode-se dizer que o conceito de procura turística centra-se em três aspetos fundamentais: a existência de um local de destino, a disponibilidade de bens e serviços e o preço a pagar ao longo de todo o processo turístico. Dentre as diferentes maneiras de

⁹ Oferta turística é “o conjunto dos fatores naturais, equipamentos, bens e serviços que provoquem a deslocação de visitantes, satisfaçam as suas necessidades de deslocação e de permanência e sejam exigidos por estas necessidades” Cunha (1997, p.150).

entender a procura turística, neste trabalho é adotado o de Lim (1997), pois o âmbito desta investigação é a previsão da procura turística utilizando dados do número de dormidas nos estabelecimentos turísticos.

Entretanto existem três tipos de procura (Buhalis, 2004): **Efetiva** (em que os turistas estão efetivamente envolvidas no processo turístico, ou seja, tendo meios e vontade, participam como consumidores dos produtos turísticos); **Suprimida** (em que existem pessoas que devido a situações extremas não conseguem viajar para o destino no período planeado, mas que depois podem envolver-se no processo – chamado de *Procura potencial* – e aquelas que são impedidas de viajar devido a problemas no destino – chamado de *Procura diferida ou protelada*); e **Latente** (em que a procura é expressa a nível espacial e temporal em um determinado Sítio na Web. Por exemplo, a procura por alojamento turístico. Em suma, não viajam).

Mas, dado à dificuldade de medir o segundo tipo de procura, ela não tem sido levada em consideração (Buhalis, 2004; UNWTO, 2001). Em síntese, a procura turística pode ser dividida entre aqueles que viajam (Procura Efetiva) e os que não viajam (Procura Latente) a um determinado destino.

A procura turística pode ser medida de diferentes formas, dependendo da abordagem ser, por exemplo, física ou monetária. A procura turística física pode ser mensurada por meio de variáveis de volume, nomeadamente através das “chegadas às fronteiras de cada país”/“número de visitantes” e/ou pelo “n.º de dormidas”/“n.º de hóspedes” nos meios de alojamento. Por outro lado, a procura turística, numa abordagem pelo lado monetário, pode ser obtida através do “consumo turístico” e/ou “despesas turísticas”.

3.2.1. Características e Fatores Determinantes

Algumas características como elasticidade, sensibilidade e sazonalidade são vistas como inerentes ao estudar a procura turística. A elasticidade, segundo Lage (2009) e Jesus (2010), é a variação percentual da quantidade de produtos turísticos existentes no destino, dada a variação percentual dos preços dos mesmos produtos. Para Dias (2005), a sensibilidade está relacionada com pequenas variações, como epidemias, guerras, crise

política, transporte, etc., que influenciam na escolha do destino a ser visitado. Quanto a sazonalidade, esta é explicada com detalhes mais adiante neste capítulo.

Segundo Middleton (2002) fatores relacionados a aspetos económicos, tecnológicos, sociais e políticos influenciam a procura turística; e preços comparativos, demografia, geografia, atitudes socioculturais, mobilidade, governação/regulamentação, meios de comunicação e a tecnologia de informação e comunicação a determinam. Fatores como o tempo e a sazonalidade, segundo Poon (1994, 1993) influenciam a procura turística em relação à sua escala e magnitude.

Uysal (1998) classificou os fatores em: psicossociológicos, económicos e exógenos. Cunha (2013), por sua vez categorizou-os em três: estruturais, conjunturais e psicossociológicos. UNWTO (2008) já definem três tipos: fatores *push*, fatores *pull* e fatores de resistência. Divisekera (2013) cria duas categorias: fatores que afetam as viagens (*ex ante*) e fatores que influenciam a escolha dos destinos (*ex post*). Entretanto as variáveis que constituem cada tipo de classificação/categorização dos fatores estão resumidas na Tabela 3.1 e encontram-se entre os pilares e variáveis de competitividade turística (WTTC – 2017).

Mas, por detrás desses fatores e determinantes existem diferentes motivos nomeadamente: lazer, recreação e férias; visitar amigos e parentes; compromissos comerciais e profissionais; tratamentos de saúde; peregrinações religiosas e outras; outros motivos mais pessoais (Nelwamond, 2009) que podem ser classificados por diferentes tipos de atividades.

Entretanto, se pudesse ser resumido tudo o que já se disse, recorrer-se-ia à Declaração de Manila sobre o Turismo Mundial feita em 1980, que diz que o objetivo final do turismo era “(...) a melhoria da qualidade de vida e a criação de melhores condições de vida para todas as pessoas” (Cooper et al., 1993, p. 14). Esse sentimento se reflete na visão do turismo do Governo de Cabo Verde conforme as Grandes Opções do Plano Estratégico de Desenvolvimento do Turismo para 2018-2030 (GOPEDS-Turismo, 2019).

Tabela 3.1: Variáveis que constituem cada tipo de classificação/categorização dos fatores.

Fatores	Variáveis
Psicossociológicos	Fatores demográficos; Motivações; Preferências da viagem; Imagem dos destinos; Perceção dos destinos; Oportunidade; Distância cognitiva; Atitudes sobre os destinos; Tempo de lazer; Férias remuneradas; Experiências anteriores; Fase do ciclo de vida; Capacidade física e saúde; Similaridades culturais e afiliações; Circunstâncias familiares.
Económicos	Rendimento disponível; Produto Interno Bruto per capita; Consumo privado; Índice de preços do consumidor; Preços turísticos; Custo de vida no país de destino; Diferenças nas taxas de câmbio; Preços relativos entre destinos competitivos; Despesas de promoção e eficácia do marketing.
Exógenos	Avaliação dos recursos da oferta; Crescimento económico e estabilidade; Ambiente económico, social e político; Avanços tecnológicos; Acessibilidade; Nível de desenvolvimento de infraestruturas; Desastres naturais; Fatores de segurança; Condições de saúde; Guerra e terrorismo; Atrações sociais e culturais; Grau de urbanização; Eventos especiais; Obstáculos à circulação; Restrições, regras e leis.
Estruturais	Rendimento disponível; demografia; duração do tempo de trabalho; Densidade populacional e taxa de urbanização; Progresso científico e técnico; Transportes.
Conjunturais	Variações cambiais; A inflação.
<i>Push</i>	Tamanho da população; Rendimento (PIB e Rendimento Nacional Bruto); Distribuição da educação; Distribuição da idade; Tempo de lazer; Estrutura familiar; Condições meteorológicas (no país de origem).
<i>Pull</i>	Amigos e familiares; Clima (no destino); Relações comerciais; Relações socioculturais; Marketing; Eventos especiais; Costumes.
Resistência	Preço-produto; Necessidades; Taxas de câmbio; Ações dos competidores; Capacidades da oferta; Tempo de viagem/distância; Taxas; Ameaças à segurança pessoal.
<i>ex ante</i>	Rendimento disponível; Tempo de lazer; Fatores institucionais; Popularidade crescente das viagens ao estrangeiro.
<i>ex post</i>	Custos dos transportes; Custos dos produtos e serviços no destino; Fatores sociopolíticos; Singularidade dos atributos turísticos do destino.

Fonte: Elaboração própria

A melhoria referida no parágrafo anterior traria mudanças na procura turística como previsto por Schwaninger (1989), como, por exemplo, o crescimento do turismo devido a sua diferenciação, especialização, segmentação e mais personalizado; mas tal crescimento não seria estático, ou seja, continuaria mudando (Prosser, 1994). Essas previsões têm vindo a tornarem-se realidade para o turismo em Cabo Verde, onde a ilha do Sal se insere. O país definitivamente tem os recursos para se concentrar num turismo mais diferenciado, como se pode constatar no Capítulo I - Turismo: o Mundo e Cabo Verde.

3.3. ESTADO DA ARTE: PREVISÃO DA PROCURA TURÍSTICA

3.3.1. Métodos e modelos de previsão no período entre 1960 a 2018

Segundo Goh e Law (2011) a partir de 1985 as técnicas de modelação e previsão da procura turística começaram a ganhar mais qualidade e a contribuir de forma mais significativa para a indústria do turismo, trazendo avanços principalmente em técnicas econométricas, nomeadamente o ECM (*Error Correction Model*) e o TVP (*Time Varying Parameter*); nas séries temporais como GARCH (*Generalised Autoregressive Conditional Heteroskedastic Model*) e o MARIMA (*Multivariate Autoregressive Integrated Moving Average*); e na Inteligência artificial usando redes neuronais.

Dos 211 estudos analisados por Song, Qiu e Park (2019), a maioria (55,9%) fez uso de modelos de séries temporais. Wu, Song e Shen (2017) consultaram 145 estudos, mas não se mencionou a quantidade de modelos segundo grupo.

Na revisão da literatura sobre o estado da arte feito por Goh e Law (2011) no período entre 1995 a 2009, eles agruparam os métodos e técnicas em três grupos: Modelos de Séries Temporais, Modelos Econométricos e Inteligência Artificial. Observaram 384 modelos em 155 trabalhos consultados em que 52,6% eram referentes às Séries Temporais, 41,7% a modelos econométricos e 5,7% a modelos de Inteligência Artificial.

Song e Li (2008) consultaram 121 artigos no período entre 2000 a 2008. Destes, 59,5% utilizaram modelos de séries temporais e 58,7% modelos econométricos e 9,1% outros modelos, principalmente baseados na Inteligência Artificial.

Witt e Witt (1995) analisaram 49 estudos no período de 1965 a 1995 totalizando 251 modelos utilizados. Destes, a maioria (65,7%) foram aplicados modelos de séries temporais seguindo-se os modelos econométricos.

Segundo Song, Qiu e Park (2019) os métodos para a previsão da procura turística foram divididos em quatro categorias e reconhecido em todo o mundo, sendo elas: os modelos

de séries temporais, os modelos econométricos, os modelos baseados em Inteligência Artificial e os métodos de julgamento. Os três primeiros se aplicam para dados quantitativos e o outro modelo para dados tanto qualitativos como quantitativos. Uma comparação feita de diferentes modelos mencionados nesta secção se encontra no Anexo II.

3.3.2. Séries Temporais

A maioria dos estudos revisados por Goh e Law (2011) utilizaram modelos de séries temporais, um total de 202 na qual 34,0% fizeram uso dos métodos Box-Jenkins, 18,0% do método Naïve, 11,0% de vários métodos de suavização exponencial (**SE** – *Smoothing Exponential*), 7,0% modelos autorregressivos, 7,0% modelos estruturais de séries temporais, 3,0% modelos de média móvel e 20,0% restantes. Dos 211 estudos revisados por Song, Qiu e Park (2019), a maioria (55,9%) fez uso de modelos de séries temporais, sendo que 55 destes utilizaram modelos básicos. Especificamente 60,0% dos modelos de séries temporais eram os do tipo ARIMA (*Autoregressive Integrated Moving Average*). Quanto a revisão de Witt e Witt (1995), mais de 54,0% dos modelos basearam-se em Séries Temporais, com destaque para o modelo Naïve.

Os modelos **Naïve** aplicados à procura turística demonstraram ser tanto eficientes como não. Tal inconstância na previsão pode ser devido ao tipo de dados, por exemplo, ter que utilizar dados sazonais ao invés de anuais (Goh & Law, 2011). Em síntese, um método Naïve gera um valor futuro igualando-o ao último valor da série. Existe o **Naïve 1** (*No-Change Model*) e **Naïve 2** (*Constant Growth Rate Model*) cuja diferença reside na utilização de dados sazonais ajustados em vez de dados brutos no Naïve 2.

No estudo de Song, Qiu e Park (2019) foi observado que os modelos Naïve 1 e Naïve 2 foram os mais utilizados dentre os modelos básicos, com destaque para Naïve 1 que se mostrou mais preciso, principalmente nas previsões a curto prazo. Witt e Witt (1995) já tinham observado tal precisão em termos da magnitude do erro em horizontes de um ano, bem como seu desempenho superior com relação a outros métodos mais complexos. Ainda é mencionado o modelo **sazonal-Naïve** que tem sido utilizado em

muitos estudos e cujo papel é projetar o último período sazonal. Para Song e Li (2008) os modelos Naïve 1 e Naïve 2 que surgiram repetidas vezes em estudos a partir do ano 2000 foram utilizados como *benchmarks* na avaliação da precisão da previsão.

O modelo do tipo **SE Holt-Winters** que integra a componente sazonal revelou ser superior a outros tipos de SE (suavização exponencial), segundo Song, Qiu e Park (2019).

O modelo **AR** (*Autoregressive*) também foi observado em uma dúzia de estudos, utilizando dados anuais, trimestrais e mensais, tendo apresentado um bom desempenho nas previsões (Goh & Law, 2011). Segundo Witt e Witt (1995) o seu desempenho é melhor no horizonte de dois anos. Este modelo calcula o valor atual da variável dependente com base em seus valores em períodos de tempo anteriores e acrescido do valor de erro aleatório. Entretanto, segundo Song, Qiu e Park (2019), uma versão sazonal deste modelo (**AR sazonal**) tem sido adotada na procura turística, porém não foi mencionado o seu desempenho.

Poucos dos estudos analisados por Goh e Law (2011) utilizaram o modelo **MA** (*Moving Average*) não observando um bom desempenho nas previsões. Este modelo calcula o valor atual da variável dependente com base nos valores das inovações (ε_t) passados usando um processo de média móvel.

Dos estudos observados por Goh e Law (2011), a abordagem de Box-Jenkins foi a mais utilizada nos modelos de séries temporais, especificamente o modelo **ARIMA** que foi considerado como superior aos outros modelos na precisão em prever a procura turística. Segundo Song, Qiu e Park (2019) os modelos ARIMA têm-se tornado muito flexíveis e amplamente aplicados à previsões turísticas. Dos estudos analisados, por estes investigadores, que utilizaram as técnicas de séries temporais, 62,7% adotaram os modelos ARIMA. A maioria destes modelos se mostraram superiores, principalmente quando era avaliado pelo menos um destino. Mas do levantamento feito por Song e Li (2008) observaram contradições quanto à superioridade do modelo, comparando com o SARIMA (*Seasonal Autoregressive Integrated Moving Average*).

O modelo **SARIMA**, que é o ARIMA sazonal, foi observado em vários estudos revelando bom desempenho tendo vindo a aumentar a sua aplicação (Goh & Law, 2011). Segundo

Song e Li (2008) a sua popularidade deve-se ao fato de que muitos decisores têm interesse em conhecer a variação sazonal da procura turística.

Um outro modelo observado pelos investigadores é o **ARIMAX**, uma extensão do ARIMA, sendo X a representar as variáveis exógenas. Este modelo concentra principalmente em discernir a dinâmica da procura turística, e tem mostrado superior em previsões de longo prazo, ao menos com relação ao modelo SARIMA. Também traz resultados satisfatórios quando combinado com modelos econométricos do tipo **VP** (*static varying parameters*) e **MIDAS** (*Mixed-Data Sampling*) (Song, Qiu & Park, 2019).

O **MARIMA**, que é um ARIMA sazonal multivariado (ou SARIMA multivariado), também foi utilizado na previsão da procura turística quando acontecem eventos externos ao destino. Neste contexto foi considerado superior aos demais outros modelos de séries temporais (Goh & Law, 2011). Este modelo veio melhorar, de forma significativa, o desempenho do SARIMA padrão. O que não acontece com o ARIMA multivariado não conseguindo superar o ARIMA univariado (Song & Li, 2008).

É mencionado o modelo **ARFIMA** (*Fractionally Integrated Autoregressive Moving Average*) e o modelo **SARIMA-In** (*SARIMA with Intervention*) no estado da arte elaborado por Song, Qiu e Park (2019). Entretanto, nada se mencionou sobre o desempenho dos mesmos. Mas, Wu, Song e Shen (2017) já haviam observado que o modelo ARFIMA se mostrou superior aos modelos SARIMA e ao ARARMA (*ARMA autoregressive*).

Nos estudos analisados foi também observado a aplicação do modelo **SARIMAX**, que é uma extensão da SARIMA com variáveis exógenas, e que mostrou ser superior ao modelo padrão SARIMA e o SE Holt-Winters (Song, Qiu & Park, 2019).

Foi observado um outro modelo avançado de séries temporais em Wu, Song e Shen (2017), a **ARARMA** (*ARMA autoregressive*), mas sem qualquer menção sobre seu desempenho.

Um outro modelo e pouco utilizado foi o **REG-ARIMA**. É um modelo de regressão com erros ARIMA onde os regressores são predefinidos ou definidos pelo utilizador. A sua aplicação mostrou-se superior aos modelos Naïve e ARIMA (Goh & Law, 2011).

Também foi utilizado, em meia dúzia de estudos, um **modelo trigonométrico** com funções de onda senoidal. Foi constatado uma limitação na sua aplicação no tocante a sazonalidade estocástica e ciclos indeterministas, o que sugere o uso de dados sazonais ajustados (Goh & Law, 2011).

Os modelos **ARCH** (*Autoregressive Conditional Heteroskedastic*) e **GARCH** têm sido também aplicados, principalmente o GARCH dado a sua capacidade de capturar a volatilidade e a existência de heterocedasticidade (Goh & Law, 2011). Também foi observado em Song, Qiu e Park (2019) e em Wu, Song e Shen (2017) e na versão multivariada em Song e Li (2008), mas sem menção do seu desempenho. Também é mencionado (apenas) em Wu, Song e Shen (2017) a aplicação do modelo **ARIMA-ARCH**.

Segundo Song, Qiu e Park (2019), a Análise de Espectro e a Análise de Wavelets têm vindo a surgir no campo de procura turística. Ainda a análise denominada por **SSA** (*Singular Spectrum Analysis*) foi observado em Wu, Song e Shen (2017) e é uma técnica não paramétrica que também decompõe a série em diferentes componentes (ex. sazonalidade e tendência). O SSA tem mostrado ser superior ao SE, SARIMA e STSM (Wu, Song & Shen, 2017).

Nas séries temporais aparecem os modelos estruturais **STSM** (*Structural Time Series Model*) básicos que decompõem as séries em quatro, sendo uma tendência estocástica, um ciclo periódico, um componente sazonal e outro irregular; assumindo média zero e sem correlação serial. Ainda a **M-STSM** associada a estrutura multivariada da série. Segundo Goh e Law (2011), tanto o básico como a multivariada foram utilizadas várias vezes a título de comparação nas previsões. Constatou-se que todas as séries temporais utilizadas nesses estudos foram testadas quanto a estacionariedade utilizando o teste de raiz unitária ADF (*Argumented Dickey Fuller*).

Song, Qiu e Park (2019) acrescentam que este modelo estrutural investiga a influência das variáveis exógenas com destaque para tendências e componentes sazonais e de ciclo. O desempenho do modelo não foi mencionado por esses investigadores. Mas em Wu, Song e Shen (2017) é mencionado que quando são acrescentados ao modelo os chamados indicadores de sentimento de negócios, ele resulta em previsões melhores.

Um outro modelo introduzido no estudo das previsões da procura turística é o **ISS** (*Inovations State Space Models*) que tem demonstrado grande desempenho em vários estudos (Wu, Song & Shen, 2017). Este modelo utiliza a noção de suavização exponencial e lida com apenas uma única fonte de erro, contrariamente ao STSM (que tem a mesma especificação do ISS) que permite que cada equação lide com seu próprio termo de erro.

Modelos não lineares de séries temporais também têm sido aplicados e observados em Wu, Song e Shen (2017) como o modelo limiar auto-excitante autorregressivo e o modelo de comutação de Markov. Este se mostrou superior ao primeiro e ao modelo AR.

A combinação do modelo STSM com o TVP também foi aplicado e mostrou-se superior a outros modelos tanto de séries temporais como econométricos (Wu, Song & Shen, 2017)

Abordagens integrativas como a ARIMA sazonal combinado com a regressão de onda senoidal e o modelo de volatilidade combinado com o modelo exponencial foram utilizados, porém seus resultados não indicam melhorias na previsão comparando com os modelos não combinados (Goh & Law, 2011).

Estes investigadores apontam outros métodos de séries temporais utilizados na previsão da procura turística, nomeadamente: o modelo **ARAR** (*Autoregressive Autoregressive*), modelo de decomposição, modelo de mudança proporcional, modelo de Gompertz em forma de S, modelo sazonal fracionário, modelo de quebra de tendência sequencial, modelo de espaço de estados e o modelo polinomial cúbico. Mas não se fez menção de seus desempenhos.

Neste trabalho é aplicado o modelo SARIMA e o método de Holt-Winters, pois a sazonalidade é uma característica do turismo na ilha do Sal e em Cabo Verde, e por ser um dos modelos mais utilizados nos estudos de previsão da procura turística em séries temporais com presença de sazonalidade.

3.3.3. Modelos Econométricos

Nos modelos econométricos, a maioria dos estudos observados por Goh e Law (2011) e Song e Li (2008), adotaram como método de estimativa os Mínimos Quadrados

Ordinários (MQO). Song e Li (2008) comentam que métodos de estimação mais eficazes deveriam ter sido utilizados ao invés dos MQO devido a sua regressão espúria e falta de dinâmica.

O modelo **SR** (*Static Regression*) onde se envolve uma única regressão estática, segundo Song, Qiu e Park (2019), foi muito utilizado inicialmente apesar de problemas obtidos pela ausência de estacionariedade, tem vindo a ser reutilizado nos últimos anos. Geralmente é mais utilizado no sentido de determinar a influência de diversos fatores nos valores atuais.

Tem sido utilizado o modelo **DL** (*Distributed Lag*) que determina a influência de diversos fatores tanto nos valores atuais como nos valores anteriores, porém tem demonstrado ser muito limitado na previsão turística. Aparece o modelo de correção de erros **ECM** que tem produzido resultados muito significativos, segundo Song, Qiu e Park (2019). Song e Li (2008) observaram que em alguns estudos o ECM supera alguns modelos como por exemplo o SARIMA e o Naïve.

O **TVP** tem superado em muitos casos os outros modelos econométricos. A combinação dos modelos ECM e TVP revelou-se mais preciso dentre os demais em estimar a procura turística (Goh & Law, 2011). Esta integração mostrou ter bom desempenho em previsões a curto prazo apesar de ser ainda muito pouco utilizado (Song & Li, 2008). A combinação do TVP com o ADLM funciona tão bem quanto a sua combinação com a AIDS e modelos SR (Song, Qiu & Park, 2019). Entretanto quando se trata de previsões a um passo a frente o TVP obteve um desempenho melhor que o ECM, VAR e ADLM, superando também o ARIMA e o Naïve, em alguns estudos. Quando a previsão é de longo prazo passa a ser, em uma avaliação geral, o ECM o que demonstra melhor desempenho (Song & Li, 2008).

Uma forma reduzida do modelo ADLM tem sido também utilizada, bem como a sua integração com recursos **MIDAS** que permite a previsão com base em dados de frequência mista. O MIDAS tende a fornecer melhores precisões de previsão (Song, Qiu & Park, 2019).

Quanto aos modelos estáticos, vários estudos utilizaram a abordagem de regressão de multi-equações tendo como destaque o modelo de sistemas de equações quase ideal da

procura turística – **AIDS** (Goh & Law, 2011). Segundo Song, Qiu e Park (2019), têm sido utilizadas, neste sistema, versões da AIDS para fazer a previsão da procura turística e, para além de ser utilizado num sistema estático, pode ser estendido para um sistema dinâmico. Segundo Song e Li (2008), é utilizado normalmente para analisar a procura em diferentes destinos com base em uma origem, ou seja, quando um turista escolhe um destino entre vários outros destinos alternativos (Wu, Song & Shen, 2017). O modelo AIDS num sistema estático não se mostrou muito eficiente, segundo Goh e Law (2011). Mas Song e Li (2008) observaram em sua lista de artigos analisados estudos em que mostra a superioridade do AIDS face ao VAR padrão. Tal superioridade também foi observada em Wu, Song e Shen (2017). Entretanto alguns estudos onde aparece o modelo inserido num sistema dinâmico integrando o ECM e em outros estudos que incorporaram nele o TVP nas formas de correção de erro de longo e curto prazo (TVP-LR-AIDS e TVP-EC-AIDS) produziram melhores resultados (Song, Qiu & Park, 2019; Goh & Law, 2011; Song & Li, 2008).

Um outro modelo que tem sido utilizado é o de equações estruturais (**SEM** – *Structural Equation Model*), que analisa a estrutura de relacionamentos entre variáveis, em que cada equação no sistema é composta de conjuntos de variáveis tanto exógenas como endógenas (Goh & Law, 2011). Dada a sua capacidade de representar as relações causais entre as variáveis, tem sido adequado na pesquisa da procura turística (Song & Li, 2008), porém o seu desempenho não tem sido analisado.

Na revisão feita por Wu, Song e Shen (2017) foi observado um modelo não linear **STR** (*Smooth Transition Regression*). Apesar de ser muito vantajoso em termos técnicos, sua aplicação foi encontrada em apenas um estudo de procura turística em 2014.

Em relação aos modelos dinâmicos, foi observado em vários estudos analisados por Goh e Law (2011) a aplicação de um modelo dinâmico de equação única denominado **ADLM**. Segundo Song, Qiu e Park (2019) este modelo tem produzido resultados muito significativos a exemplo do modelo ECM e muitas vezes é combinado com a abordagem **GETS** (*General-To-Specific*), onde o ADLM começa por remover as variáveis que são desnecessárias e seguidamente entra o GETS para reduzir a quantidade de variáveis,

eliminando aquelas insignificantes até chegar a um modelo considerado ótimo. Esta abordagem de previsão econométrica se baseia na modelação geral para a específica. Modelos GETS de agregação *bootstrap* foram observados em apenas um estudo de 2018 em que foi constatado que este modelo se mostrou com melhor desempenho nas previsões comparando com GETS padrão.

Outros modelos dinâmicos mencionados por Goh e Law (2011) incluem os modelos de difusão, função transferência (usado em alguns estudos na comparação com outros modelos econométricos e de séries temporais), regressão de comutação gradual e o ARMAX (*Autoregressive Moving Average Exogenous*).

Um dos modelos dinâmicos de multi-equações de nome **VAR** foi utilizado em mais de duas dúzias de estudos, mas não ficou clareado o seu desempenho e resultados mesmo quando supera outros modelos. Entretanto apenas um estudo mostrou que o VAR padrão pode ser promissor na previsão da procura turística e noutros estudos pode ser superado por técnicas econométricas concorrentes mais recentes. Para especificar corretamente o modelo criou-se o **VEC** (*Vector Error Correction*). Estes modelos são uma extensão do modelo de equação estática única e são capazes de capturar a interdependência de várias series temporais (Song, Qiu & Park, 2019; Goh & Law, 2011; Song & Li, 2008).

Dado a essa incerteza quanto ao desempenho do VAR, Song, Qiu e Park (2019) observaram, nos estudos que analisaram, a aplicação do modelo **BVAR** (*Bayesian VAR*) como um melhoramento tendo apresentado melhores resultados que o VAR não bayesiano. A mesma observação foi feita por Song e Li (2008) não tendo registrado comparações com outros modelos econométricos.

Um outro estudo analisado apresentou o **GVAR** (*Global Vector Autoregressive*), que é uma estrutura global do VAR e outro estudo introduziu a técnica bayesiana **BGVAR** (*Bayesian Global Vector Autoregressive*) para prever a procura turística (Song, Qiu & Park, 2019).

Outros modelos dinâmicos utilizados na procura turística são o **EC-LAIDS** (*Error Correction-Linear AIDS*) e o **TVP-EC-AIDS** que incorpora o TVP no modelo EC-LAID (Goh &

Law, 2011). Este modelo, em comparação com sua versão estática, se mostrou superior (Song & Li, 2008).

Segundo Song, Qiu e Park (2019) os modelos ADLM e ECM foram os mais utilizados dentre os artigos analisados que utilizaram a abordagem econométrica; e eles são enfáticos em medir a relação de causa e efeito entre os fatores que influenciam a procura turística.

Dados agrupados e em painéis também foram aplicados nos estudos observados. O primeiro analisa uma amostra em diferentes momentos e no segundo as mesmas amostras são acompanhadas e analisadas ao longo do tempo. Tais aplicações foram vistas como vantajosas e sem qualquer obstáculo apresentado. Entretanto, existe alguma preocupação quanto à possibilidade de se violar a suposição de ser constante a variância dos erros nos dados agrupados e no caso dos dados em painéis, de ignorar as condições necessárias que determina a escolha de um modelo, como por exemplo a homogeneidade dos coeficientes e o nível de correlação existente entre os efeitos da seção transversal e as variáveis explicativas (Goh & Law, 2011).

Também Song e Li (2008) observaram que a análise de dados em painéis é adequada na previsão da procura turística pois reduz a multicolinearidade e aumenta o grau de liberdade no modelo. Esta vantagem também foi observada em Wu, Song e Shen (2017) acrescentando que é mais eficiente ainda quando a frequência dos dados das séries temporais é curta.

Segundo Song, Qiu e Park (2019) o **PDR** (regressão de dados em painéis) tem vindo a ser raramente aplicado na previsão da procura turística, porém pode ser um recurso integrado a análise de dados em painéis. Todavia, seu desempenho não foi comentado.

Nos modelos econométricos aplicados nos estudos analisados por Song, Qiu e Park (2019) e Song e Li (2008) observou-se um conjunto de fatores mais importantes que determinam a procura turística: nível de rendimento dos turistas, taxas de câmbio, preço relativo dos produtos turísticos, preço substituto de produtos turísticos (preços concorrentes medidos pelo IPC – índice de preço do consumidor), mudança climática, estabilidade política, eventos pontuais, ataques terroristas e crises financeiras. Entretanto, outros fatores

tradicionais que também influenciam a procura foram observados por Wu, Song e Shen (2017) nomeadamente taxa de desemprego, custo de transporte (medido pelo preço de combustível), gastos com publicidade, número de população na origem, volume de negócio entre a origem e o destino e ainda fatores sociais e culturais incluindo variáveis fictícias para capturar o impacto da sazonalidade. Outros fatores não tradicionais que tem vindo a valorizar os modelos de previsão da procura turística são o índice climático, índice de confiança e a restrição de vistos.

O uso desses modelos econométricos pressupõe a utilização de variáveis exógenas. Neste trabalho não é feita nenhuma aplicação dos mesmos, dada a inexistência de dados mensais ao longo do horizonte temporal da série (janeiro de 2000 a dezembro de 2018) em todas as variáveis exógenas socioeconómicas inicialmente propostas no Plano da Tese.

3.3.4. Inteligência Artificial

Segundo Song, Qiu e Park (2019) e Goh e Law (2011), a Inteligência Artificial (IA) aplicada à previsão turística divide-se em dois tipos de categorias principais: métodos causais baseados em IA e os métodos de séries temporais baseados em IA. Foi observado por Goh e Law (2011) mais de uma dúzia de estudos que fizeram uso da IA e alguns destes utilizaram em simultâneo os dois tipos de métodos.

Dentro da IA o primeiro método a ser utilizado foram as redes neuronais artificiais ou simplesmente RN (redes neuronais). Estudos sobre a previsão da procura turística, mostraram que as RN são, na maioria das vezes, superiores aos modelos de séries temporais e modelos de regressão múltipla (Song & Li, 2008).

3.3.4.1. Métodos causais baseados em IA

Dentro da categoria dos métodos causais foi utilizado uma rede neuronal supervisionada de *feedforward* para prever a procura turística tendo variáveis socioeconómicas como independentes. O desempenho da rede superou os modelos econométricos mais em uso, bem como os modelos de regressão múltipla, Naïve 1, média móvel (MA) e suavização

exponencial (SE). Também foi construída uma rede neuronal de retropropagação que se mostrou superior a outras abordagens. Um dos algoritmos citados que permite descrever o mapeamento da rede é o *perceptron* multicamada (Goh & Law, 2011). Este tipo de rede, bem como a RBF (*Radial Basis Function*) também foi observado por Song, Qiu e Park (2019). Em Wu, Song e Shen (2017) é mencionada também a rede neuronal de regressão generalizada (GRRN) e a rede neuronal de Elman. Entretanto, segundo Wu, Song e Shen (2017), o *perceptron* é o modelo de rede mais utilizado, se bem que o seu processo de aprendizagem é mais longo que, por exemplo, o RBF.

De uma forma geral, as redes neuronais têm superado várias técnicas de previsões incluindo as de séries temporais (Naïve 1, MA e SE) e o seu desempenho também é notado mesmo quando os dados da série temporal são de qualidade duvidosa (Song, Qiu & Park, 2019; Goh & Law, 2011).

Segundo estes investigadores ainda não está claro o tipo de modelo de IA a ser aplicado na previsão do turismo.

3.3.4.2. Modelação de séries temporais baseados em IA

Dentro da categoria dos modelos de séries temporais baseados em IA, também foi feito uso da rede neuronal *feedforward* em previsões mensais resultando num melhor desempenho face a outros métodos. Também foi aplicada a rede neuronal de Elman na previsão de chegada de turistas de diferentes locais de origem, mostrando-se superior a outros métodos. Esta rede também foi observada por Song, Qiu e Park (2019).

Modelos inovadores como **séries temporais nebulosas** e **teoria cinzenta** foram também utilizadas (Song, Qiu & Park, 2019; Goh & Law, 2011), porém os resultados não são claros quanto a ser melhor que os restantes em precisão (Goh & Law, 2011). O método das séries temporais nebulosas é mais vantajoso para séries temporais com dimensões curtas, e o foco da teoria cinzenta está na incerteza do modelo e na insuficiência de informação. Estes modelos foram observados por Song e Li (2008) em apenas um estudo e não demonstraram ser superiores entre eles, dado que o desempenho dependia do país de origem em análise (Song & Li, 2008).

Cho (2009) utilizou a rede neuronal de retropropagação com dois modelos de séries temporais: Holt-Winters e SARIMA. Ao ter realizado três experimentos concluiu que tal procedimento supera o desempenho de precisão do modelo da rede neuronal. O que fez, num primeiro momento, foi incluir chegadas mensais X_{t-1} e X_{t-2} para prever a chegada mensal \hat{X}_t ; num segundo momento utilizou as chegadas mensais de um ano antes $-X_{t-12}$ e de dois anos antes $-X_{t-24}$ para prever as chegadas mensais \hat{X}_t ; e num terceiro momento utilizou todos os pontos de tempo X_{t-1} , X_{t-2} , $-X_{t-12}$ e $-X_{t-24}$ para prever \hat{X}_t .

Ainda modelos de IA não lineares, como o **SVR** (*Support Vector Regression*) conhecidas também como redes neuronais de retropropagação foram combinados com modelos Naïve, SE e ARIMA (que são modelos lineares). O SVR combinado com algoritmos genéticos (GA – *Genetic Algorithm*) deu origem ao **GA-SVR** também observado nos estudos de procura turística (Song, Qiu & Park, 2019).

O modelo **GFS** (*Genetic Fuzzy System*) que é um sistema híbrido de IA também é utilizado e tem tido bom desempenho (Song, Qiu & Park, 2019).

Ainda foi aplicado um modelo de máquina de vetores de suporte (**SVM** – *Support Vector Machines*) com algoritmos genéticos. Segundo Goh e Law (2011) esta abordagem superou os modelos de redes neuronais e o ARIMA com base no erro quadrado médio normalizado (NMSE – *Normalized Mean Square Error*) e no erro percentual absoluto médio (MAPE – *Mean Absolute Percentage Error*). Song e Li (2008) acrescentam que o SVM também superou o SARIMA.

De frisar que nos artigos analisados por Song e Li (2008) foi observado que o desempenho das previsões teve por base, para além do MAPE, a raiz do erro percentual quadrado médio (RMSPE – *Root Mean Square Percentage Error*).

Foi observado por Wu, Song e Shen (2017) a aplicação do modelo **GPR** (*Gaussian Process Regression*), mas sem qualquer análise sobre seu desempenho. Entretanto uma versão do mesmo, o **GPR esparsa**, é visto como melhor quanto à complexidade computacional.

Wu, Song e Shen (2017) observaram também a aplicação do modelo **fuzzy** baseado em rede que se mostrou superior ao modelo fuzzy baseado em séries temporais. Ainda tem sido utilizado com alguma frequência a combinação do modelo de sistema difuso com algoritmos genéticos.

3.3.5. Outras abordagens quantitativas

O modelo de **conjuntos aproximados** também foi aplicado, utilizando variáveis socioeconómicas independentes bem como variáveis não económicas alcançando resultados muito satisfatórios (Goh & Law, 2011). Esta abordagem também foi observada por Song e Li (2008). Esse tipo de modelo prevê níveis da procura turística ao invés de prever valores exatos, pelo que é considerado útil e complementar aos modelos econométricos na análise da procura numa perspetiva micro (Song & Li, 2008). Os métodos de julgamento se caracterizam por um conjunto de opiniões subjetivas de especialistas, quer seja individual ou em grupo de modo a gerar consensos sobre uma determinada matéria. Na previsão da procura turística Song, Qiu e Park (2019) também observou a aplicação do modelo **Delphi** usada geralmente para a previsão de longo prazo. Entretanto, o seu desempenho é questionável devido à subjetividade das opiniões envolvido num debate de previsão. Este método também foi observado nos estudos analisados por Witt e Witt (1995) que na altura começara a atrair muita atenção.

De uma forma geral os métodos de julgamento têm vindo a ser pouco aplicados quando individuais, mas ao serem integrados com as técnicas estatísticas têm trazido resultados encorajadores. Esta forma de integração, denominada de **quase-Delphi**, foi observado por Song e Li (2008), porém é inconclusiva quanto ao seu desempenho.

3.3.6. Tendências e desafios futuros das previsões da procura turística

Nesta subsecção é apresentado um leque de tendências e desafios para o futuro mencionados em diferentes trabalhos (Song, Qiu & Park, 2019; Wu, Song & Shen, 2017; Song & Li, 2008) que serviram de base ao estado da arte:

- Os dados mensais são os que, na prática, interessam mais os decisores e políticos no setor turístico. Entretanto, na maioria dos estudos analisados, no âmbito do estado da arte, foi utilizado dados anuais, trimestrais, semanais e diários. Poucos são os estudos que utilizam dados mensais, sendo, portanto algo que precisa ser mais trabalhado.
- A utilização mais frequente da análise da procura turística com base em dados desagregados [separação das informações coletadas em seus diferentes componentes. Por exemplo, ao invés de utilizar dados de dormidas em que contempla tanto internacionais e nacionais, pode-se analisar apenas as dormidas internacionais, separadas das nacionais] sempre que estes estiverem disponíveis.
- As estimativas *bayesianas* podem vir a proporcionar melhores resultados quanto à previsão, assim como o uso de extensões dos modelos de séries temporais com variáveis e dimensões multivariadas.
- A continuação da aplicação de modelos combinados e híbridos, é uma outra tendência, pois, tem trazido resultados bastante significativos. Também devem ser levadas em conta as técnicas avançadas de TVP, pois tem vindo a ganhar popularidade e possibilita mudanças estruturais ao longo do tempo.
- O modelo TVP-ECM, apesar de seus resultados promissores, tem sido muito pouco utilizado, pelo que se espera um grande esforço em sua aplicação.
- Será de interesse futuro avaliar (i) o desempenho do modelo BVAR em comparação com outros modelos econométricos, nomeadamente o ECM e o TVP, pois não se registou tal informação no estado da arte e (ii) a comparação do desempenho do modelo AIDS em sistemas dinâmicos com os modelos VAR cointegrado.
- O modelo AIDS tende a vir a ser muito mais explorado, devido a inter-relação que existe na procura por diferentes destinos.
- Dada a pouca utilização do modelo SEM na procura turística, recomenda-se a sua aplicação para diferentes abordagens bem como a análise de seu desempenho.
- Estudos futuros no tocante ao desempenho do SVM na previsão para comparação com modelos econométricos modernos e que pudesse ser proposto uma extensão do modelo de forma a permitir a incorporação de variáveis explicativas.

- A aplicação dos modelos VAR e MIDAS com dados de frequência mista não foi encontrado em nenhum estudo, pelo que a pesquisa nesse aspeto é encorajada. Por exemplo, o VAR, segundo trabalha apenas com as séries de frequências mais altas considerando as mais baixas como sendo *outliers*.
- A combinação dos métodos de julgamento com os modelos de previsão quantitativa, do pouco que tem sido aplicado, tem mostrado promissora. Este tipo de combinação pode vir a ser uma forma alternativa de ajustar as previsões, já que, em média tais previsões ajustadas têm sido alcançadas de forma imparcial. Porém é necessário um estudo mais aprofundado sobre este assunto.
- A combinação de modelos tem vindo a ganhar popularidade, porém é necessário desenvolver técnicas mais complexas de combinação, métodos individuais de previsão mais avançados e uma forma de gerar previsões precisas independentemente do horizonte.
- Uma possível direção de investigação futura diz respeito às previsões combinadas. Isto é, o desenvolvimento de métodos inovadores de combinação e, também, o número ideal de modelos individuais a serem combinados.
- Como não foi observado, nos estados da arte analisados, nenhuma aplicação da versão multivariada do método SSA, é necessário haver trabalhos futuros a fim de examinar sua precisão de previsão. Ainda desafia os investigadores a considerarem outros métodos espectrais como o *multitaper* e entropia máxima.
- Dadas as características não lineares da procura turística, os modelos STR podem vir a ser muito utilizados por serem tecnicamente muito vantajosos. Ainda, estudos devem ser feitos no sentido de descobrir regras de relacionamentos não lineares entre as variáveis da procura turística e seus determinantes com base em técnicas de IA.
- Devido a um grande volume de dados sobre o turismo disponível na Internet, espera-se um grande desenvolvimento dos modelos baseados em IA, pois esta metodologia tem demonstrado grande capacidade em lidar com *Big Data*, embora esteja em sua fase inicial.
- Desenvolvimento de novos modelos, técnicas e métodos de previsão e estimativas para trabalhar junto os modelos de séries temporais com o *Big Data*, bem como

conciliar as pesquisas de previsão turística com as teorias de comportamento do consumidor do destino.

- Dados oriundos de pesquisas no Google e no Baidu têm mostrado ser muito úteis na previsão da procura turística, pelo que sua popularidade e uso farão grande diferença num futuro próximo.
- Na era do *Big Data* o aspecto computacional faz toda a diferença. Assim, sinergias devem ser criadas com investigadores das áreas de computação e engenharia de forma a obter-se mais eficiência e menor custo computacional na procura de previsões mais precisas.
- Desenvolver mais pesquisas para compreender melhor a consistência do desempenho da previsão dos modelos *fuzzy* de séries temporais.
- O desafio permanece na tentativa de generalizar os modelos econométricos. Resultados até então sugerem que a precisão das previsões depende, por exemplo, do tipo de destino, porém é necessário trabalhar mais as questões dos indicadores da procura.
- Crises e desastres trazem grandes prejuízos ao sector do turismo, pelo que se considera necessário o desenvolvimento de métodos que levem em conta esses eventos imprevisíveis na previsão da procura turística.
- Considerar e pesquisar novas variáveis não tradicionais com o apoio de teorias de diferentes áreas do saber para explicar a procura turística. Uma outra opção é usar técnicas de modelação de frequência mista.
- Obter um modelo universal capaz de superar todos os outros modelos, existentes até então, quanto à precisão e previsão da procura turística, não contrariando Song e Li (2008) que admite não haver um modelo supremo para todas as condições, mas que seja para as condições principais.

Em Cabo Verde não tem havido investigação a nível da previsão da procura turística utilizando as técnicas, métodos e modelos descritos neste estado da arte. Assim, e no intuito de iniciar este tipo de investigação, e dar um passo de cada vez, convinha começar, neste trabalho, por explorar os modelos de séries temporais mais conhecidos e comumente utilizados em séries que possuem características sazonais (como o SARIMA e

o Holt-Winters) e os modelos das redes neuronais, para futuramente seguir as tendências, fazer as devidas comparações entre modelos e lançar aos desafios.

3.4. SÉRIES TEMPORAIS

3.4.1. Definição

O que é uma *série temporal*? Segundo Ehlers (2005) é um conjunto de observações dispostas ao longo do tempo de forma sequencial. Mas essa sequência, no entender de Morettin (1981), não tem que ser, necessariamente e igualmente espaçada. Reis (2014) diz ser em intervalos sucessivos e/ou regulares. Quanto ao *tempo*, este pode ser assumido como uma variável do tipo contínuo ou discreto, sendo este último, segundo Kirchgassner e Wolters (2007) o mais geralmente assumido.

Mas, para que uma determinada série seja classificada como uma série temporal é necessário que preencha outro pré-requisito: os dados também devem apresentar uma dependência serial entre eles. Ainda, existem duas formas de se analisar uma série temporal, que segundo Morettin e Tolo (2006) são: a análise no domínio da frequência e a análise no domínio do tempo, sendo este o utilizado neste trabalho.

De forma mais formal, Shumway e Stoffer (2016), considera uma série temporal como um conjunto sequencial de variáveis aleatórias, X_1, X_2, X_3, \dots , onde cada uma dessas variáveis possuem valores que correspondem a um determinado período de tempo $t = 0, \pm 1, \pm 2, \dots$. No geral, este conjunto de variáveis $\{X_t\}$ indexadas ao tempo é referido como um processo estocástico (ver secção 3.4.5). Segundo Maddala (1992), normalmente denota-se o $\{X_t\}$ por X_t se for discreta e por $X(t)$ se for contínua.

Neste presente trabalho assume-se o tempo como uma variável do tipo discreta, devido ao período de observação dos dados da variável em estudo (Número de dormidas mensais), registados entre os anos 2000 e 2018; e portanto é utilizada a notação X_t . Na prática, segundo Maddala e Lahiri (2009), em uma série temporal discreta X_t , existe a possibilidade de infinitas observações x no tempo t indexadas pela data de cada

observação e optou-se por trabalhar com apenas uma amostra de tamanho n , $\{x_1, x_2, \dots, x_n\}$, isto é:

$$\{X_t\}_{t=-\infty}^{\infty} = \left\{ \dots, x_{-1}, x_0, \underbrace{x_1, x_2, x_3, \dots, x_n}_{\substack{\text{amostra de} \\ \text{tamanho } n}}, \dots \right\} \quad (6)$$

3.4.2. História

Resumidamente, a história parece começar com o trabalho de George Udny Yule (1927), que segundo Xavier (2016) deu uma grande contribuição no desenvolvimento de ferramentas e métodos para a análise de séries temporais. Mas Kirchgassner e Wolters (2007) relatam que Evgenij Evgenievich Slutsky também deu esses passos e que esses dois estatísticos desenvolveram o modelo auto-regressivo (AR) para representar as séries temporais, na qual o valor previsto dependia dos valores passados (Xavier, 2016). Em 1938, Herman Wold sistematizou e generalizou essas abordagens e em 1970 George E.P. Box e Gwilym M. Jenkins desenvolveram métodos empíricos para aplicar esses modelos (Kirchgassner & Wolters, 2007).

3.4.3. Importância

A importância das séries temporais pode ser vista nos vários ramos de aplicação ou campos de conhecimento. Ehlers (2005) enumera alguns: na Economia (*e.g.* preços diários de ações, taxa mensal de desemprego, produção industrial, ...), na Medicina (*e.g.* eletrocardiograma, eletroencefalograma, ...), na Epidemiologia (*e.g.* número mensal de novos casos de uma determinada doença, ...), na Meteorologia (*e.g.* velocidade do vento, temperatura diária, ...), etc. Mas, a sua aplicação tem sido mais notada no campo da econometria (Kirchgassner & Wolters, 2007). Entretanto aplicações no estudo da previsão da procura turística é também uma realidade como se pode constatar no estado da arte. Nesses campos de conhecimento é importante: (i) encontrar um modelo estatístico apropriado que possa explicar o comportamento da série num determinado período observado; (ii) estimar valores futuros, tendo como base os valores passados (Xavier, 2016).

3.4.4. Padrões

Segundo Reis (2014), os dados das séries temporais podem ser compostos de quatro padrões ou componentes: Tendência (que indica o comportamento da série a longo prazo), Ciclicidade (ciclos nas séries que ocorrem com alguma periodicidade de duração superior a um ano, em geral uma década ou mais); Sazonalidade (oscilações nas séries que ocorrem num determinado período específico, geralmente não superior a um ano); e, Irregularidade (oscilações inexplicáveis, fruto, por exemplo, de catástrofes naturais). Esta irregularidade é conhecida como componente restante, aleatória ou erro, que segundo Morettin e Toloi (2006) é de média zero e variância constante.

Esses componentes podem ser relacionados com a variável da série através do modelo aditivo (7) ou multiplicativo (8):

$$X_t = S_t + T_t + E_t \quad (7)$$

ou

$$X_t = S_t \times T_t \times E_t \quad (8)$$

onde

X_t – valor da série no tempo t ; S_t – componente sazonal no tempo t ; T_t – componente do ciclo de tendência no tempo t ; E_t – componente restante no tempo t .

O modelo aditivo é preferível caso a oscilação no padrão sazonal não varie com o nível da série temporal (entende-se como nível da série o valor ou faixa típica de valores que a variável pode assumir, se não for observado comportamento crescente ou decrescente no longo prazo), ou seja, as variações periódicas têm uma amplitude que se mantém constante; enquanto o modelo multiplicativo é preferível quando a variação no padrão sazonal parece ser proporcional ao nível da série (Reis, 2014).

3.4.5. Conceitos Gerais sobre Processos Estocásticos

Um dos objetivos da análise de séries temporais consiste em identificar o tipo de evolução da série ao longo do tempo do fenómeno em estudo e matematicamente

modelar essa evolução. Tal fenómeno depende do tempo e do acaso e o modelo matemático que o descreve é denominado de *processo estocástico*.

Um processo estocástico, de notação $\{X_t, t \in T\}$, é uma sucessão de variáveis aleatórias indexadas no tempo (ou no espaço) pelo parâmetro t . Neste trabalho é dado ênfase ao tempo. Entende-se variável aleatória (v.a.) como uma função $X(w)$ que associa a cada acontecimento $w \in \Omega$ (Ω é o espaço amostral) um valor pertencente a \mathbb{R} . Entende-se que o valor assumido por X depende do ponto w .

De modo semelhante a do conceito de v.a., um processo estocástico é uma função $X_t(w)$ que associa a cada acontecimento $w \in \Omega$ e a cada ponto $t \in T$ um valor E a que se chama *espaço dos estados do processo*. A T chama-se *espaço parâmetro*.

Uma forma de descrever um processo estocástico é através da distribuição de probabilidade conjunta das n v.a. $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ para qualquer conjunto de tempos t_1, t_2, \dots, t_n pertencentes a T (Muller, 2007), que segundo Ehlers (2005) não seria na prática uma tarefa fácil, sendo tal descrição feita geralmente através das funções de média μ , variância σ^2 e autocovariância γ :

$$\mu_t = E(X_t) \quad (9)$$

$$\sigma_t^2 = Var(X_t) \quad (10)$$

$$\gamma(t_1, t_2) = E(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2}) \quad (11)$$

Se o tempo t for fixo, então $X_t(w) \equiv X(w) \equiv X$ é uma v.a., e se $X_t(w) \equiv X(t)$ estamos diante de uma *realização* ou *trajetória* do processo.

O tempo, ou um conjunto qualquer de índices T bem como o espaço de estados E podem ser discretos ou contínuos. Assim, os processos estocásticos classificam-se:

- Quanto à natureza do espaço parâmetro – se T é finito ou infinito numerável, na qual se considera $T \equiv \mathbb{Z} = \{1, \dots, t\}$, o processo estocástico diz-se ser de *parâmetro discreto*. Caso T for infinito não-numerável, diz-se ser de *parâmetro contínuo*;

- Quanto à natureza do espaço de estados – se E é finito ou infinito numerável, o processo estocástico diz-se *discreto*. Caso E for infinito não-numerável, o processo estocástico diz-se *contínuo*;
- Quanto ao tipo de relação de dependência entre as suas variáveis – um processo estocástico pode ser estacionário (forte ou fraco) ou não estacionário.

Dentro dos processos estocásticos existem os chamados processos estacionários que servirão de base para a análise de estacionariedade das séries temporais abordado mais à frente. A seguir duas definições (3.4.1 e 3.4.2) formais sobre a estacionariedade forte e fraca (Muller, 2007; Wei, 2006; Ehlers, 2005):

Definição 3.4.1

Um processo estocástico $\{X_t, t \in T\}$ diz-se fortemente (ou estritamente) estacionário se $\forall n \geq 1, \forall k \geq 0$ tal que $t_i + k \in T, i = 1, \dots, n$ e qualquer escolha de índices $(t_1, t_2, \dots, t_n) \in T$ a distribuição dos vectores $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ e $(X_{t_1+k}, X_{t_2+k}, \dots, X_{t_n+k})$ é idêntica, ou seja, possuem a mesma distribuição conjunta de probabilidades.

Definição 3.4.2

Um processo estocástico $\{X_t, t \in T\}$ diz-se fracamente estacionário, ou de segunda ordem, ou seja, preserva suas características ao longo do tempo, $X_t \sim \text{Dist}(\mu, \sigma^2)$, se para todo o $t, E[X_t^2] < +\infty, \forall k \geq 0$ e

1. $E[X_t] = \mu(\text{constante}), t \in T \equiv \mathbb{Z} = \{1, \dots, t\}$
2. $\text{Var}[X_t] = \sigma^2(\text{constante}), t \in T$
3. $\text{Cov}(X_t, X_s) = \text{Cov}(X_{t+k}, X_{s+k}) = \gamma(|t - s|), t, s \in T; \text{ ou seja, depende da distância temporal entre as variáveis.}$

Segundo Ehlers (2005) esta segunda definição é a mais utilizada na prática, devido à sua facilidade de verificação. Poderíamos citar como um exemplo de processo estocástico estacionário, o chamado *Ruído Branco* que se define a seguir (Ehlers, 2005):

Definição 3.4.3

Um processo estocástico $\{X_t, t \in T\}$ diz-se um processo de Ruído Branco (RB) se:

1. $E[e_t] = \mu_e$
2. $Var[e_t] = \sigma_e^2$
3. $Cov(e_t, e_{t+k}) = \gamma(k) = 0, k = \pm 1, \pm 2, \dots$

Geralmente, neste tipo de processo, considera-se $\mu_e = 0$, sendo abreviadamente representado como $e_t \sim RB(0, \sigma_e^2)$.

3.4.6. Autocorrelação

A função de autocorrelação (**ACF** – *Autocorrelation Function*) fornece correlações entre a série X_t e os valores defasados da série para *lags* de 1, 2, 3, ... períodos de tempo. É usada geralmente para determinar se a série temporal é estacionária ou não. Uma série temporal estacionária fará com que a autocorrelação amostral caia a zero rapidamente, mas para uma série não estacionária ela cai de forma lenta. Segundo Ehlers (2005) a função de autocorrelação (ACF) tem maior efeito nas séries estacionárias e é dada pela expressão (12), considerando os pares de valores $(x_1, x_2), \dots, (x_{n-1}, x_n)$, sendo x_1, \dots, x_{n-1} e x_2, \dots, x_n como duas variáveis, obtendo assim, à semelhança da correlação de Pearson, a correlação entre x_t e x_{t+1} :

$$r_1 = \frac{\sum_{t=1}^{n-1} (x_t - \bar{x}_1)(x_{t+1} - \bar{x}_2)}{\sqrt{\sum_{t=1}^{n-1} (x_t - \bar{x}_1)^2 \sum_{t=1}^{n-1} (x_{t+1} - \bar{x}_2)^2}} \quad (12)$$

onde as médias amostrais são dadas por

$$\bar{X}_1 = \sum_{t=1}^{n-1} x_t / (n - 1) \text{ e } \bar{X}_2 = \sum_{t=2}^n x_t / (n - 1) \quad (13)$$

Uma generalização e simplificação para calcular a correlação amostral entre os valores defasados de k períodos de tempo é dado por:

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}; \quad -1 \leq r_k \leq 1 \quad (14)$$

Geralmente, na prática, os coeficientes de autocorrelação são obtidos por meio de

$$r_k = c_k/c_0, \quad k = 1, 2, \dots \quad (15)$$

onde $c_0 = Var(x_t)$ e os primeiros coeficientes de autocovariância $\{c_k\}$ são dados por

$$c_k = Cov(x_t, x_{t+k}) = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{n} \quad (16)$$

sendo que c_k é o estimador do coeficiente de autocovariância teórico γ_k .

A função de autocorrelação parcial (**PACF** – *Partial Autocorrelation Function*) é a correlação da série temporal com um atraso de si mesmo, com a dependência linear de todos os desfasamentos entre eles removidos. É definida pela correlação entre as variáveis x_t e x_{t+k} sem qualquer influência das variáveis intermédias $x_{t+1}, \dots, x_{t+k-1}$ e é apresentada pela expressão (17).

$$r(k, k) = \frac{Cov(x_t, x_{t+k} | x_{t+1}, \dots, x_{t+k-1})}{\sqrt{Var(x_t | x_{t+1}, \dots, x_{t+k-1})} \sqrt{Var(x_{t+k} | x_{t+1}, \dots, x_{t+k-1})}} \quad (17)$$

3.4.7. Correlograma

O correlograma é um gráfico que apresenta os k primeiros coeficientes de autocorrelação em função de k (Cryer & Chan, 2008). Permite identificar algumas características da série temporal, como a aleatoriedade, estacionariedade e sazonalidade. Em caso de aleatoriedade espera-se que $r_k = 0$. Aqui convém estabelecer formalmente as hipóteses:

H_0 : se $r_k = 0$, as observações da série são independentes

H_1 : se $r_k \neq 0$, as observações da série não são independentes

A Figura 3.1 é um exemplo de dois correlogramas que representa as funções de autocorrelação simples (ACF) e parcial (PACF). As linhas horizontais tracejadas (cor azul) indicam os limites inferiores e superiores, calculados para um determinado nível de confiança; por exemplo, quando o coeficiente de autocorrelação amostral é normalmente distribuído, os limites de confiança aproximados de 95% são definidos por $\pm Z_{\alpha/2}/\sqrt{n}$ relativamente à linha horizontal zero (Cryer & Chan, 2008).

Assim, todos os valores em cada k , quer da ACF, quer da PACF, que estiverem fora dos limites são considerados diferentes de zero, e os que estiverem dentro são considerados nulos. Se a maioria dos valores estiverem dentro dos limites é uma indicação da não rejeição da hipótese nula de que as observações são independentes. No caso da Figura 3.1, o $n = 228$ e portanto os limites são aproximadamente $\pm 0,13$. Ainda no correlograma podemos constatar dois tipos de correlação:

- *As correlações de curto-prazo* – quando existe um valor alto de r_1 seguido por valores que tendem a ficar sucessivamente menores aproximando-se de zero a partir de uma certa defasagem k . Caso esta aproximação a zero tende a decair de forma lenta é uma indicação que a série é não estacionária. Na Figura 3.1, no gráfico que corresponde a ACF constata-se que os valores r_k estão com tendência para decair para zero lentamente, mas ainda em defasagens significativas. Tal ocorre devido a presença da componente tendência na série, o que é fácil visualizar pois as observações seguem todos do mesmo lado da média. Este decaimento lento dos valores também é um indicativo da utilização da componente MA; caso o decaimento for exponencial, é um sinal da utilização da componente AR. Na ACF o decaimento exponencial, caso exista, acontece após a defasagem q (ordem não sazonal da componente MA) enquanto que na PACF o mesmo comportamento acontece após a defasagem p (ordem não sazonal da componente AR). De salientar que o ACF é útil para localizar o valor de q e o PACF para identificar o valor de p ;
- *As correlações negativas* – quando existe uma alternância entre os r_k , estando uns acima e outros abaixo da linha zero.

Com relação à sazonalidade, o correlograma irá permitir visualizar as flutuações sazonais na mesma frequência da série.

Uma visualização gráfica do ACF é útil também na identificação dos parâmetros do modelo SARIMA, de ordem não sazonal p, d, q e os de ordem sazonal P, D, Q , modelo este definido na subsecção 3.4.9.1. No caso de ter sido aplicada a diferenciação sazonal na série (definido na subsecção 3.4.10) e se verificar que o ACF possui um pico positivo é sinal

da utilização da componente AR sazonal; caso tiver um pico negativo utiliza-se um termo MA sazonal.

De salientar que a análise do correlograma da ACF e da PACF devem ser feitas em conjunto, sempre que justificar, para melhor interpretação a respeito das características da série, principalmente se houver termos autorregressivos no modelo.

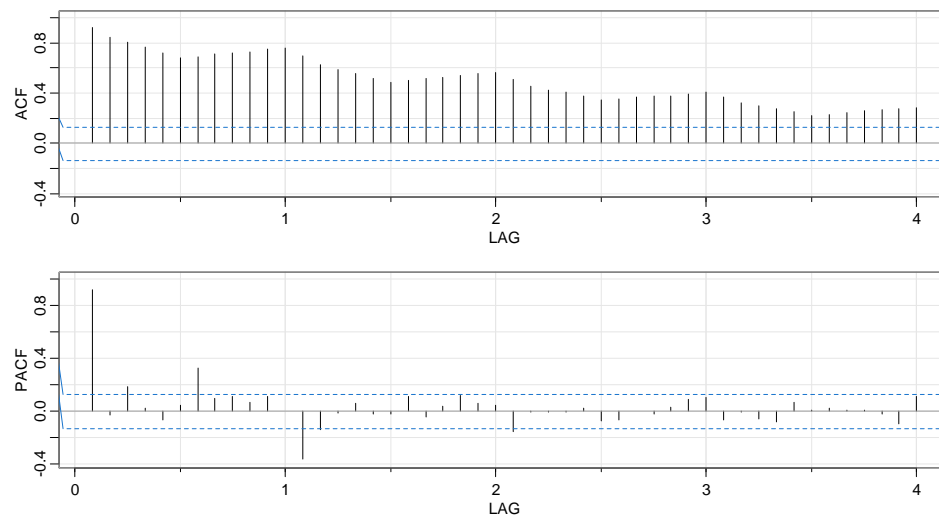


Figura 3.1: Exemplo de um correlograma.

Fonte: Elaboração própria (no R)

3.4.8. Sazonalidade

Segundo Xavier (2016) a sazonalidade de uma série temporal pode ser determinística (o padrão sazonal é regular e estável, sendo necessário introduzir *dummies* sazonais nos modelos para contabilizar as variações) ou estocástica (padrão sazonal varia com o tempo, sendo necessário que a série seja diferenciada sazonalmente de forma a contabilizar as raízes da unidade sazonal). Entretanto, para Kim e Moosa (2001) não existem evidências conclusivas que nos permitem optar por um tratamento determinístico em detrimento do estocástico e vice-versa.

Chen *et al.* (2017) observam que o tratamento determinístico é o mais considerado nos métodos tradicionais de previsão no contexto turístico, porém métodos mais recentes já

sugerem a sazonalidade estocástica na modelação da previsão da procura turística, nomeadamente o método Box-Jenkins e seu modelo SARIMA. Para decidir sobre a presença da componente sazonal na série geralmente são utilizados os testes de Kruskal-Wallis, Friedman e o teste F. Nestes testes caso o $p\text{-value} < \alpha$ (α – nível de significância) então se pode concluir sobre a presença da sazonalidade na série.

Com base em Ollech (2021), o teste de Friedman e o teste F são específicos para o contexto de *dummies* sazonais, enquanto o teste de Kruskal-Wallis e o teste WO não. Assim, para este trabalho se poderia considerar estes dois últimos testes, porém o teste WO é o escolhido e utilizado mais pela função *isSeasonal* que o suporta. Esta função permite colocar em seu argumento *test* os outros testes aqui mencionados e além do mais possui menos argumentos do que a função que retorna o resultado do teste de Kruskal-Wallis e por isso mais simples de implementar. O comando básico é dado por (Ollech, 2021):

$$isSeasonal(x, test = "wo", freq = NA)$$

onde x é a série temporal, **test** é o teste de Webel-Ollech a utilizar e **freq** é a frequência da série temporal. Neste teste caso o $p\text{-value} < \alpha$ então se pode concluir sobre a presença da sazonalidade na série.

No contexto turístico, a sazonalidade, segundo Butler (1994, 2001) pode ser considerada como quaisquer altos e baixos sistemáticos ao longo do tempo em atividades específicas do turismo, incluindo chegadas de turistas, gastos por turistas, visitas a destinos específicos, etc. Para Bidermarn (2008), uma característica permanente da sazonalidade é a mudança repentina da procura em um determinado período do ano.

Butler ainda apontou duas causas importantes de sazonalidade no turismo, a saber: (1) **natural**, resultado de variações regulares nas condições climáticas, como temperatura, precipitação, neve e luz do dia e (2) **institucional**, resultado de decisões humanas numa combinação de fatores, por exemplo, religiosos, culturais, étnicos e sociais.

Além das causas, a sazonalidade produz alguns impactos na procura turística. Carnas (2012) agrupou os impactos em três categorias principais, a saber: (i) **impactos**

económicos – como a perda de receitas/lucros devido ao uso não económico de recursos e outras instalações do destinatário (Sutcliffe & Sinclair, 1980; Marning & Powers, 1984; William & Shaw, 1991), dificuldades com relação a questões de recrutamento e retenção de pessoal em tempo integral no setor de viagens e turismo (Yacoumis, 1980) e consequente manutenção de padrões de qualidade do produto (Baum, 1999), oportunidades de emprego temporário para algumas pessoas, como estudantes, artistas e donas de casa (Carnas, 2012); (ii) **impactos socioculturais** – para Marning e Powers (1984) crimes, congestionamento, preços altos (apenas para citar alguns) colocam "uma pressão sobre a capacidade de suporte social do destino", o que pode "resultar em ressentimento da comunidade local em relação a todas as atividades turísticas"; (iii) **impactos ecológicos** – como erosão física de caminhos e outros recursos naturais, problemas de lixo, perturbação da vida selvagem e congestionamento de áreas rurais (Carnas, 2012) devido ao alto grau de concentração de turistas durante a temporada alta nos destinos.

3.4.9. Modelos de Séries Temporais

Segundo Song e Li (2008) o modelo SARIMA tem vindo a ganhar popularidade pelo fato de que muitos decisores têm interesse em conhecer a variação sazonal da procura turística. E sendo uma versão sazonal do modelo ARIMA, primeiramente é abordado este tipo de modelo e seguidamente vai detalhar-se um pouco o modelo em foco – o SARIMA.

O modelo ARIMA é um modelo que, segundo Hyndman e Athanasopoulos (2018), possui uma característica distinta de outros modelos (e métodos) pelo fato de atribuir os dados próximos uma maior relação entre si do que dados que se encontram distantes uns dos outros por grandes intervalos de tempo. Esta característica é denominada autocorrelação serial, que seria o grau de associação linear entre os valores defasados da série, e que é utilizado para prever os valores futuros da série baseado em seus valores passados e presente. Por exemplo a autocorrelação na defasagem 1 seria a associação entre X_t e X_{t-1} , na defasagem 2 entre X_t e X_{t-2} e assim sucessivamente. Este tipo de modelo, pertencente à metodologia de Box-Jenkins, é capaz de descrever de maneira satisfatória séries estacionárias e não estacionárias.

Segundo Xavier (2016) é necessário identificar quais as ordens para cada parâmetro do modelo ARIMA que melhor se ajustam aos dados, antes de se avançar com a modelação ARIMA. Para tal, recorre-se geralmente às funções de autocorrelações (ACF) e autocorrelações parciais (PACF) estimadas. O modelo utiliza os termos autorregressivo (AR), integrado (I) e de média móvel (MA), sendo que os termos AR correspondem a defasagem da série e MA é a defasagem dos erros aleatórios. Segundo Lemos (2006) o ARIMA é um modelo que tende a ajustar bem os dados com um número mínimo de parâmetros.

Antes de descrever cada um dos modelos que compõe o ARIMA, convém conhecer alguns parâmetros a ser utilizados: entende-se por p o número de termos autorregressivos, d o número de diferenciações e q o número de termos de médias móveis; p , d e $q \geq 0$.

Assim, um modelo **AR** (p) não sazonal de ordem p é dado por:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t \quad (18)$$

onde ϕ_i é o parâmetro que descreve a relação da série X_t com X_{t-i} , $i = 1, 2, \dots, p$, e ε_t o erro aleatório, em que X_{t-1}, X_{t-2}, \dots , é independente do ε_t .

Neste caso, esses parâmetros devem ser inferiores a unidade $|\phi_i| < 1$ e as autocovariâncias (γ_k) sejam independentes para que o modelo seja estacionário.

Um modelo **AR** (P) sazonal de ordem P e período sazonal s é dado por (Cryer & Chan, 2008):

$$X_t = \Phi_1 X_{t-s} + \Phi_2 X_{t-2s} + \dots + \Phi_P X_{t-Ps} + \varepsilon_t \quad (19)$$

sendo que o seu polinómio característico é dado por

$$\Phi(B^s) = 1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_P B^P \quad (20)$$

onde Φ_i é o parâmetro que descreve a relação da série X_t com X_{t-is} , $i = 1, 2, \dots, P$, e ε_t o erro aleatório, em que X_{t-s}, X_{t-2s}, \dots , é independente do ε_t . Para ser estacionário, as raízes de $\Phi(B^s)$ devem ser, em valor absoluto, maiores que 1, ou seja, $|\Phi_i| > 1$.

Segundo Etuk (2012), neste modelo a PACF é interrompida após o *lag p* e o ACF é uma combinação de senoidais que caem lentamente.

Um modelo de médias móveis **MA (q)** não sazonal de ordem **q** é dado por:

$$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \dots + \theta_q\varepsilon_{t-q} \quad (21)$$

onde θ_i é o parâmetro que descreve a relação da série X_t com ε_{t-i} , $i = 1, 2, \dots, q$.

Um modelo de médias móveis **MA (Q)** sazonal de ordem **Q** e período sazonal **s** é dado por (Cryer & Chan, 2008):

$$X_t = \varepsilon_t - \Theta_1\varepsilon_{t-s} - \Theta_2\varepsilon_{t-2s} - \dots - \Theta_Q\varepsilon_{t-Qs} \quad (22)$$

sendo que o seu polinómio característico dado por

$$\Theta(B^s) = 1 - \Theta_1B - \Theta_2B^2 - \dots - \Theta_QB^Q \quad (23)$$

Neste tipo de modelo a série é sempre estacionária e a ACF é diferente de zero apenas nas defasagens sazonais de $s, 2s, 3s, \dots, qs$.

Segundo Etuk (2012), neste modelo o ACF é interrompido após o *lag q* e o PACF cai lentamente.

O modelo ARMA (p, q) incorpora uma parte AR(p) e uma parte MA(q) e é dado pela expressão (24).

$$X_t = \phi_1X_{t-1} + \phi_2X_{t-2} + \dots + \phi_pX_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \dots + \theta_q\varepsilon_{t-q} \quad (24)$$

Quando se inclui o termo de ordem de integração o modelo utilizado é o ARIMA (p, d, q). Os modelos ARIMA são identificados através das ACF e PACF (Xavier, 2016) e são construídos seguindo três etapas (Reis, 2014): (i) identificação; (ii) estimação; e, (iii) verificação, descritos mais adiante neste capítulo. Caso o modelo conseguido não for satisfatório as etapas são repetidas. Tendo chegado a um modelo satisfatório, o próximo passo é a realização das previsões.

3.4.9.1. Modelo SARIMA

O modelo SARIMA, ou também conhecido como ARIMA sazonal, incorpora fatores não sazonais e sazonais num modelo multiplicativo representado, simbolicamente, na forma de $SARIMA(p, d, q) \times (P, D, Q)_s$; onde p é a ordem autorregressiva não sazonal, d é a diferenciação não sazonal, q é a ordem da média móvel não sazonal, P é a ordem autorregressiva sazonal, D é a diferenciação sazonal, Q é a ordem da média móvel sazonal, e s é o intervalo de tempo de repetição do padrão sazonal. Segundo Cryer e Chan (2008) para uma série temporal mensal, $s=12$.

Este modelo é popular devido à sua capacidade de lidar com séries estacionárias e não estacionárias. No caso de séries temporais não estacionárias, os valores da variável são tomados em suas primeiras diferenças para estimar o modelo. Formalmente diz-se que uma série temporal $\{X_t\}$ segue o modelo SARIMA se for definido como (Mishra, Rout & Pradhan, 2018):

$$\Phi(B^s)\phi(B)(1 - B^s)^D(1 - B)^dX_t = \Theta(B^s)\theta(B)e_t, \quad t \in T \quad (25)$$

em que,

$\phi(B) = 1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p$ é o polinómio característico da componente autorregressivo não sazonal;

$\Phi(B^s) = 1 - \Phi_1B - \Phi_2B^2 - \dots - \Phi_pB^p$ é o polinómio característico da componente autorregressivo sazonal;

$\theta(B) = 1 + \theta_1B + \theta_2B^2 + \dots + \theta_qB^q$ é o polinómio característico da componente média móvel não sazonal;

$\Theta(B^s) = 1 + \Theta_1B + \Theta_2B^2 + \dots + \Theta_QB^Q$ é o polinómio característico da componente média móvel sazonal;

$(1 - B^s)^D(1 - B)^d$ é a diferenciação sazonal de ordem d e diferenciação não sazonal de ordem D , e B é o operador de retardo ($B^iX_t = X_{t-i}$).

Definição 3.4.4.

Um processo estocástico $\{X(t), t \in T\}$ admite uma representação mista integrada autoregressiva e médias móveis e de componente sazonal de período s de ordens (p, d, q) e (P, D, Q) , abreviadamente SARIMA $(p, d, q) \times (P, D, Q)_s$, se existirem números reais $(\phi_1, \phi_2, \dots, \phi_p)$, $(\theta_1, \theta_2, \dots, \theta_q)$, $(\Phi_1, \Phi_2, \dots, \Phi_P)$, $(\Theta_1, \Theta_2, \dots, \Theta_Q)$ e um processo de ruído branco $\{\varepsilon_t, t \in T\}$ de variância $\sigma_\varepsilon^2 > 0$, tais que o processo X_t verifica a equação (25) (Pires, 2001).

3.4.10. Análise de estacionariedade

Segundo Morettin e Tolo (2006) a estacionariedade de uma série temporal tem a ver com a seguinte suposição “(...) ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável” (Morettin & Tolo, 2006, p. 4). Chatfield (2000) afirma que a maioria das séries temporais é estocástica, mas, segundo Morettin e Tolo (1981), são estacionárias se, em diferentes fases, mantiverem a mesma média, variância e autocovariância. Mas é não-estacionária se a média e/ou a variância variarem com o tempo.

O comportamento temporal da série poderá ser analisado graficamente, tendo em conta a variação dos dados ao redor de uma linha horizontal ao longo do tempo. Caso se verifique a não-estacionariedade procede-se à sua transformação utilizando a **técnica de diferenciação**. Isto consiste em tomar, geralmente, uma (ΔX_t) , duas $(\Delta^2 X_t)$ ou mais $\Delta^d X_t$ diferenças sucessivas da série original até obter uma série estacionária.

Citando Maddala e Lahiri (2009) e Reis (2014) a 1ª diferença é dada por

$$\Delta X_t = X_t - X_{t-1} \tag{26}$$

e a 2.ª diferença dada por

$$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1} = X_t - 2X_{t-1} - X_{t-2} \tag{27}$$

No caso do modelo sazonal, o SARIMA, essa primeira diferença é dada por

$$\Delta_s X_t = X_t - X_{t-s} \tag{28}$$

em que Δ é o operador diferença.

Computacionalmente, a função `ndiffs()` e `nsdiffs()` permite determinar quantas diferenças não sazonais e sazonais, respetivamente, são necessárias para obter a estacionaridade da série. Estas funções fazem parte do *Software R* versão 4.0.5, em uso neste trabalho, e descrito em Hyndman *et al.* (2021).

Entretanto, segundo Cryer e Chan (2008), esta técnica tem sua desvantagem, pois ao realizar as diferenças sucessivas os dados são perdidos; isto é, ao se aplicar, por exemplo, diferenciações numa série mensal de comprimento N cuja componente sazonal é de período s vai ter-se uma série de diferenças sazonais de comprimento $N - s$.

3.4.11. Identificação do modelo

Eis alguns passos importantes da metodologia de Box-Jenkins a ter em conta no processo de identificação de um modelo SARIMA (Shumway & Stoffer, 2016; Etuk, 2012; Cryer & Chan, 2008; Pires, 2001):

1. Criar uma representação gráfica da série temporal em estudo, para uma inspeção cuidada;
 - Em caso de haver um aumento de variação ao longo do tempo, é preferível estabilizá-la, utilizando por exemplo as transformações de Box-Cox ou a transformação logaritmica $\log(X_t)$. Com relação a Box-Cox, a decisão de estabilizar ou não a variância da série é avaliada, computacionalmente, pelo valor do parâmetro λ (*Lambda*) de ser, ou não, igual a 1. Caso $\lambda = 1$ então a transformação não é necessária. Matematicamente a transformação de Box-Cox é definida por

$$g(X_t) = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log(X_t) & \text{se } \lambda = 0 \end{cases} \quad (29)$$

sendo que, para efectuar a transformação, $\lambda \in [-1, 1[$.

- Pode fazer uso da visualização gráfica da ACF. Caso se confirme a não estacionariedade, aplicar a técnica de diferenciação simples. Esta diferença

de ordem d , $D^d = (1 - B)^d$ permite eliminar uma grande parte de tipo de tendências. Atualizar a visualização gráfica da ACF;

- Aplicar testes para identificar a integração sazonal (p.ex. o teste HEGY) e não sazonal (p.ex. o teste Augmented Dickey-Fuller).
 - Em caso de haver a componente sazonal, esta deve ser removida por meio da diferenciação sazonal. Esta diferença sazonal de período s e de ordem D , $D_s^D = (1 - B^s)^D$, permite eliminar os movimentos estritamente periódicos com período s .
 - A soma das ordens de diferenciação de d e D devem ser no máximo 2 (ou seja, $0 \leq d + D \leq 2$).
 - Caso a presença da componente sazonal estiver bem marcada na série, é preferível aplicar primeiro a diferenciação sazonal e só depois a simples.
 - Identificar a componente cíclica de período desconhecido e isolá-la da série durante o processo de estabilização da média. A constatação da componente cíclica é feita observando as flutuações não regulares em torno da tendência, isto é, se houver alternâncias sistemáticas entre valores maiores e menores do que 0 (ponto neutro no modelo aditivo) e do que 1 (ponto neutro no modelo multiplicativo) ao longo dos períodos, e se os valores permanecerem predominantemente maiores/menores do que 0 (ou 1) durante pelo menos um ano.
 - A utilização da função `decompose()` presente no *Software R* versão 4.0.5 permite identificar quais componentes estão presentes na série.
2. Depois de se obter uma série livre da tendência e da sazonalidade, é necessário identificar os valores preliminares dos parâmetros de ordem não sazonal p, d, q e os de ordem sazonal P, D, Q . Segundo Murteira (1993) os valores de p, q, P, Q raramente ultrapassam o valor 2.
 3. Obter o modelo sazonal provisório. Este modelo está sujeito à revisão na fase de diagnóstico.

Depois de identificar o modelo provisório, seguem outros passos:

4. Estimar os parâmetros do modelo provisório da forma mais eficiente possível;
5. Fazer o diagnóstico;
6. Escolher o modelo;

que serão descritos a seguir.

3.4.12. Estimação dos parâmetros

Existem diferentes métodos para estimar os parâmetros ϕ , Φ , Θ , θ e σ_ε^2 que a seguir se descrevem de uma forma resumida:

- **Método dos Momentos** – segundo Cryer e Chan (2008), o Método dos Momentos é um dos mais eficientes para estimar os parâmetros, porém não é muito conveniente para o modelo MA(q) já que geralmente não produz boas estimativas. Este método consiste em igualar os momentos da população aos momentos amostrais de forma a resolver as equações resultantes para obter as estimativas dos parâmetros (Cryer & Chan, 2008). No caso do modelo AR (p) as equações são

$$\begin{cases} \phi_1 + r_1\phi_2 + r_2\phi_3 + \dots + r_{p-1}\phi_p = r_1 \\ r_1\phi_1 + \phi_2 + r_1\phi_3 + \dots + r_{p-2}\phi_p = r_2 \\ \vdots \\ r_{p-1}\phi_1 + r_{p-2}\phi_2 + r_{p-3}\phi_3 + \dots + \phi_p = r_p \end{cases} \quad (30)$$

Essas equações lineares são resolvidas para $\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p$.

O estimador da constante μ e da variância do erro σ_ε^2 é dado por

$$\hat{\mu} = (1 - \hat{\phi}_1 - \hat{\phi}_2 - \dots - \hat{\phi}_p)\bar{X}_t \quad (31)$$

$$\hat{\sigma}_\varepsilon^2 = (1 - \hat{\phi}_1r_1 - \hat{\phi}_2r_2 - \dots - \hat{\phi}_pr_p)s^2 \quad (32)$$

onde s^2 a variância amostral e \bar{X}_t a média da série.

Considerando o modelo MA(1), o estimador do parâmetro θ é dado por

$$\hat{\theta} = \frac{-1 + \sqrt{1 - 4r_1^2}}{2r_1} \quad (33)$$

Se $r_1 = \pm 0,5$, soluções reais únicas existem. Mas para MA(q) de ordem superior, as estimativas θ_i são resolvidas por meio de métodos numéricos.

Pode ainda estimar σ_ε^2 dada por

$$\hat{\sigma}_\varepsilon^2 = \frac{s^2}{1 + \hat{\theta}_1^2 + \hat{\theta}_2^2 + \dots + \hat{\theta}_q^2} \quad (34)$$

- **O Método dos Mínimos Quadrados Condicionados** – com relação aos mínimos quadrados condicionais consideremos para exemplificação o modelo AR(p), em que o estimador dos parâmetros desconhecidos $\vec{v} = (\phi_1, \phi_2, \dots, \phi_p, \mu)$, dada uma série temporal de dimensão N, (X_1, X_2, \dots, X_N) , consiste em minimizar a soma dos quadrados

$$Q_n(\vec{v}) = \sum_{t=p+1}^n (X_t - \mu - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \dots - \phi_p X_{t-p})^2 \quad (35)$$

Particularizando a expressão (35)-ao modelo AR(1) temos os estimadores dos mínimos quadrados para ϕ , μ e σ_ε^2 :

$$\hat{\phi} = \frac{\sum_{t=2}^n (X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=2}^n (X_{t-1} - \bar{X})^2} \quad (36)$$

$$\hat{\mu} = \frac{1}{n-1} \sum_{t=2}^n X_t - \hat{\phi} \left\{ \frac{1}{n-1} \sum_{t=2}^n X_{t-1} \right\} \quad (37)$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2} \sum_{t=2}^n (X_t - \hat{\mu} - \hat{\phi} X_{t-1})^2 \quad (38)$$

- **O Método da Máxima Verossimilhança Condicionada** – quanto ao estimador de máxima verossimilhança condicionada, considerando o processo AR(p), a função de verossimilhança condicionada aos valores observados do processo escreve-se do seguinte modo

$$L(\vec{v}) = \left(\frac{1}{2\pi\sigma_\varepsilon^2} \right)^{\frac{n-p}{2}} \exp \left[-\frac{1}{2\sigma_\varepsilon^2} \sum_{t=p+1}^n \varepsilon_t^2 \right] \quad (39)$$

e portanto

$$\log(L(\vec{v})) = \frac{-(n-p)}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} Q_n(\vec{v}) \quad (40)$$

em que o logaritmo é máximo quando a soma dos quadrados $Q_n(\vec{v})$ for mínima.

O estimador da variância do erro é dado por (Cryer & Chan, 2008; Pires, 2001)

$$\hat{\sigma}_\varepsilon^2 = \frac{Q_n(\hat{\phi}, \hat{\mu})}{n} \quad (41)$$

3.4.13. Avaliação do diagnóstico

O diagnóstico tem como objetivo testar a qualidade do ajuste do modelo e Pires (2001) propõe cinco passos a saber:

1. Significância estatística dos parâmetros do modelo

Considerando os seguintes parâmetros estimados, por exemplo, para o modelo ARIMA, $\hat{\beta}_i = (\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p, \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_q)$, pretende-se saber qual(ais) deste(s) parâmetro(s) deve(m) ser eliminado(s) do modelo. Esta avaliação consiste em construir hipóteses sobre os parâmetros β_i

$$H_0: \beta_i = 0 \text{ vs } H_1: \beta_i \neq 0, \quad i = 1, 2, \dots, p + q + 1 \quad (42)$$

e realizar o teste baseada na distribuição amostral dos respetivos estimadores, sendo a estatística de teste dada por

$$T = \sqrt{n} \frac{\hat{\beta}_i}{\sqrt{\text{Var}(\hat{\beta}_i)}} \sim t_{(n-(p+q+1))}. \quad (43)$$

Quanto à regra de decisão, a hipótese nula H_0 é rejeitada caso $|T| > t_{(n-(p+q+1)), \alpha/2}$ para um determinado nível de significância α . Caso H_0 não seja rejeitada, implica a eliminação do parâmetro β_i do modelo em avaliação (Pires, 2001).

2. Estacionariedade do modelo

Uma das técnicas para verificar a estacionariedade é o uso dos testes estatísticos de raiz unitária, sendo o mais utilizado o de Dickey-Fuller (Reis, 2014).

Por exemplo, considerando um modelo AR(1) viu-se que para ser estacionário, as suas raízes devem ser, em valor absoluto, menores que 1, ou seja, $|\phi_i| < 1$. Se o valor da estimativa de máxima verosimilhança para o parâmetro ϕ for próximo de 1, neste caso para averiguar se o modelo é estacionário, a formulação das hipóteses seria (Shumway & Stoffer, 2016):

$$H_0: \phi = 1 \text{ vs } H_1: \phi < 1 \quad (44)$$

sendo a estatística de teste, baseada na distribuição amostral dos respetivos estimadores, dada por (Shumway & Stoffer, 2016; Pires, 2001)

$$T = \sqrt{\frac{(\hat{\phi} - \hat{\theta})^2}{(1 - \hat{\phi}^2)(1 - \hat{\phi}\hat{\theta})^2}} (\hat{\phi} - 1) \sim t_{n-2} \quad (45)$$

Quanto à regra de decisão, a hipótese nula H_0 é rejeitada caso $T < -t_{n-2}$ para um determinado nível de significância α . A nível computacional se rejeita H_0 se o *p-value* $< \alpha$.

Entretanto é aconselhável o uso de mais testes para se poder confirmar a estacionariedade, sendo os testes de Phillips-Perron e de KPSS (Kwiatkowski–Phillips–Schmidt–Shin) mais utilizados (Xavier, 2016).

3. Estabilidade do modelo

Neste passo o objetivo é verificar se o modelo é adequado para descrever o comportamento futuro da série. Para tal, faz-se uso da correlação entre os estimadores dos parâmetros do modelo. Caso as correlações forem, em valor absoluto, superiores a 0,7 é uma indicação de sobreparametrização do modelo e, portanto, modelos alternativos devem ser identificados (Pires, 2001). No *Software R* versão 4.0.5 (<https://www.r-project.org/>) o valor destas correlações está identificado no parâmetro ACF1.

Para evitar o problema de sobreparametrização (*overfitting*) é necessário ajustar modelos de ordens similares, porém mais baixas. Este processo costuma ser iterativo, isto é, na base da tentativa e erro seguindo algum dos critérios de informação apresentados no subtópico 3.4.14 (Escolhas de modelos).

4. Redundância dos parâmetros do modelo

Quando se tem, por exemplo, dois modelos em que um deles possui um número maior de parâmetros em relação ao outro, e a variância, o ACF, e o PACF são muito semelhantes, então está-se diante de uma redundância de parâmetros. Caso for, pelo princípio da parcimónia, escolhe-se o modelo com o menor número de parâmetros.

5. Avaliação da qualidade do modelo

Segundo Ehlers (2005), assim que for ajustado um modelo, deve-se verificar a sua adequação face aos dados. Para tal, é feita a verificação do comportamento dos resíduos, sendo estes definidos como a diferença entre os valores observados e os previstos. A nível gráfica, caso a dispersão dos resíduos se dê em torno da linha horizontal zero [se distribuem aleatoriamente em torno de zero] e sem nenhuma tendência observada [com variância aproximadamente constante], então é sinal de que o modelo pode ser adequado (Cryer & Chan, 2008). Neste passo, o modelo é avaliado diante de seus pressupostos teóricos e através do estudo dos resíduos estimados. Assim sendo, devem-se verificar três hipóteses sobre a série $\{\varepsilon_t, t \in T\}$, segundo Pires (2001):

a) *Se é uma série de ruído branco*

Para verificar esta hipótese, deve-se analisar a ACF e PACF dos resíduos e ver se as correlações estão ou não dentro dos limites (linhas a tracejadas). Caso estejam dentro dos limites, pode-se concluir que os resíduos se comportam como um ruído branco.

Ainda se pode verificar tal correlação usando a estatística de teste de Box-Pierce

$$Q = n \sum_{k=1}^m r_k^2 \quad (46)$$

em que m é o número de autocorrelações amostrais ($15 \leq m \leq 30$) e r_k é o coeficiente de autocorrelação amostral. Caso o modelo ajustado for satisfatório então $Q \hat{\sim} \lambda_{m-p-q}^2$. No caso, a formulação das hipóteses é dada por

$$\begin{aligned} H_0: r_1(\varepsilon) = r_2(\varepsilon) = \dots = r_m(\varepsilon) = 0 \\ \text{vs} \\ H_1: \exists r_k(\varepsilon) \neq 0, \quad k = 1, 2, \dots, m \end{aligned}$$

e a hipótese nula é rejeitada, segundo Pires (2001), para valores elevados da estatística de teste. A nível computacional se pode utilizar a função `Box.test()` do *Software R versão 4.0.5*, e se rejeita H_0 se o *p-value* $< \alpha$.

Outros testes, que avaliam se os valores observados dos resíduos são oriundos de uma série de variáveis independentes e identicamente distribuídas com distribuição absolutamente contínua, são (Pires, 2001): *o teste do sinal* (verifica a nulidade da mediana da distribuição dos resíduos), *o teste do sinal das diferenças* (aplica-se o teste do sinal à série das diferenças dos resíduos), *o teste dos instantes de inflexão* (identifica comportamentos de tipo não aleatório) e *o teste das ordens* (identifica a monotonia de longo prazo).

b) Se segue uma distribuição normal

Aqui geralmente faz-se uso da análise do cronograma da série das estimativas dos resíduos padronizados $\{\hat{\varepsilon}_t/S_{\hat{\varepsilon}}^2\}$, do histograma e do gráfico Q-Q *plot*. Caso não for detetado nenhum desvio acentuado relativamente ao modelo normal, é sinal da presença de normalidade. Entretanto o teste de ajustamento de Kolmogorov-Smirnov (K-S) ou o teste de Shapiro-Wilk (S-W) também pode ser aplicado (Pires, 2001).

Segundo Maroco (2003) e Hill e Hill (2005) pode-se utilizar o teste S-W no caso de amostras pequenas, inferior a 50. Neste trabalho é utilizado o teste K-S, pois o número de observações que é composta a variável “Número de dormidas” ultrapassa 50, ou seja são 228 observações.

Entretanto convém estar atento à questão da eficiência do teste, pois, segundo Zar (1999) e Torman, Coster e Ribolti (2012), em muitas situações, o teste K-S não consegue rejeitar a hipótese nula quando os dados realmente não são oriundos de uma distribuição normal.

Para formular as hipóteses do teste K-S, primeiro convém mencionar a ideia de distribuição empírica. Isto é, dada uma certa amostra (x_1, x_2, \dots, x_n) de uma determinada população X , o teste compara a distribuição de frequências acumuladas (representada por $F_0(x)$), que deveria ocorrer sob a suposta distribuição (sob a hipótese nula, H_0), com a distribuição de frequências acumuladas dos valores observados (representada por $S_n(x)$). Como a hipótese nula supõe que a amostra tenha sido obtida da distribuição $F_0(x)$, então espera-se que, para cada valor de X , a diferença entre $S_n(x)$ e $F_0(x)$ seja mínima. A estatística de teste toma o máximo destas diferenças em módulo e é denominada de desvio máximo, dada por

$$D = \max_{x \in \mathbb{R}} |F_0(x) - S_n(x)| \quad (47)$$

Assim, a formulação das hipóteses são:

$$H_0: S_n(x) = F_0(x)$$

$$H_1: S_n(x) \neq F_0(x)$$

Conhecendo a distribuição amostral de D e por meio da respectiva tabela de valores e dado um nível de significância α , a regra de decisão é a seguinte: Se o D calculado é menor que o D tabelado, então rejeita-se H_0 ; caso contrário aceita-se H_0 (Reis *et al.*, 2016).

c) *Se tem variância constante*

Para validar esta hipótese espera-se que, numa representação gráfica, os resíduos se distribuam aleatoriamente em torno zero com variância constante e correlação igual a zero. Tal seria um indício de um “bom” ajuste do modelo. Uma outra forma seria subdividir a série dos resíduos em diferentes intervalos e calcular a variância amostral

em cada um. Caso os resultados forem muito diferentes entre si é sinal de inconstância e, portanto, novos modelos devem ser identificados (Pires, 2001).

3.4.14. Escolha de modelos

Para selecionar o melhor modelo são utilizados alguns critérios de informação como **BIC** (Critério de Informação Bayesiano de Akaike), **HQC** (Critério de Informação Harnan-Quinn) e **AIC** (Critério de Informação de Akaike), sendo este o mais utilizado. Segundo Xavier (2016), a estimativa do AIC para um determinado modelo é dada por

$$AIC = -2L + 2m \quad (48)$$

em que m o número de parâmetros do modelo e L a log-verossimilhança máxima do modelo ajustado dado por

$$\ln L(\underline{\beta}; x_1, x_2, \dots, x_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \left| \sum_n \cdot \right| - \frac{1}{2} (\underline{x}_n - \underline{\mu}_n)' \sum_n^{-1} (\underline{x}_n - \underline{\mu}_n) \quad (49)$$

em que $\underline{\mu}_n$ e \sum_n são o vetor dos valores esperados e a matriz de covariâncias, respectivamente; $\underline{\beta}$ o vetor dos parâmetros do modelo com $\underline{x}_n = (x_1, x_2, \dots, x_n)$ (Pires, 2001).

Assim, o modelo que apresentar com menor valor AIC é considerado aquele que melhor se ajuste aos dados. Entretanto, a desvantagem deste critério é ser influenciado pelo número de parâmetros no modelo, o que originou em um AIC Corrigido (**AICc**) definido por (Cryer & Chan, 2008):

$$AICc = AIC + \frac{2(m+1)(m+2)}{n-m-2} \quad (50)$$

e no BIC dado por (Pires, 2001),

$$BIC = n \ln S_{\hat{\epsilon}}^2 + m \ln n - (n-m) \ln \left(1 - \frac{m}{n}\right) + m \ln \left[\frac{1}{m} \left(\frac{S_{\hat{X}_t}^2}{S_{\hat{\epsilon}}^2} - 1 \right) \right] \quad (51)$$

onde $S_{x_t}^2$ é a variância amostral da série, S_ε^2 a variância dos resíduos e n o número de observações da série. Porém, o BIC, em relação aos outros critérios, penaliza muito mais a inclusão de parâmetros, o que leva a seleção de modelos mais parcimoniosos (Elers, 2005; Pires, 2001). Neste trabalho se destaca apenas o AIC, o AICc e o BIC por ser os que aparecem nos resultados computacionais do *Software R* versão 4.0.5 utilizado.

3.4.15. Performance de previsão

Os métodos de previsão podem usar transformações matemáticas para tornar estacionária uma série e fazer previsões nesta, para depois inverter as transformações e obter as previsões para a série original. As transformações mais utilizadas são: tomar diferenças sucessivamente; deflacionar; aplicar o logaritmo, para estabilizar as variâncias; fazer ajustes sazonais, *de-trending*; etc. (Maddala & Lahiri, 2009).

Segundo Chatfield (2000), os procedimentos estatísticos de previsão podem ser de três tipos: a) as previsões *Judgemental*; b) os métodos univariados; e, c) métodos multivariados.

Para se obter uma melhor previsão, Brida e Garrido (2009), propõe o uso da estatística U de Theil como uma medida relativa de erro de previsão de forma a se encontrar os melhores parâmetros no modelo SARIMA que possam minimizar tal erro e, conseqüentemente, a qualidade da previsão. A estatística é dada por:

$$U = \frac{\sqrt{\frac{1}{h} \sum_{t=1}^h (x_t - f_t)^2}}{\sqrt{\frac{1}{h} \sum_{t=1}^h (x_t)^2 + \frac{1}{h} \sum_{t=1}^h (f_t)^2}} \quad (52)$$

onde, f_t é o valor previsto pelo modelo no período t ; h é o número de períodos previstos; x_t é o valor observado desejado na previsão; U é o coeficiente de Theil e varia entre 0 (ajuste perfeito) e 1 (previsão muito pobre). Quanto mais próximo o coeficiente for de zero, melhor o resultado da previsão.

Para além da estatística U de Theil, existem outros indicadores que permitem avaliar a qualidade dos métodos e modelos de previsão de séries temporais: o Erro Quadrático Médio (MSE – *Mean Square Error*); a raiz quadrada do MSE (o RMSE – *Root Mean Square Error*), o Erro Médio Percentual Absoluto (MAPE – *Mean Absolute Percentage Error*), o Erro Médio (ME- *Mean Error*), o Erro Absoluto Médio (MAE – *Mean Absolute Error*) e o Erro Percentual Médio (MPE – *Mean Percentage Error*), dados por:

$$MSE = \frac{1}{n} \sum_{t=1}^n (x_t - \hat{x}_t)^2 \quad (53)$$

$$RMSE = \sqrt{MSE} \quad (54)$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{x_t - \hat{x}_t}{x_t} \right| \times 100\% \quad (55)$$

$$ME = \frac{1}{n} \sum_{t=1}^n (x_t - \hat{x}_t) \quad (56)$$

$$MAE = \frac{1}{n} \sum_{t=1}^n |x_t - \hat{x}_t| \quad (57)$$

$$MPE = \frac{1}{n} \sum_{t=1}^n \left(\frac{x_t - \hat{x}_t}{x_t} \right) \times 100\% \quad (58)$$

onde n o número de observações da série, x_t o valor real e o \hat{x}_t o valor previsto; sendo que quanto menor for o valor destes indicadores, melhor é a previsão. Segundo Brida e Garrido (2009), o RMSE não permite fazer a comparação entre diferentes séries temporais e nem em diferentes intervalos de tempo, porém é capaz de ignorar qualquer super e subestimação. O MAPE para além de permitir fazer essas comparações (pois os erros percentuais permitem comparar o desempenho da previsão em diferentes conjuntos de dados, segundo Hyndman e Koehler (2005)), é útil quando as medidas de x_t são relativamente grandes.

Neste trabalho é feito uso do indicador MAPE por três razões: por considerar os valores da série relativamente grandes; pelo fato dos dados de dormidas mensais estarem numa escala de proporção (ou seja, dados com um zero significativo), pois segundo Armstrong e Collopy (1992) o MAPE é mais relevante para este tipo de dados; e ainda por possuir um

conjunto de critérios, proposto por Lewis (1982), para avaliar a qualidade das previsões (Tabela 3.2).

Tabela 3.2: Critério MAPE para a Avaliação da qualidade das previsões.

MAPE (%)	Classificação das Previsões
<10	Alta precisão
10 - 20	Boa precisão
20 - 50	Precisão razoável
>50	Precisão Não confiável

Fonte: Adaptado de Lewis (1982, p.40)

3.5. MÉTODOS DE SUAVIZAÇÃO EXPONENCIAL

Os métodos de suavização exponencial são aqueles que possuem na sua formulação equações de médias móveis, utilizando uma ponderação distinta para cada valor observado na série temporal, onde os valores mais recentes recebam pesos maiores. Estes pesos variam num intervalo de $[0; 1]$ e o conjunto de pesos atribuídos forma um decaimento exponencial a partir dos valores mais recentes. É um dos métodos mais empregados na previsão de curto prazo de séries temporais (Gardner, 2006).

Este tipo de métodos, segundo Souza, Samohyl e Miranda (2008), não depende de variáveis externas na sua aplicação, ou seja, funciona apenas com os dados da própria série [assim como os modelos ARIMA/SARIMA do método de Box-Jenkins]. Segundo Hyndman *et al.* (2002, 2008) os métodos de suavização exponencial incorporam três elementos: o nível, a tendência e a sazonalidade. Estes dois últimos elementos já são conhecidos da seção anterior e entende-se o nível como sendo a média da série no tempo t .

Estes métodos, que para Makridakis e Hibon (2000) possuem boa precisão nas previsões, estão classificados em três tipos ou três variações básicas: Suavização Exponencial Simples (que considera apenas o nível da série); Suavização Exponencial Dupla (conhecida como método de Holt, que adiciona a componente tendência em sua formulação); e,

Suavização Exponencial Tripla (que adiciona ainda a componente sazonal, podendo ser na forma aditiva ou multiplicativa).

Por meio da formulação inicial simples (equação (59)) são construídas as variantes Dupla e Tripla. Para este trabalho destaca-se a Suavização Exponencial Tripla dado que a série temporal de dormidas mensais em estudo possui a componente tendência e sazonalidade, como confirmado no capítulo dos resultados.

$$\hat{X}_{t+1} = \alpha X_t + (1 - \alpha)\hat{X}_t, \quad 0 \leq \alpha \leq 1 \quad (59)$$

onde,

\hat{X}_{t+1} é o valor previsto para o período $t + 1$

X_t é o valor observado actual no período t

\hat{X}_t é o valor previsto no período t , baseado na média de todos os dados históricos da série.

α é a constante de suavização que pode assumir qualquer valor entre 0 e 1. O melhor valor para o α é feita iteração por iteração comparando os resultados por meio do MSE, MAPE ou outro critério (Makridakis & Hibon, 2000).

3.5.1. Suavização Exponencial Tripla

A Suavização Exponencial Tripla, também conhecida por Método de Holt-Winters, é uma expansão do método de Holt (1957) desenvolvida por Winter (1960) para aplicar em séries temporais que possuem as componentes de tendência e sazonalidade. Este método considera dois tipos de efeitos sazonais: o multiplicativo e o aditivo.

O método para efeito sazonal multiplicativo

Seja um modelo de série sazonal, de período s , fator sazonal multiplicativo e tendência aditiva, isto é,

$$X_t = L_t S_t + T_t + \varepsilon_t, \quad (60)$$

que descreve o comportamento da série. Os valores futuros previstos da série são calculados através da função de previsão do método dada por (Makridakis & Hibon, 2000):

$$\hat{X}_{t+h} = (L_t + hT_t)S_{t-s+h}, \quad h = 1, 2, \dots \quad (61)$$

onde, \hat{X}_{t+h} é a previsão para h passos à frente ($t+h$). Além desta função em (61), existem outras utilizadas para estimar o nível (62), a tendência no período atual (63) e os valores do fator sazonal correspondente ao último período de sazonalidade (64).

$$L_t = \alpha \frac{X_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} + T_{t-1}), \quad 0 \leq \alpha \leq 1 \quad (62)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}, \quad 0 \leq \beta \leq 1 \quad (63)$$

$$S_t = \gamma \frac{X_t}{L_t} + (1 - \gamma)S_{t-s}, \quad 0 \leq \gamma \leq 1 \quad (64)$$

onde,

α , β e γ são as constantes de suavização que controlam o peso relativo ao nível (L_t), a tendência (T_t) e a sazonalidade (S_t), respectivamente. As equações L_t e S_t assumem que, no período t , existe uma estimativa do índice sazonal no período de tempo $t-s$ ou que existe um valor S_{t-s} . Assim, e para utilizar as equações de suavização, é necessário estimar os valores para S_1, S_2, \dots por meio da equação (65).

$$S_t = \frac{X_t}{\frac{\sum_{k=1}^s X_k}{s}}, \quad t = 1, 2, \dots, s \quad (65)$$

O método para efeito sazonal aditivo

Este método é utilizado quando a amplitude do ciclo sazonal permanece constante com o passar do tempo, ou seja, não depende do nível local da série. No caso, um modelo cuja série sazonal, de período s , é formada pela soma do nível, tendência, um fator sazonal e um erro aleatório, é dada por

$$X_t = L_t + T_t + S_t + \varepsilon_t, \quad t = 1, 2, \dots \quad (66)$$

Os valores futuros previstos da série são calculados através da função de previsão do método dada por

$$\hat{X}_{t+h} = L_t + hT_t + S_{t+h-s}, \quad h = 1, 2, \dots \quad (67)$$

onde, \hat{X}_{t+h} é a previsão para h períodos à frente ($t+h$). Além desta função em (67), existem outras utilizadas para estimar o nível (68), a tendência no período atual (69) e os valores do fator sazonal correspondente ao último período de sazonalidade (70):

$$L_t = \alpha(X_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1}), \quad 0 \leq \alpha \leq 1 \quad (68)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}, \quad 0 \leq \beta \leq 1 \quad (69)$$

$$S_t = \gamma(X_t - L_t) + (1 - \gamma)S_{t-s}, \quad 0 \leq \gamma \leq 1 \quad (70)$$

onde,

α , β e γ são as constantes de suavização que controlam o peso relativo ao nível (L_t), a tendência (T_t) e a sazonalidade (S_t), respectivamente. As equações L_t e S_t assumem que, no período t , existe uma estimativa do índice sazonal no período de tempo $t-s$ ou que existe um valor S_{t-s} . Assim, e para utilizar as equações de suavização, é necessário estimar os valores para S_1, S_2, \dots por meio de (Makridakis & Hibon, 2000):

$$S_t = X_t - \sum_{k=1}^s \frac{X_k}{s}, \quad t = 1, 2, \dots, s \quad (71)$$

A nível computacional o Método de Holt-Winters pode ser utilizado através do *Software* R versão 4.0.5, por meio da função `HoltWinters()` para verificar o ajustamento dos dados, cujo comando mais simples se apresenta da seguinte forma (Meyer, 2021):

```
HoltWinters(x, seasonal=c("additive", "multiplicative"))
```

em que x é o conjunto de dados da série temporal, **seasonal** o tipo de sazonalidade a escolher (o aditivo ou o multiplicativo).

Entretanto esta função não permite realizar as devidas previsões, pelo que é útil o uso do pacote **forecast** e sua função de Holt-Winters `hw()` cujo comando é

```
hw(x, h, seasonal= c("additive",
"multiplicative", initial="optimal")
```

em que h é o período de previsão.

3.6. A Lei de Benford

No século XIX, o astrónomo e matemático Simon Newcomb encontrou um padrão que ocorria com o manuseio de livros de tabelas logarítmicas. Observou, em 1881, que as primeiras páginas dos livros eram mais gastas, ou seja, mais utilizadas, e correspondiam às páginas cujos números iniciavam com 1, 2 ou 3 quando comparadas às últimas páginas cujos primeiros dígitos eram os números 7, 8 ou 9 (Newcomb, 1881). Ele deduziu, portanto, que em muitos fenómenos quantitativos, ou seja, em um conjunto de números que apareceu na natureza, mais números começam com o dígito 1 do que com o dígito 2, e mais com 2 do que com 3, e mais com 3 do que com 4 e mais com 4 do que com 5 e mais com 5 do que com 6 e assim por diante. Ele projetou a distribuição de probabilidade de ocorrência do primeiro dígito dos números em um conjunto de dados, visualizando assim uma distribuição logarítmica (Newcomb, 1881).

O tópico foi retomado e estudado, e a teoria foi formalmente demonstrada e disseminada em 1938 pelo físico americano Frank Benford, ficando conhecida como Lei de Benford, ou Lei do Primeiro Dígito, ou Lei de Newcomb-Benford, ou Lei dos Números Anômalos. Entretanto, para muitos, esta Lei pode parecer um tanto intrigante, como podemos ver: em um jogo que envolve lançar um dado de seis lados, a probabilidade de que saía um dos lados é $1/6$, e a probabilidade de obter cara ou coroa ao jogar uma moeda equilibrada é $1/2$. Nessa lógica, nosso senso comum diria que em um conjunto de dados qualquer número que escolhermos aleatoriamente teria 1 como o primeiro dígito com probabilidade $1/9$ e 2 como o primeiro dígito com a mesma probabilidade, e assim por diante. Ou seja, todos os dígitos (1, 2 ..., 9) teriam a mesma frequência, ou seja, ocorreriam praticamente na mesma quantidade, porém uma distribuição uniforme não ocorre em muitos fenómenos estudados por Benford (Benford, 1938). Ele analisou vários conjuntos de dados de várias fontes, como comprimentos de rios, número de populações

da cidade, etc., provando assim o que Newcomb havia descoberto cinquenta e sete anos atrás.

Na verdade, ultimamente, a Lei de Benford tem sido usada para detetar evidências de irregularidades e manipulação de dados, fraude eleitoral e fiscal, dados económicos e psicológicos. O trabalho de Saville (2014) mostrou que é possível detetar relatórios falsos ou fraudulentos de dados de contabilidade usando a Lei de Benford sobre dados retirados de empresas listadas na Bolsa de Valores de Johannesburg. Essas informações foram então usadas por auditores e serviços de fiscalização, mas também por analistas financeiros como entrada para modelos de previsão.

Em relação à fraude eleitoral, há publicações apontando em diferentes direções, o trabalho de Deckert, Myagkov e Ordeshook (2011) mostrou como, neste contexto, a Lei de Benford pode ser uma ferramenta forense problemática. A Lei de Benford foi aplicada com sucesso como um detetor de mentiras, apresentando um avanço na psicologia cognitiva (Gauvrit, Houillon & Delahaye, 2017). O trabalho de Gürnel e Tödter (2007) mostrou que as frequências relativas do primeiro e do segundo dígito na pesquisa económica são amplamente consistentes com a Lei de Benford. Os dados da procura turística seguem padrões de crescimento económico, a revisão feita por Gürnel e Tödter (2007) apoia a questão de que a Lei de Benford poderia ser aplicada a esta área.

Deve-se notar que a Lei de Benford tem evidência empírica, investigação e aplicação prática; porém, como qualquer lei/modelo, não tem uma explicação universal. Logo, a Lei de Benford pode ser usada como um teste para saber se um conjunto de dados contém números fidedignos (Gabaix, 2016).

A Lei de Benford tem sido cada vez mais usada como uma ferramenta de diagnóstico para isolar grandes conjuntos de dados com irregularidades que merecem uma inspeção mais detalhada. Executar uma varredura sistemática é extremamente árduo; portanto, esses dados são um candidato natural para a triagem inicial em comparação com as distribuições de Benford (Cho & Gaines, 2007).

A Lei de Benford pode ser definida como uma coleção de evidências empíricas relacionadas com a distribuição de frequência dos dígitos iniciais em conjuntos de dados

numéricos. A versão mais conhecida da lei afirma que, nesses conjuntos de dados que representam uma coleção de dados "naturais", a probabilidade de ver um dígito específico na primeira posição está inversamente relacionada à sua classificação (Kopczewski & Okhrimenko, 2019).

Quanto à sua aplicação no contexto turístico, pelo que se sabe, existe apenas um trabalho recente que explora a aplicação desta lei no contexto do tráfego turístico na Croácia, realizado por Matakovic (2021). Neste estudo a Lei de Benford foi aplicada ao número de dormidas domésticas e internacionais, numa série mensal durante um período de três anos na Croácia. Neste estudo, apenas as dormidas domésticas não seguiram a Lei de Benford. Porém, os autores não identificaram as causas da adaptação e não adaptação do número de dormidas à Lei de Bendford, eles presumiram que isso ocorreu principalmente devido à falta de precisão e ausência de registos em alojamento turístico.

A proposta concebida por Newcomb e demonstrada por Benford, mostra que o primeiro dígito dos números, num conjunto de dados, segue uma distribuição logarítmica, onde a probabilidade de ocorrência do primeiro dígito (d) num universo de números aleatórios é dada por (Newcomb, 1881; Benford, 1938):

$$P(\text{primeiro dígito} = d) = \log_{10} \left(1 + \frac{1}{d} \right), \quad d = 1, 2, \dots, 9. \quad (72)$$

Empregando a equação (72), é então possível determinar a probabilidade de ocorrência dos números cujo primeiro dígito é 1, 2, ... ou 9, conforme mostrado na Tabela 3.3. Por alguma razão, os números que começam com os dígitos 1 e 2 são muito mais frequentes do que números que começam com os dígitos 8 e 9. Indicando que os números que começavam com 1 e 2 ocorriam com mais frequência na natureza do que números que começavam com 8 e 9 (Fewster, 2009).

Tabela 3.3: Lei de Benford para o primeiro dígito de números em um conjunto de dados.

Primeiro Dígito (d)	1	2	3	4	5	6	7	8	9	Total
Probabilidade $P(d)$	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046	1
Percentual (%)	30,1%	17,6%	12,5%	9,7%	7,9%	6,7%	5,8%	5,1%	4,6%	100%

Fonte: Elaboração própria a partir do modelo teórico da Lei

Por meio da equação (72) e da Tabela 3.3, é fácil entender que, de acordo com a Lei de Benford, a distribuição dos dígitos iniciais em números está longe de ser uniforme. Um número selecionado aleatoriamente deve começar com o dígito 1 cerca de 30% das vezes: mais precisamente, a proporção deve ser 0,301, ou $\log_{10}(2)$. A frequência de números com 2 como o primeiro dígito deve ser cerca de 17,6% (obtido do $\log_{10}\left(\frac{3}{2}\right)$), aqueles com 3 como o primeiro dígito devem ser cerca de 12,5% (do $\log_{10}\left(\frac{4}{3}\right)$), e sucessivamente até que a frequência do dígito 8 seja cerca de 5,1% e para o dígito 9 cerca de 4,6% (Fewster, 2009). A distribuição logarítmica mencionada anteriormente é representada na Figura 3.2.

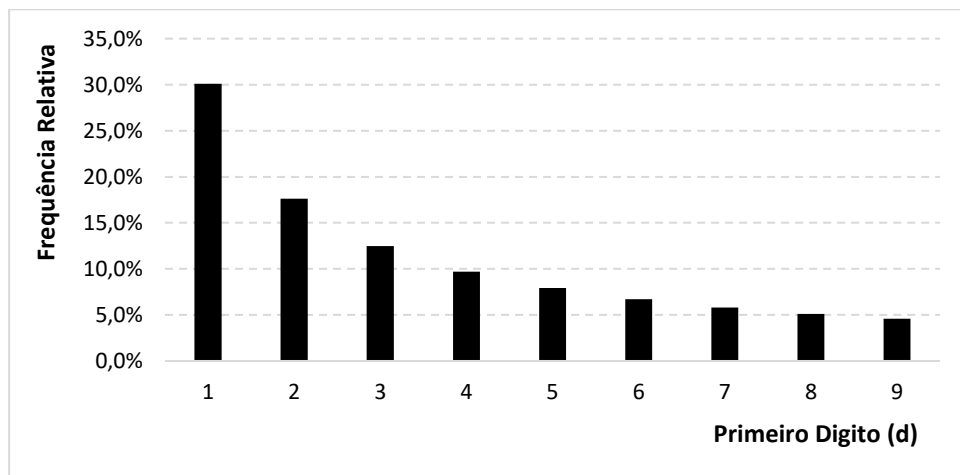


Figura 3..2: Percentual da frequência do primeiro dígito de números em um conjunto de dados, segundo a Lei de Benford.

Fonte: Elaboração própria

3.7. INTELIGÊNCIA ARTIFICIAL: REDES NEURONAIS

Redes Neurais Artificiais ou simplesmente Redes Neurais (RN) são uma metodologia muito abrangente que tem por detrás o objetivo de tentar "reproduzir" o funcionamento da rede de neurónios do cérebro humano. Vista por Matich (2001), é mais uma forma de imitar certas características do ser humano, como a capacidade de memorizar e de associar coisas. Entretanto o funcionamento e desempenho das RN depende da escolha adequada das suas diferentes componentes, nomeadamente: função de ativação; função

erro; parâmetros de controle do gradiente descendente; tempo de treino; e, a arquitetura. Estes e outros aspetos específicos das RN são abordadas ao longo desta secção.

3.7.1. Breve história das redes neuronais

Tudo começa com Alan Turing, em 1963, quando resolveu estudar o cérebro como uma forma de entender o mundo computacional. Entretanto Warren McCulloch e Walter Pitts foram os que conceberam os primeiros fundamentos da computação neuronal e mostraram, em 1943, a forma como trabalhavam os neurónios, modelando através de circuitos elétricos uma rede neuronal simples (Barreto, 2002; Moreno, 2002).

Em 1949 Donald Hebb surge como pioneiro ao explicar os processos de aprendizagem mostrando que estas ocorriam quando certas mudanças num neurónio aconteciam. Em 1957 Frank Rosenblatt desenvolve a rede neuronal *Perceptron* com a capacidade de reconhecer padrões semelhantes depois de ter havido aprendido outros padrões. Mas a primeira rede neuronal aplicada a um problema real foi desenvolvida em 1960 por Bernard Widrow e Marcial Hoff, a chamada modelo *Adaline* (Matich, 2001).

Em 1964 M. Hu apresenta o primeiro modelo de RN para a previsão de séries temporais. Cinco anos depois, o *Perceptron* foi colocado em cheque por Marvin Minsky e Seymour Papert que provaram que essa rede neuronal não era capaz de resolver problemas simples onde envolvia funções não lineares. Este período é chamado de “Época das Trevas”, ou seja, o abandono (declínio) das investigações em RN (Moreno, 2002).

Depois das críticas sobre as limitações na rede *Perceptron*, as RN voltam ao destaque em 1985 por John Hopfield com o livro “Computação neuronal de decisões em problemas de optimização”. Um ano mais tarde é redescoberto o algoritmo *backpropagation* por David Rumelhart e G. Hinton na publicação do livro “*Parallel Distributed Processing*” (Rumelhart & Hinton, 1987). Redescoberto porque inicialmente foi Paul Werbos quem havia trabalhado esse algoritmo em 1974. E a partir de 1986 as redes neuronais ganham interesse no mundo da investigação (Batista, 2009; Moreno, 2002; Matich, 2001).

Segundo Batista (2009) tem havido uma certa confusão quanto às terminologias usadas nas RN e nos modelos estatísticos clássicos, dado que as pesquisas nesses campos eram desenvolvidas de forma independente. Assim, tendo em conta as semelhanças que existem entre eles, um glossário foi criado por W. Sarle (1994) para facilitar a comunicação entre essas duas áreas, e está resumido na Tabela 3.4.

Tabela 3.4: Glossário Redes Neurais x Estatística.

Redes Neurais	Estatística
Pesos	Parâmetros
Conjunto de treino	Amostra
Entradas	Variáveis independentes
Saídas	Valores previstos
Retropropagação	Aproximação estocástica
Treino ou aprendizagem	Estimação
Sinal de entrada	Valor das variáveis exógenas
Sinal de saída	Valor estimado
Alvo	Valor da variável endógena
Padrões	Observações

Fonte: Batista (2009, p.52)

3.7.2. Conceito de Redes Neurais

Uma rede neuronal é um modelo artificial e simplificado do cérebro humano capaz de ‘imitar’ certas características do ser humano para resolver problemas complexos por meio de sistemas de computação (algoritmos), cuja inspiração está na estrutura e funcionamento dos sistemas nervosos tendo como elemento principal o neurónio (Tépan, 2013; Matich, 2001).

Outras definições podem ser encontradas em Matich (2001). Segundo Cortez (1997) esses modelos são capazes “(...) de aprender a partir do conhecimento de situações passadas, dando respostas a novas situações, especialmente em problemas com uma forte componente de ruído e/ou informação incompleta” (Cortez, 1997, p.3). A Figura 3.3 apresenta um modelo padrão de um neurónio artificial.

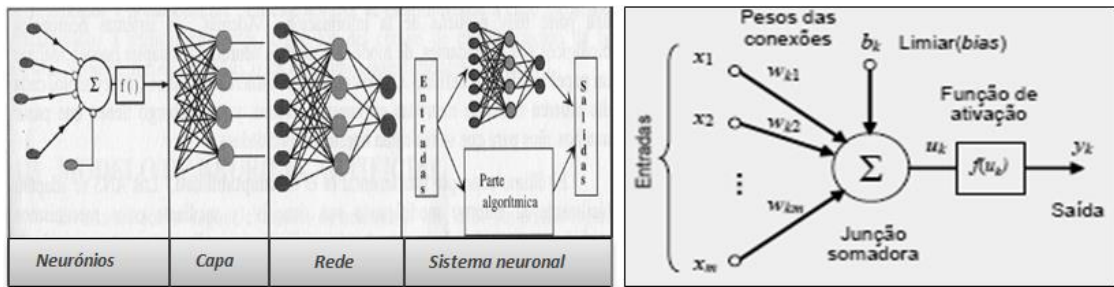


Figura 3.3: Modelo padrão de um neurónio artificial.

Fonte: Larrañaga, Inza e Moujahid (2015, p.3); Batista (2009, p.43)

3.7.3. Funcionamento

Uma RN é composta por um conjunto de neurónios artificiais que através de um vetor de entrada gera uma resposta única (Tepán, 2013). Numa rede neuronal artificial as ligações entre os neurónios são feitas por meio de sinapses que por sua vez possuem ligações direcionais, ou seja, a informação propaga-se apenas num único sentido. Da mesma forma que biologicamente se considera que o conhecimento está relacionado com a ligação entre os neurónios (em média cada neurónio pode ligar-se com 10.000 outros) também se considera que nas RN o conhecimento se encontra representado nos *pesos* das ligações entre neurónios (Larrañaga, Inza & Moujahid, 2015; Matich, 2001).

Normalmente os neurónios agrupam-se em unidades estruturais denominadas *camadas*. Estas definem-se como um conjunto de neurónios cujas entradas de informação são provenientes da mesma fonte e as saídas de informação dirigem-se para o mesmo destino. O conjunto de uma ou mais *camadas* formam a rede neuronal. Cada neurónio está ligado aos outros por um tipo de ligação na qual tem associado um determinado *peso*. Este representa a informação que será usada pela rede na resolução de um determinado problema. Existem três tipos de *camadas*: de entrada (sensorial - recebe a informação do exterior), de saída (resposta da rede ao exterior) e ocultas (processamento - as entradas e saídas encontram-se dentro do sistema) (Larrañaga, Inza & Moujahid, 2015).

Assim, quanto a estrutura, temos as redes *monocamada* (conhecidas por redes autoassociativas) – os neurónios estão agrupados em apenas uma camada – e *multicamadas* (conhecidas por redes heteroassociativas) – os neurónios estão agrupadas em pelo menos duas camadas. Quanto à propagação de dados tem-se as redes unidireccionais (*feedforward*) – único sentido – e redes recorrentes (*feedback*) – em qualquer sentido (Larrañaga, Inza & Moujahid, 2015; Palmer & Montano, 1999).

As RN funcionam com algoritmos de aprendizagem ou de treino que procuram valores numéricos adequados para os chamados *pesos* sinápticos. Através do algoritmo os *pesos* vão sendo ajustados pela rede a fim de emitir uma resposta (saída) de uma informação de entrada. Consoante forem os valores dos *pesos* assim as ligações dos neurónios podem: destruir-se (*peso=0*), modificar-se ou criar-se (*peso≠0*) (Passari, 2003).

O processo de aprendizagem termina quando os *pesos* se mantiverem estáveis. Os algoritmos de treino (regras de aprendizagem) estão divididos em duas classes: *supervisionados* e *não supervisionados*. A diferença reside no fato de que na primeira existe um conjunto de casos de treino¹⁰ em que a aprendizagem é feita por ajustamentos dos *pesos* das ligações e passa pela minimização do erro entre a saída da rede e uma saída desejada; e na segunda apenas um conjunto de entradas onde os dados são agrupados a partir de suas propriedades ou padrões, ou seja, é capaz de identificar padrões nos valores de entrada e não recebe informações adicionais (Tepán, 2013; Passari, 2003; Cortez, 1997).

Em suma, os neurónios recebem os sinais de entrada dos neurónios vizinhos com uma certa ponderação dada pelos *pesos* das ligações. A soma desses sinais ponderados proporciona a entrada do neurónio e mediante a aplicação de uma função matemática sobre a entrada se calcula o valor de saída, que por sua vez é enviado a outros neurónios. Tanto os valores de entrada como de saída podem tomar valores negativos (estado de inibição) ou positivos (estado de excitação) (Palmer & Montano, 1999).

¹⁰ Um caso de treino consiste num vetor de entrada e no correspondente vetor de saída (Cortez, 1997).

Pode-se notar na Figura 3.3 (à direita) a presença da componente b_k (o *bias*) cuja função, segundo Neves (2017) “(...) é realizar uma transformação afim no somatório de modo que, quando a função de ativação for maior que zero, o neurónio produz uma saída. Caso contrário permanece desativado” (Neves, 2017, p. 8).

Matematicamente a associação de todas as entradas origina a entrada total net_i dada por (Miguel, 2011):

$$net_i = \sum_{j=1}^n a_{ij} y_j(t) + \sum_{k=1}^{m-1} b_{ik} u_k(t) + b_{im} bias_i \quad (73)$$

em que y_j são as saídas de outros neurónios, u_k entradas externas, $bias_i$ entrada constante, a_{ij} e b_{ik} os pesos correspondentes e m número de entradas; e a saída do neurónio k é dada por (74) podendo ser simplificada por (75) de forma a incluir o *bias* definindo $x_0 = 1$ com peso associado $w_{k0} = b_k$ (Zuben & Boccato, 2017):

$$y_k = f \left[\sum_{i=1}^m x_i w_{ki} + b_k \right] = f(u_k) \quad (74)$$

e

$$y_k = f \left[\sum_{i=0}^m x_i w_{ki} \right] = f(u_k) = f(w^T x) \quad (75)$$

onde $x_i, i = 1, \dots, m$, as entradas; w^T a matriz de pesos; f a função de ativação.

Para a produção da saída do neurónio é feita a atualização na ativação dos neurónios por meio de funções de ativação. As funções de ativação mais conhecidas estão apresentadas na Tabela 3.5. Sendo γ o ponto de inflexão em que variando-o se obtêm diferentes inclinações da função e α, β previamente definidos.

Tabela 3.5: Funções de ativação mais utilizadas.

Nome da função	Função de activação	Contradomínio	Derivada
Degrau	$y = f(u_k) = \begin{cases} 1 & \text{se } u_k > b_k \\ 0 & \text{se } u_k \leq b_k \end{cases}$	$] - \infty, \infty[$	
Linear	$y = f(u_k) = \begin{cases} 0, & \text{se } u_k \leq \alpha \\ \frac{x - \alpha}{\beta - \alpha}, & \text{se } \alpha < u_k \leq \beta \\ 1, & \text{se } u_k > \beta \end{cases}$	$] - \infty, \infty[$	$\frac{dy}{du_k} = 1$
Logística (sigmóide)	$y = f(u_k) = \frac{1}{1 + \exp(-\gamma u_k)}$	$[0; 1]$	$\frac{dy}{du_k} = \gamma y(1 - y) > 0$
Tangente hiperbólica	$y = f(u_k) = \frac{\exp(\gamma u_k) - \exp(-\gamma u_k)}{\exp(\gamma u_k) + \exp(-\gamma u_k)}$	$[-1; 1]$	$\frac{dy}{du_k} = \gamma(1 - y^2) > 0$

Fonte: Elaboração própria, a partir de Batista (2009, p. 45), Schopf (2007, p. 20), Haykin (2008, p. 195)

3.7.4. Vantagens e desvantagens

A capacidade de aprender mediante treino ou por uma experiência inicial (aprendizagem), ter a sua própria representação da informação (auto-organização), ser muito robusto (tolerância a falhas), e responder em tempo real são as principais vantagens de uma RN. Quanto às desvantagens, temos: complexidade de aprendizagem para grandes tarefas, muito tempo de aprendizagem, não permite interpretações e grande quantidade de informação para realizar o treino (Matich, 2001). Ainda poderíamos citar como vantagem o facto de a RN não requerer um conhecimento *a priori* da relação entre as variáveis de entrada e saída, poder detetar automaticamente a relação de dependência entre as variáveis, capacidade de aproximar qualquer função contínua e boa capacidade de generalização (Passari, 2003).

3.7.5. Tipo de redes neuronais

Vários são os tipos de redes neuronais, como por exemplo: *Perceptron*, *Adaline*, *Perceptron* Multicamada, Rede de Aprendizagem Associativa, Redes de Kohonen, LVQ

“*Learning Vector Quantization*”, Redes recorrentes, *Backpropagation*, RBF (*Radial Basis Function*), *Dynamic Bayesian*, *Predictive modular* e *Feedforward*. Como se pode verificar muitos são os tipos de arquiteturas de RN, sendo que cada um apresenta suas próprias vantagens e desvantagens. Entretanto, segundo Cortez (1997), a grande diferença é feita entre as redes *feedforward* e *recorrentes*. Na primeira as ligações são unidirecionais, ou seja, o fluxo de sinal é apenas em um sentido e na segunda as ligações são feitas entre quaisquer neurónios, ou seja, existe pelo menos um ciclo de retroalimentação, em que o sinal retorna para uma camada anterior (Schopf, 2007). Mas, Machado (2010) junta a essas duas redes uma terceira – as redes radiais – que em seu entender apresentam melhores resultados.

Entretanto neste trabalho damos atenção às redes *feedforward*, pois além de serem as mais amplamente utilizadas (Cortez, 1997) e dependendo do número de neurónios, possuem a capacidade de aproximar qualquer função não linear (Miguel, 2011), também têm sido utilizadas em previsões mensais no contexto turístico, resultando num melhor desempenho face a outros métodos utilizados neste contexto.

3.7.6. As redes *Feedforward*

Especificamente, trataremos da rede neuronal *feedforward* multicamada mais conhecida por MLP (*MultiLayer Perceptron*). Ela é constituída por um conjunto de neurónios dispostos em várias camadas, sendo uma de entrada, uma ou mais intermédias e uma de saída (Figura 3.4). Segundo Carvalho (2005) o neurónio pode estar total ou parcialmente conectado por *pesos*, sendo que a rede é alimentada pela camada de entrada e são processadas progredindo para frente, da esquerda para a direita, de camada em camada, até chegarem a camada de saída, sem realimentação.

Os neurónios da primeira camada fazem a distribuição das entradas para todos os neurónios da camada seguinte e assim sucessivamente. As camadas intermédias fazem o processamento não linear dos padrões recebidos e a última camada atua como a saída da rede (Santana, 2006).

Normalmente as funções de ativação mais utilizadas para este tipo de rede, são as *lineares*, as *sigmoidais* e a *tangente hiperbólica* (Machado, 2010; Santana, 2006; Cortez, 1997). A função de ativação tem o papel de impedir a saída de valores elevados que poderá atrasar o processo de convergência do algoritmo de treino (Santana, 2006). Segundo Matich (2001), quando as funções de ativação não são lineares é necessário calcular a derivada da função em relação ao *peso*. Com esta derivada os *pesos* são modificados durante o processo de aprendizagem.

Entretanto para o cálculo do gradiente local de cada neurónio da rede é requerido que a função de ativação não-linear a utilizar seja continuamente diferenciável. A função de ativação mais utilizada neste tipo de rede é a não-linearidade sigmóide da *função logística* ou da *função tangente hiperbólica* (Haykin, 2008).

A título de exemplo, a Figura 3.4 ilustra uma rede *feedforward* MLP totalmente conectada e com duas camadas intermédias. Segundo Pommerenzenbaum (2014) é o mais utilizado nos trabalhos práticos encontrados na literatura.

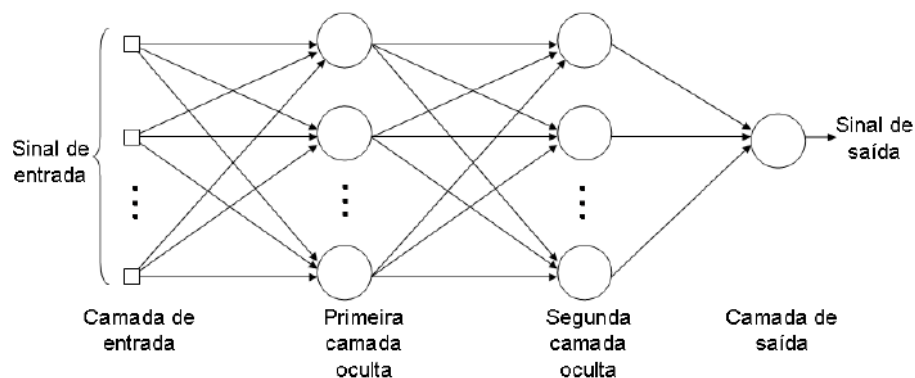


Figura 3.4: Rede Neuronal feedforward MLP com duas camadas intermédias.

Fonte: Schopf (2007, p.23)

De salientar que esta rede possui três características importantes: o modelo de cada neurónio inclui uma função de ativação não-linear (uma não-linearidade suave, como por

exemplo a sigmóide); possui uma ou mais camadas ocultas de nós computacionais; e exibe um alto grau de conectividade (Hykin, 2008).

Importante salientar que o número de nós aconselhável na camada oculta, quando esta é única, segundo Bishop (1995), varia de 5 a 25.

3.7.7. O Algoritmo Error *Backpropagation*

O algoritmo de treino associada à rede *feedforward* é o *error backpropagation* [propagação reversa ou retropropagação de erro], ou simplesmente *backpropagation*, dada a sua simplicidade de implementação e provada eficiência (Pommerenzenbaum, 2014), por ser o mais popular dentro da aprendizagem supervisionada, segundo Cortez (1997) e por fornecer um método computacional eficiente para o treino da rede (Haykin, 2008). A origem do nome está relacionada com a forma como o erro se propaga, isto é, o erro se propaga de traz para frente a partir da camada de saída, permitindo que os *pesos* nas conexões dos neurónios nas camadas ocultas mudem durante o treino. Esta mudança influencia tanto a entrada global como a ativação e a saída de um neurónio (Matich, 2001).

A aprendizagem

Especificamente, a aprendizagem é feita em dois passos: uma, a chamada *propagação* (passo para frente), no qual o vetor de entrada é aplicado aos nós sensoriais da rede e se propaga para frente camada por camada e mantendo fixo os *pesos* sinápticos; a outra, a chamada *retropropagação* (passo para traz), que começa na camada de saída, na qual a resposta produzida no primeiro passo é subtraída de uma resposta desejada produzindo um sinal de *erro* e este é propagado para traz contra a direção das conexões sinápticas, camada por camada, através da rede, e recursivamente calculando-se o gradiente local de cada neurónio.

Aqui os *pesos* sinápticos são todos ajustados, de acordo com a regra *delta* (δ) da fórmula (86), de forma que a resposta da rede se aproxima o mais possível da resposta desejada.

Para calcular as modificações dos pesos de todas as conexões que alimentam a camada de saída utilizamos a fórmula (86). Para calcular os gradientes locais para todos os neurónios na penúltima camada é feito o uso da fórmula (88). A recursividade continua camada por camada durante todo o ciclo que engloba o passo para frente seguido pelo passo para trás. Convém salientar a importância de que todos os neurónios devem aprender com a mesma taxa de aprendizagem. O parâmetro da taxa de aprendizagem mencionado na fórmula (86) deve possuir valores menores nas últimas camadas da rede (devido a estas possuírem gradientes locais maiores) e principalmente em caso de neurónios com muitas entradas, de forma a se manter algum equilíbrio no tempo de aprendizagem (Haykin, 2008).

A matemática por detrás do algoritmo

Seguidamente é descrita, resumidamente, a matemática, dada por Haykin (2008), que sustenta este algoritmo, tendo em consideração a seguinte notação: i , j e k são neurónios diferentes na estrutura da rede, em que, caso o neurónio j seja uma unidade oculta, então o i e o k se encontram a esquerda e a direita do j respetivamente:

- O sinal de erro na saída do neurónio j , na iteração n é dado por

$$e_j(n) = d_j(n) - y_j(n) \quad (76)$$

onde $d_j(n)$ é a resposta desejada para o neurónio j e $y_j(n)$ é o sinal funcional que aparece na saída do neurónio j .

- O valor instantâneo da energia total do erro é dado por

$$\xi(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n) \quad (77)$$

onde o conjunto C inclui todos os neurónios da camada de saída da rede.

- A energia média do erro quadrado é dada por

$$\xi_{méd} = \frac{1}{N} \sum_{n=1}^N \xi(n) \quad (78)$$

onde N é o número total de padrões contidos no conjunto de treinamento. Para um dado conjunto de treino, $\xi_{méd}$ serve como uma medida do desempenho de aprendizagem, isto é, quando se ajusta os parâmetros livres da rede (pesos sinápticos e *bias*) o objetivo é minimizar o $\xi_{méd}$.

- O campo local induzido (que é a soma ponderada de todas as entradas sinápticas mais o *bias*) que constitui o sinal aplicado à função de ativação associada ao neurónio j é dado por

$$v_j(n) = \sum_{i=0}^m w_{ji}(n)y_i(n) \quad (79)$$

onde m é o número total de entradas aplicadas ao neurónio j , w_{ji} é o peso sináptico conectando a saída do neurónio i à entrada no neurónio j . Mas para $i = 0$, w_{j0} (que corresponde à entrada fixa $y_0 = +1$) é igual ao *bias* b_j , ou seja, diferenciando a equação (79) em relação a $y_i(n)$ temos,

$$b_j = w_{j0}(n) = \frac{\partial v_j(n)}{y_0(n)} \quad (80)$$

- Assim, o sinal funcional que aparece na saída do neurónio j , na iteração n é

$$y_j(n) = \varphi_j(v_j(n)) \quad (81)$$

onde φ_j é a função de ativação, que descreve a relação funcional não linear da entrada-saída associada ao neurónio j .

Normalmente as funções de activação mais utilizadas neste tipo de algoritmo são: a função logística e a função tangente hiperbólica. Neste trabalho é feito uso da função logística, pois é a que surge nas funções computacionais do *Software R* utilizadas no Capítulo IV. Esta função já apresentada na Tabela 3.5 é reescrita agora com os símbolos em uso neste subtópico, da seguinte forma

$$\varphi_j(v_j(n)) = \frac{1}{1 + \exp(-\alpha v_j(n))} \quad \alpha > 0; v_j(n) \in] -\infty, \infty[\quad (82)$$

Ao derivar esta equação em relação a $v_j(n)$ e fazendo $y_j(n) = \varphi_j(v_j(n))$ obtém-se

$$\varphi'_j(v_j(n)) = \alpha y_j(n)[1 - y_j(n)] \quad (83)$$

Esta derivada permite expressar o gradiente local para o neurónio j localizado na camada de saída em que o sinal funcional na saída deste neurónio j ($o_j(n)$) é igual a $y_j(n)$. Este gradiente é dado por

$$\begin{aligned} \delta_j(n) &= e_j(n)\varphi'_j(v_j(n)) \\ &= \alpha[d_j(n) - o_j(n)o_j(n)][1 - o_j(n)] \end{aligned} \quad (84)$$

em que $d_j(n)$ é a resposta desejada para o neurónio j .

Com relação a um neurónio j oculto arbitrário, o gradiente local é expressa como

$$\delta_j(n) = \alpha y_j(n)[1 - y_j(n)] \sum_k w_{jk}(n)\delta_k(n) \quad (85)$$

- Ao peso sináptico $w_{ji}(n)$ é aplicada uma correção definida pela regra *delta* (δ),

$$\Delta w_{ji}(n) = -\eta \frac{\partial \xi(n)}{\partial w_{ji}(n)} = \eta \delta_j(n) y_i(n) \quad (86)$$

onde η é o parâmetro da taxa de aprendizagem do algoritmo e o sinal negativo indica a descida do gradiente no espaço de pesos, isto é, a correção procura fazer mudanças nos pesos de forma a reduzir $\xi(n)$.

Entretanto convém salientar que caso queiramos aumentar o η para acelerar a taxa de aprendizagem, e de forma que não haja instabilidade na rede, a fórmula a utilizar é

$$\Delta w_{ji}(n) = \alpha \Delta w_{ji}(n-1) + \eta \delta_j(n) y_i(n) \quad (87)$$

onde α é a *constante de momento*, que controla o laço de realimentação que age em torno de $\Delta w_{ji}(n)$.

Como se pode verificar, a fórmula (87) é uma generalização da (86), sendo que esta é obtida quando $\alpha = 0$.

- O gradiente local que aponta para as mudanças necessárias nos pesos sinápticos é dado por

$$\delta_j(n) = -\frac{\partial \xi(n)}{\partial v_j(n)} = e_j(n)\varphi'_j(v_j(n)) \quad (88)$$

- Quando o neurónio j é um nó de saída, utiliza-se a fórmula (76) para calcular o seu sinal de erro, bem como a fórmula (87) para o cálculo do gradiente local. Mas quando o neurónio j é um nó oculto (pertence a uma camada oculta), a fórmula de retropropagação para o gradiente local é dada por

$$\delta_j(n) = \varphi'_j(v_j(n)) \sum_k \delta_k(n) w_{kj}(n) \quad (89)$$

O treino

Quando se trata de um conjunto de treino, a aprendizagem do algoritmo pode ser feita de duas formas: modo *Sequential* (ou modo padrão) ou modo por *Lote*. A diferença é que no primeiro a atualização dos pesos é feita após cada treino, enquanto que na segunda forma o ajuste dos pesos é realizado no final de todos os treinos. A par das vantagens e desvantagens de cada uma das formas, o modo *sequencial* tem sido mais utilizado devido à simplicidade na implementação do seu algoritmo, por fornecer soluções efetivas diante de problemas grandes e difíceis e é computacionalmente mais rápido (Haykin, 2008).

Crítérios de paragem

Aqui uma questão crucial é saber quando parar o processo de aprendizagem, ou seja, em que momento se deve parar de fazer ajustes aos pesos. Um dos critérios seria considerar a **norma euclidiana do vetor gradiente**. Porém o tempo de aprendizagem poderá ser longo quando se quer alcançar um limiar suficientemente pequeno da norma. Um outro seria a **taxa absoluta de variação do erro médio quadrado**. Neste caso como se deseja que esta taxa seja suficientemente pequena, tal pode parar o processo de aprendizagem prematuramente.

Existe ainda um outro critério de convergência, e que foi adotada neste trabalho, que é o critério de **validação cruzada**, cujo procedimento padrão é dividir aleatoriamente o conjunto de dados da série em um conjunto de treino e em um conjunto de teste. O conjunto de treino, por sua vez, é dividido em dois subconjuntos: um de estimação (para seleccionar o modelo) e um outro de validação (para testar ou validar o modelo). Assim, a partir do conjunto de treino se pode avaliar o desempenho de diferentes modelos e escolher o que melhor satisfaz sob uma determinada condição (por exemplo o menor

erro MAPE). Entretanto o modelo seleccionado pode ajustar em demasia o subconjunto de validação. No caso, a generalização do modelo seleccionado é feita sobre o conjunto de teste (Haykin, 2008).

Mas dentro da validação cruzada, se pode realizar também o treino com **paragem antecipada**. Isto é, o treino é interrompido após um número determinado de épocas (ou seja, periodicamente), em que os pesos sinápticos e os níveis de *bias* são todos fixos e então para cada padrão de treino do subconjunto de validação é medido o erro. Após a fase de validação, o treino é reiniciado para um novo período, e o processo todo é repetido.

Simplificação das etapas de execução do algoritmo

As etapas que requerem a aplicação/execução do algoritmo são cinco:

1. *Inicialização*. Escolher os valores iniciais dos pesos sinápticos e limiares de uma distribuição uniforme de média zero e uma variância (igual ao recíproco do número de conexões sinápticos de um neurónio), de tal modo que o desvio padrão do campo local induzido de um neurónio caia na área de transição entre as partes linear e saturada da sua função de activação sigmóide (definida pela função logística ou a função tangente hiperbólica). Esses valores iniciais não devem ser nem muito elevados e nem tão pouco muito baixos, pois ambos os extremos levam à saturação dos neurónios e perda de velocidade do processo de aprendizagem bem como a operacionalização do algoritmo fica posta em causa por operar em uma área muito plana em torno da superfície de erro. A sugestão é usar valores que estejam entre os dois extremos (Haykin, 2008).
2. *Apresentação dos padrões de treino*. Para cada conjunto de treino é feita uma sequência de computações para frente e para trás para cada padrão do conjunto. A ordem desta apresentação deve ser aleatória, de um conjunto para outro conjunto de treino. Esta aleatoriedade permite (garante) que os exemplos apresentados em um conjunto de treino dificilmente pertencem à mesma classe.

3. *Computação para frente (propagação)*. Aqui o campo local induzido dado pela fórmula (79) agora é representado assim

$$v_j^{(l)}(n) = \sum_{i=0}^{m_l} w_{ji}^{(l)}(n) y_i^{(l-1)}(n) \quad (90)$$

onde, m_l é o tamanho (o número de nós) da camada l ; $y_i^{(l-1)}(n)$ é o sinal de saída do neurónio i na camada $l-1$, na iteração n ; $w_{ji}^{(l)}(n)$ é o peso sináptico do neurónio j da camada l , que é alimentado pelo neurónio i da camada $l-1$.

Para $i=0$, temos $y_0^{(l-1)}(n) = +1$ e $w_{j0}^{(l)}(n) = b_j^{(l)}(n)$ que é o *bias* aplicado ao neurónio j na camada l .

O sinal de saída do neurónio j na camada l , usando a função sigmóide (definida pela função logística ou a função tangente hiperbólica), é

$$y_j^{(l)} = \varphi_j(v_j(n)) \quad (91)$$

Caso o neurónio j esteja na primeira camada oculta, o sinal de saída é

$$y_j^{(0)} = x_j(n) \quad (92)$$

representando assim o j -ésimo elemento do vetor de entrada $x(n)$.

Caso o neurónio j esteja na camada de saída, o sinal de saída é

$$y_j^{(L)} = O_j(n) \quad (93)$$

onde L é a profundidade da rede ($l = 0, 1, \dots, L$) e $O_j(n)$ é o j -ésimo elemento do vetor de saída global.

O sinal do erro calcula-se assim

$$e_j(n) = d_j(n) - O_j(n) \quad (94)$$

4. *Computação para trás (retropropagação)*. Aqui calculam-se os gradientes locais da rede tanto para o neurónio j da camada de saída L , dada pela fórmula (88) e agora

representada pela (95), como para o neurónio j na camada oculta l , dada pela fórmula (89) e agora representada pela (96)

$$\delta_j^{(l)}(n) = e_j^{(l)}(n)\varphi_j'(v_j^{(l)}(n)) \quad (95)$$

$$\delta_j^{(l)}(n) = \varphi_j'(v_j^{(l)}(n))' \sum_k \delta_k^{(l+1)}(n)w_{ki}^{(l+1)}(n) \quad (96)$$

O ajuste dos pesos sinápticos na camada l é feita de acordo com a regra *delta* (δ) generalizada mencionada na fórmula (87) e representada agora por

$$w_{ji}^{(l)}(n) + \alpha[w_{ji}^{(l)}(n-1)] + \eta\delta_j^{(l)}(n)y_i^{(l-1)}(n) \quad (97)$$

5. *Iteração.* Aqui se trata de iterar as computações para frente e para traz, apresentando novos conjuntos de exemplos de treino, até satisfazer um dos critérios de paragem abordados anteriormente.

3.7.8. Redes Neurais na Previsão de Séries temporais

Segundo Ehlers (2005) uma série temporal é um conjunto de observações dispostas ao longo do tempo de forma sequencial. Mas essa sequência, no entender de Morettin (1981), não tem que ser necessariamente igualmente espaçada. Reis (2014) diz ser em intervalos sucessivos e/ou regulares. Quanto ao *tempo*, este pode ser assumido como uma variável do tipo contínuo ou discreto, sendo este último, segundo Kirchgassner e Wolters (2007) o mais geralmente assumido.

Para Cortez (1997), os métodos convencionais [p.ex. o método de Box-Jenkins] que têm sido utilizados na previsão de séries temporais tornam-se inadequados quando as séries apresentam ruídos ou componentes não lineares. Recordemos que os métodos de Box-Jenkins consistem em ajustar os modelos autorregressivos (AR), médias móveis (MA), modelos mistos autorregressivos médias móveis (ARMA) e modelos integrados autorregressivos médias móveis (ARIMA), sendo que os modelos AR são feitos usando as funções de autocorrelação (ACF) e autocorrelação parcial (PACF).

Assim, uma das propostas é a aplicação das Redes Neurais na previsão das séries cujo propósito é descobrir padrões nos dados do passado de modo a extrapolar para o futuro. Para Machado (2010) a RN surge como uma ferramenta que se coaduna bem com os problemas relacionados com a previsão de séries temporais. Mas convém ressaltar que numa situação linear o modelo ARIMA apresenta melhores resultados para previsões de um período à frente do que as RN, mas para previsões de longo prazo as RN têm sido as mais destacadas. Segundo Kao e Huang (2000) citado por Bruni (2000), as RN produzem resultados superiores em qualidade de previsão, pois apresentam menor erro médio. Quanto a Batista (2009) a utilização das redes MLP e RBF melhorou em muito a previsão de séries temporais, dado que elas são estruturas não-lineares.

Quanto aos modelos de previsão modelados por RN, Cortez (1997) apresenta alguns cuidados a ter em consideração: a escolha do conjunto de variáveis de entrada, a arquitetura da rede, as funções de ativação, o ajuste da rede, a escolha dos pesos iniciais e o escalonamento dos dados. Entretanto cita quatro fatores importantes na previsão das séries temporais usando as RN: (i) o número de nodos de entrada, (ii) a função de ativação, (iii) um dos parâmetros do algoritmo de treino RPROP (*Resilient backpropagation*) e (iv) o número de nodos intermédios.

Quanto ao ajuste dos modelos de RN em séries temporais Batista (2009) apresenta uma lista de passos necessários:

- i. Depois de se obter a série temporal, definir se a predição será de um ou vários passos à frente e normalizar os dados, caso necessário;
- ii. Definir quais os dados da série que serão considerados na previsão e separá-los em três conjuntos: treino, validação e teste. Segundo Barros (2014), à medida que os dados do treino vão sendo submetidos à RN, esta, de acordo com o erro, vai ajustando os pesos. A evolução do erro é monitorizada pelo conjunto de validação, que serve também como critério de paragem. Quanto ao conjunto de teste, este serve para avaliar a capacidade de generalização da RN;
- iii. Escolher o tipo de treino da rede e estabelecer os critérios de paragem. Esta acontece quando (A) for atingido o valor mínimo do erro quadrático médio (EQM)

- para os dados de validação, (B) for alcançado um número pré-determinado de épocas de treino ou ambos (A ou B);
- iv. Executar o preditor para fazer previsões utilizando como entrada o conjunto separado para treino;
 - v. Aplicar o critério do EQM para a escolha do melhor modelo que se ajusta aos dados (o melhor modelo é aquele que apresentar com menor erro); [Este critério, bem como outros de uso frequente, nomeadamente o RMSE e o MAPE, estão apresentados na subsecção 3.4.15];
 - vi. Avaliar o preditor com os dados de teste.

A nível computacional as previsões das séries com RN podem ser realizadas, ou com o software IBM SPSS Statistics® através do comando *Analyses-->Neuronal Networks* apresentando apenas duas aplicações de redes: o *Multilayer Perceptron* (MLP) e o *Radial Basis Function* (RBF). Existe também duas opções para seleccionar a estrutura da rede, sendo uma delas de selecção automática. Quando se opta por esta, as funções de activação também são seleccionadas automaticamente pelo SPSS. São três as funções disponíveis: a *tangente hiperbólica*, *sigmóide* e *identidade*. Para avaliar o desempenho da rede, o SPSS calcula duas medidas de erro: a soma dos quadrados dos erros (SSE) e o erro relativo (ER). O SPSS dispõe de dois métodos de optimização do algoritmo: *gradiente conjugado escalado* e o método do *gradiente descendente* (vem por defeito) e também define por defeito a percentagem da amostra para a realização do teste (30%) e para o treino (70%).

Ainda podem ser realizadas com o *Software R*, versão 4.0.5, através, por exemplo, dos pacotes *neuralnet*, *forecast* e *nnfor*, utilizando funções como `neuralnet()`, `nnetar()` e `mlp()` respetivamente, cuja descrição completa das mesmas se encontra em Fritsch *et al.* (2019), Hyndman *et al.* (2018) e em Kourentzes (2019), respetivamente.

Resumidamente, como a série do número de dormidas mensais, em estudo, é sazonal, a função `nnetar()` utiliza a notação RNAR(p,P,k)[m], para simbolizar o modelo, onde **p** é o número de *lags* de entrada; **P** é a ordem de sazonalidade; **size (k)** é o número de neurónios na camada oculta, que por padrão é $\frac{p+P+1}{2}$ (arredondado para o inteiro mais próximo) e **m** o período sazonal. Esta função é utilizada para séries univariadas e

considera a arquitetura *feedforward* com apenas uma camada oculta, utiliza o algoritmo *backpropagation* e usa por padrão a função de ativação *sigmoidal*. O comando básico desta função é a seguinte:

```
nnetar(dados, decay=, p=, P=, size=, lambda=NULL, repeats=,
       subset=NULL)
```

onde, **dados** é a série temporal utilizado para fazer a modelação; **p**, **P** e **Size** já foram mencionados anteriormente; **decay**, é a taxa de aprendizagem; **lambda**, é para informar se vai ser utilizado a transformação de Box-Cox, no caso usa-se a palavra “auto”, do contrário usa-se NULL; o **repeats**, seria o número de redes para ajustar diferentes pesos iniciais definidos aleatoriamente; e o **subset**, é o subconjunto da série utilizada para treino (por defeito está igualado a NULL, que é no caso de não ser necessário ter um conjunto de dados para treino).

Quanto a função `mlp()`, este é um tipo de rede neuronal denominada *Perceptron* Multicamada que considera também a estrutura *feedforward* permitindo o uso de mais do que uma camada oculta. Também será utilizada a notação RNAR(p,P,k)[m] como descrito anteriormente. O seu comando a seguir, tenta especificar automaticamente entradas autorregressivas e qualquer pré-processamento necessário da série. Com os argumentos pré-especificados, por padrão ele treina 20 redes que são usadas para produzir previsões e com um determinado número de camadas ocultas. Vários argumentos estão igualados a NULL por defeito, mas tanto estes como os demais podem ser alterados.

```
mlp(y, m = frequency(y), hd = NULL, reps = 20, comb =
    c("median", "mean", "mode"), lags = NULL, keep = NULL,
    difforder = NULL, outplot = c(FALSE, TRUE), sel.lag =
    c(TRUE, FALSE), allow.det.season = c(TRUE, FALSE), det.type
    = c("auto", "bin", "trg"), xreg = NULL, xreg.lags = NULL,
    xreg.keep = NULL, hd.auto.type = c("set", "valid", "cv",
    "elm"), hd.max = NULL, model = NULL, retrain = c(FALSE,
    TRUE), ...)
```

Neste trabalho é feito uso das funções `mlp()` e `nnetar()` dado serem específicos para modelar séries temporais univariadas como é o caso da nossa série de dormidas mensais

em estudo. Já a função `neuralnet()` mencionado anteriormente, e outras funções semelhantes, não são alvos deste estudo, pois elas se aplicam ao contexto da análise de regressão onde estão presentes variáveis respostas e exógenas.

3.8. PROPOSTA DE ALGORITMO DE ATRIBUIÇÃO DO ERRO

O modelo SARIMA é aquele que traz melhores resultados dentre os modelos ARIMA da metodologia de Box-Jenkins, em se tratando de séries temporais univariadas com características sazonais. A questão, que surge de uma ideia intuitiva, diante da série em estudo, é: os resultados de um modelo SARIMA podem ser ainda melhorados usando esses mesmos resultados?

Seja M1 e M2 dois modelos SARIMA escolhidos que melhor modelam os dados de uma série de frequência mensal. Considera-se ainda que M2 seja o modelo final escolhido (utilizando os critérios de escolha como o AIC/AICc ou o erro de teste) e que melhor prevê os valores futuros dessa série, portanto pressupõe-se que possui o menor erro MAPE, ou seja, $MAPE_{M_2} < MAPE_{M_1}$.

Cada um dos modelos gera um conjunto de valores previstos. Sendo assim denominemos por $A = \{a_i\}, i = 1, \dots, 12$ e $B = \{b_i\}, i = 1, \dots, 12$ os conjuntos de valores previstos pelos modelos M1 e M2 respetivamente.

M1	jan	fev	dez
A	<i>a1</i>	<i>a2</i>	<i>a12</i>

M2	jan	fev	dez
B	<i>b1</i>	<i>b2</i>	<i>b12</i>

Tendo os dois conjuntos A e B, é criado um terceiro conjunto de valores previstos $C = \{c_i\}, i = 1, \dots, 12$ que é formado pela média dos valores a_i e b_i dos conjuntos A e B, respetivamente, isto é:

$$c_i = (a_i + b_i)/2, \tag{98}$$

	jan	fev	dez
C	<i>c1</i>	<i>c2</i>	<i>c12</i>

Antes de considerarmos o algoritmo, calcula-se o erro para cada um dos valores dos conjuntos *B* e *C*, assumindo que *B* seja o conjunto dos valores observados, bem como o erro:

$$MAPE_i = \left| \frac{b_i - c_i}{b_i} \right| \times 100\% \quad (99)$$

onde, *m* o número de meses.

E ainda a raiz quadrada do erro total.

$$MAPE_{total} = \left(\frac{1}{m} \sum_{i=1}^m MAPE_i \right)^{1/2} ; i = 1, 2, \dots, m \quad (100)$$

Obs.: A raiz do erro permite que o erro aumente de forma lenta em cada iteração, e esta lentidão contribui para melhor controlo no critério de paragem descrito mais adiante.

Tendo calculado o $MAPE_{total}$ inicia-se a aplicação do Algoritmo de Atribuição do Erro (AAE), com o objetivo de se encontrar um novo conjunto de valores previstos C^* que melhor possa aproximar aos dados a observar no futuro, ou seja de modo que exista um $MAPE_{total}^* < MAPE_{M_2}$.

Em síntese: Dado dois conjuntos iniciais de valores previstos *A* e *B*, o algoritmo tem o seu início depois que é construído um novo conjunto de valores esperados *C* e do erro calculado $MAPE_i$. A cada conjunto gerado C', C'', \dots, C^k calcula-se o erro $MAPE_{total}$, respectivo. Este erro é usado para gerar um novo conjunto de valores previstos, iteração por iteração. Cada iteração é composta por dois passos essenciais: (i) atribuição do erro ao conjunto anterior para gerar o novo conjunto de valores previstos; e, (ii) o cálculo do novo erro de previsão que será atribuído à próxima iteração.

As iterações vão sendo realizadas enquanto o erro da última iteração for menor que o da iteração anterior. Este algoritmo de atribuição do erro é um algoritmo de otimização do erro e aplica-se aos valores iniciais previstos por um modelo SARIMA.

---> O algoritmo consiste em:

1. Atribuir o erro $MAPE_{total}$ à cada valor previsto do conjunto C . Isto gera, por meio da fórmula (101) um novo conjunto de valores previstos que se denota por C' .

$$C'_i = |c_i(MAPE_{total} + 1)| \quad (101)$$

2. Calcular o erro $MAPE_i$ para o novo conjunto de valores previstos e ainda a raiz quadrada do erro total que se denota por $MAPE_{total_1}$. É a primeira iteração.

$$MAPE_i = \left| \frac{b_i - c'_i}{b_i} \right| \times 100\% \quad (102)$$

$$MAPE_{total_1} = \left(\frac{1}{m} \sum_{i=1}^m MAPE_i \right)^{1/2} \quad (103)$$

3. Atribuir o erro da fórmula (103) à cada valor previsto do conjunto C' . Isto gera um novo conjunto de valores previstos que se indica por C'' .

$$C''_i = |c'_i(MAPE_{total_1} + 1)| \quad (104)$$

4. Calcular o erro $MAPE_i$ para o novo conjunto de valores previstos e ainda o erro total que se indica por $MAPE_{total_2}$. É a segunda iteração.

$$MAPE_i = \left| \frac{b_i - c''_i}{b_i} \right| \times 100\% \quad (105)$$

$$MAPE_{total_2} = \left(\frac{1}{m} \sum_{i=1}^m MAPE_i \right)^{1/2} \quad (106)$$

E assim sucessivamente.

Critério de paragem:

O algoritmo permite gerar novos conjuntos de valores previstos na expectativa de encontrar um erro $MAPE_{total}^* < MAPE_{M_2}$. A partir do ponto de inflexão (Figura 3.5) e até um determinado momento as iterações seguintes produzem bons resultados, isto é, em

uma série de iterações, a medida que o erro da iteração aumenta é suposto diminuir o erro entre os valores previstos e observados. Mas para que o algoritmo seja finalizado e poder encontrar o melhor entre os bons resultados, é feita a comparação entre os erros de cada iteração com o erro teórico pré-definido no valor igual a 2, ou seja assim que o erro em uma determinada iteração k for maior que 2, isto é, se o $MAPE_{total}^k > 2$ (na Figura 3.5 é o primeiro ponto acima da reta horizontal tracejada), então o algoritmo finaliza-se e é encontrado o conjunto C^k cujo erro se prevê que seja o menor dos menores entre as iterações anteriores e também que seja menor que $MAPE_{M_2}$. A pré-definição do valor 2 como sendo o limite de erro gerado pelas iterações que o algoritmo deve realizar para encontrar o conjunto C^k , advém de uma constatação dos resultados obtidos em cada iteração (secção 4.2.7), ainda que seja por tentativa e erro.

3.9. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Tanto métodos quantitativos como qualitativos têm sido usados na previsão da procura turística (Formica, 2006; Song *et al.*, 2013). Tome-se como exemplo, os métodos de julgamento nomeadamente o modelo Delphi. Com relação aos métodos quantitativos temos as séries temporais, econométricos e inteligência artificial. Estes são os predominantes nos estudos e se combinados com os métodos de julgamento podem trazer melhorias na precisão de previsão.

Os modelos que mais têm sido destacados nos estudos são os de séries temporais e econométricos. Song, Qiu e Park (2019) constataram que as abordagens relacionadas com AIDS, PDR, VECM, VAR e STSM tiveram melhores desempenho. Quanto aos métodos de séries temporais, o método de Box-Jenkins e suas variáveis exógenas também obtiveram desempenho superior aos demais. No campo econométrico, o modelo ADLM também ganhou popularidade e tem sido assim como o ECM, VAR e TVP os mais aplicados e os que melhores resultados trazem, segundo Song e Li (2008). Porém não foram utilizados neste trabalho por dois motivos: (i) por falta de dados mensais ao longo do horizonte temporal da série (janeiro de 2000 a dezembro de 2018) para as variáveis exógenas propostas inicialmente, no Plano da Tese, para estudo e (ii) como não se conhece estudos de previsão da procura turística em Cabo Verde com uso de metodologias de modelação

de séries temporais, obtou-se por começar a utilizar apenas a série univariada, deixando estes métodos para serem explorados no futuro.

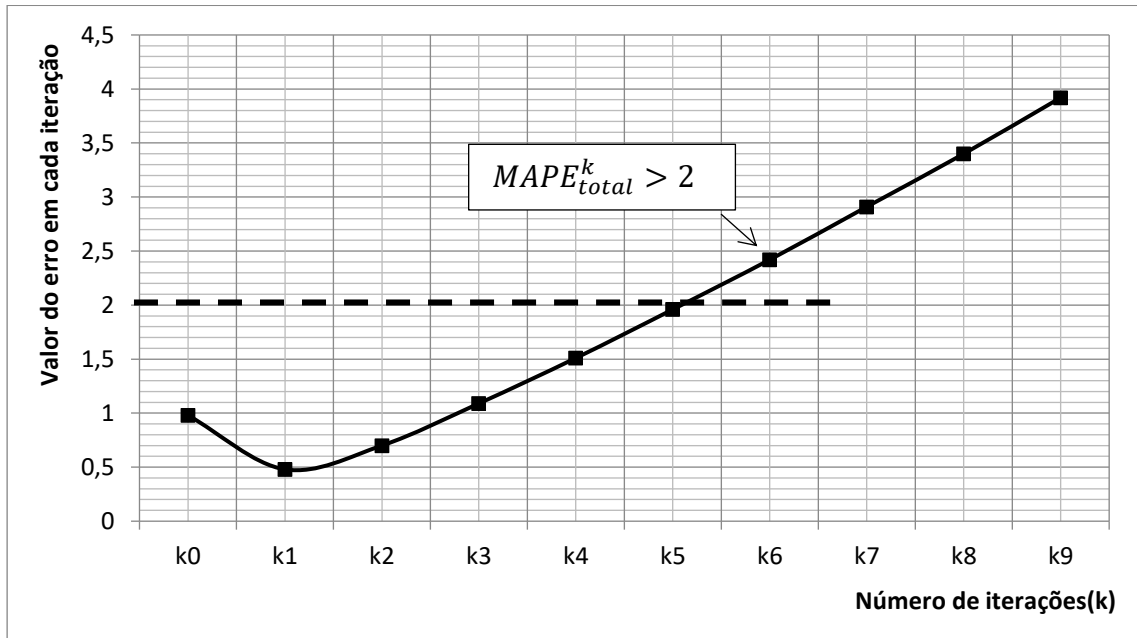


Figura 3.5: Exemplificação do efeito gráfico do comportamento do erro em cada iteração e sua escolha.

Fonte: Elaboração própria

Apesar dos modelos mais destacados, existem muitas incertezas quanto ao melhor modelo de séries temporais e a seleção dos fatores que determinam a procura turística é muito mais diversificada do que sua medição. Isto é devido aos vários objetivos de pesquisa de cada estudo.

Os modelos baseados em IA têm vindo a ganhar espaço e tem mostrado ser superior aos modelos de séries temporais e econométricos quanto a previsão da procura turística, apesar de ser muito criticado devido a suas limitações teóricas e tecnológicas e de não conseguir analisar a procura turística do ponto de vista económico; o que restringe o escopo de aplicações práticas. A rede neuronal é um dos modelos mais utilizados, com destaque para a rede *perceptron*, a rede *feedforward* e a rede de Elman. No futuro a combinação de modelos IA com modelos econométricos pode trazer grandes vantagens na previsão turística.

De todos os métodos e modelos abordados é de notar, segundo Song, Qiu e Park (2019), um crescimento do uso de dados da Internet na previsão da procura turística.

Como foi constatado, alguns modelos tendem a ser superiores aos demais, porém, segundo Song e Li (2008) o desempenho dos modelos dependem da frequência dos dados, das origens, dos destinos e do horizonte de previsão. Entretanto, diferentes modelos têm vindo a ser combinados criando assim um novo grupo de métodos de previsão, os chamados métodos híbridos. Apesar da sua aparição tímida nos estudos, têm mostrado um bom desempenho na previsão da procura turística.

O desempenho das previsões feitas pelos modelos apresentados no estado da arte foi baseado, em grande maioria, na utilização do NMSE (*Normalized Mean Square Error*), do MAPE (*Mean Absolute Percentage Error*) e do RMSPE (*Root Mean Square Percentage Error*).

Poucos foram os estudos que utilizaram dados mensais na previsão da procura turística. Desses poucos, a maioria aplicou métodos de séries temporais. Entretanto dados mensais têm vindo a ganhar mais interesse por parte dos tomadores de decisão e políticos no sector turístico devido à questão da sazonalidade.

Mesmo que os modelos de séries temporais tenham sido os mais utilizados, os modelos econométricos tendem a vir ultrapassá-los dado ao grande esforço que os formuladores de políticas e tomadores de decisão têm vindo a empreender para entender os principais fatores que determinam a procura turística. Tal objetivo não pode ser alcançado por meio de modelos não causais de séries temporais.

O método de Holt-Winters é um método a ser aplicado na série temporal em estudo dado que nele é contemplado a componente de sazonalidade, pelo que se espera, juntamente com a proposta apresentada, que tenha um bom desempenho nas previsões.

CAPÍTULO 4: RESULTADOS

4.1. INTRODUÇÃO

Neste capítulo são apresentados todos os resultados e descrita a forma como foi feita a recolha, o tratamento e a análise dos dados para cada um dos estudos realizados ao longo do presente trabalho de investigação, sendo eles:

- O Ciclo de Vida e o Índice de Desenvolvimento Turístico;
- Aplicação da Lei de Benford ao número de dormidas mensais nos estabelecimentos turísticos da ilha do Sal;
- Aplicação do método de Box-Jenkins e do seu modelo SARIMA, dado ser o método que melhores resultados trazem a nível de previsão de séries temporais, no tocante a obtenção de modelos lineares;
- Aplicação do Método de Suavização Exponencial Tripla (conhecido como Método de Holt-Winters);
- Aplicação da inteligência artificial, com Redes Neurais Artificiais, em específico a *rede feedforward*.

A Regressão Linear Múltipla não foi utilizada neste trabalho, embora tenha sido considerada, dada a inexistência de dados mensais ao longo do horizonte temporal da série (janeiro de 2000 a dezembro de 2018) para variáveis exógenas ponderadas, tais como o índice de preços no consumidor em Cabo Verde e nos mercados emissores; taxa de desemprego na ilha do Sal e nos mercados emissores; estadia média dos turistas na ilha do Sal; entre outras.

Ainda neste capítulo é apresentado o resultado da proposta de melhorar dos resultados previstos pelo modelo SARIMA. No final é feita a comparação entre todos os métodos e modelos com relação aos resultados de previsão do número de dormidas mensais, ou seja, vai concluir-se pelo modelo que melhor consegue prever o número de dormidas mensais para o ano de 2019 com o menor erro possível.

Ainda, a nível computacional é utilizada diversas ferramentas para analisar a série temporal, através dos pacotes do *Software R* versão 4.0.5, nomeadamente:

- O **forecast**, que permite utilizar diferentes modelos ARIMA, analisar as funções de autocorrelação, escolher de forma automática um determinado modelo por meio da função `auto.arima()`, elaborar gráficos, fazer testes e previsões, entre outras cuja descrição completa se encontra em Hyndman *et al.* (2021);
- O **ggplot2** que permite a apresentação de diversos tipos e estilos de gráficos, cujos detalhes se encontram em Wickham *et al.* (2021);
- O **tseries** que, dentre muitas outras funções (Trapletti, *et al.*, 2021), ajuda a estruturar os dados em um formato de séries temporais;
- O **astsa** para analisar, por exemplo, as funções de autocorrelação e ainda contém a função `Sarima` (Stoffer, 2021);
- O **urroot** que permite realizar o teste de Dikey-Fuller (De Lacalle & Boshnakov, 2020);
- O **randtests** que permite fazer testes para verificar a tendência (Caeiro & Mateus, 2014);
- O **seastests** que permite fazer testes para verificar a sazonalidade (Ollech, 2021).

As respetivas funções e seus comandos se encontram ao longo deste capítulo e no Anexo V.

4.2. RESULTADOS

4.2.1. Ciclo de Vida e o Índice de Desenvolvimento Turístico

Antes de se avançar com este estudo considerou-se dois pontos para análise que justificam ser entendidas devido a várias opiniões diferentes relativamente ao estado do turismo na ilha do Sal e conseqüentemente em Cabo Verde. Assim, estabeleceram-se os seguintes pontos a serem analisadas para o período de 2010 a 2018:

(1) a fase do ciclo de vida turística que o destino ilha do Sal se encontra individualmente, comparando-o com o destino Cabo Verde e seu posicionamento relativo a nível de competitividade com destinos com o mesmo mercado de ‘sol e praia’ e que procuram os mesmos mercados emissores, neste caso o mercado europeu;

(2) qual dos destinos utilizar para analisar a competitividade por meio do Índice de Desenvolvimento Turístico (IDT): Cabo Verde ou Ilha do Sal;

Para se conseguir encontrar as respostas face aos cenários levantados neste estudo, recorreu-se a uma pesquisa bibliográfica referente a ilha do Sal e o país onde se insere, do modelo do Ciclo de Vida de um destino turístico proposto por Butler (1980) como pelo Índice de Desenvolvimento Turístico (IDT) proposto por Coelho e Butler (2012) e por Coelho e Lourenço (2015). Foi levado a cabo a análise do ciclo de vida do destino ilha do Sal e calculado o seu IDT (expressão (5)) comparando com destinos com o mesmo tipo de turismo – ‘sol e praia’.

Os destinos escolhidos foram: Canárias, Marrocos, Tunísia e República Dominicana, por serem também as que possuem curta distância com relação ao mercado europeu. Para as análises optou-se por utilizar dados dos anos de 2010 e 2018 para cada um dos fatores presentes na fórmula (3), por forma a fazer as devidas comparações a nível de competitividade. Serão consideradas as variáveis “PIB”, “População Residente”, “Receita Turística” e “Turistas”, como definidos anteriormente.

Para o apuramento da posição relativa de cada país no seu contexto competitivo para o mercado considerado de ‘sol e praia’ utilizou-se o Microsoft Excel. Os dados para sustentar os cálculos foram recolhidos em fontes governamentais e não governamentais. Os valores monetários recolhidos nas fontes se encontravam em euros, dólares, dirhams e ECV, tendo sido convertido todos para euros utilizando o conversor de moedas do Banco de Portugal em www.bportugal.pt.

Sendo a ilha do Sal parte de Cabo Verde e ambas possuem o mesmo clima e o mesmo tipo de turismo (sol e praia), fez-se os mesmos cálculos para Cabo Verde, no sentido de verificar a importância da ilha e do arquipélago.

De acordo com MF (2019), a ilha do Sal como destino turístico, e dentro do segmento sol e praia, está na terceira fase de seu ciclo de vida, isto é, numa fase de desenvolvimento. Recorde-se que as razões características para esta fase são: em épocas altas, o número de turistas é maior que a população; já existe pressão sobre o meio ambiente, infraestrutura básica e urbana; investidores estrangeiros estão presentes na ilha; o número de ofertas está aumentando e o produto está começando a diversificar; e melhores condições de vida para a população.

Em relação ao modelo de ciclo de vida proposto por Butler (1980), descobriu-se que o ciclo de vida turístico de Cabo Verde em 2010 estava na fase de exploração. O plano estratégico para o desenvolvimento do turismo começa o processo de promoção de Cabo Verde como uma marca de turismo, e em 2018 o país já está na fase de desenvolvimento (o número de turistas é maior que a população do país; grandes cadeias de hotéis, principalmente na ilha do Sal; já existe uma certa pressão sobre o meio ambiente e a infraestrutura básica, especialmente nas ilhas mais turísticas do Sal e Boa Vista; e os padrões de vida de Cabo Verde melhoram principalmente nessas duas ilhas). Em relação à ilha do Sal, em 2010, encontrava-se na fase de Desenvolvimento, a mesma fase que em 2018 (ver Figura 4.1), pelas razões descritas no Capítulo 2.



Figura 4.1: Fase do Ciclo de Vida de Cabo Verde e da ilha do Sal como destino turístico, 2010 e 2018.

Fonte: Adaptado de Neves, Nunes e Fernandes (2020, pp.13-14)

4.2.1.1. Índice de Desenvolvimento Turístico aplicado a Cabo Verde

Com os valores obtidos em 2010 (Tabela 4.1), as Ilhas Canárias lideraram o mercado com um IDT de 0,9576, que corresponde à fase de Desenvolvimento. Cabo Verde (0,0044) estava na fase de exploração, assim como a República Dominicana (0,1796). Tunísia (0,2483) e Marrocos (0,2188) estavam na fase de Envolvimento. Em relação a 2018, Cabo Verde tem um IDT de 0,0002, revelando uma perda de posição dentro da sua fase e em pior situação competitiva; a República Dominicana também perde a posição (0,0099); e a Tunísia (0,5493) mostra uma melhoria na sua posição relativa, sendo o destino que mais cresceu em 2018, permanecendo na mesma fase, apesar de seu crescimento de cerca de 121%. O posicionamento relativo dos destinos restantes foi alterado: as Ilhas Canárias têm um IDT de 0,0265, revelando uma mudança da fase de Desenvolvimento para a fase de Exploração; Marrocos (0,0061), revelando também uma mudança negativa, passando da fase de Envolvimento para a fase de Exploração (ver os limites de cada fase na Tabela 2.3), traduzindo-se para a possibilidade de um declínio, apesar de aumentar o número de turistas em 32%.

Tabela 4.1: Determinação do IDT – Cabo Verde.

Destinos	Ano	PIB Líquido (€)	Receita Turística (€)	População (no.)	Turistas (no.)	IDT
Cabo Verde	2010	1255000	199755	491875	381831	0,0044
	2018	1676000	302027	543492	765696	0,0002
Ilhas Canárias	2010	40811000	8028000	2064876	10649727	0,9576
	2018	45720000	12895000	2207225	15559791	0,0265
República Dominicana	2010	41683000	3859063514	9479000	4125000	0,1796
	2018	72589000	6954102921	10266000	6500000	0,0099
Marrocos	2010	70305000	7580899397	32182000	9288000	0,2188
	2018	100414000	6616376776	35220000	12300000	0,0061
Tunísia	2010	33231000	3223922114	10640000	7828000	0,2483
	2018	33760000	1261010663	11565204	140000000	0,5493
Total	2010	187285000	14672112781	54857751	32272558	1,6088
	2018	254159000	14844687387	59801921	1435125487	0,5921

Fonte: Neves, Nunes e Fernandes (2020, pp.14-15)

4.2.1.2. Índice de Desenvolvimento Turístico aplicado à ilha do Sal

Analisando a Tabela 4.2, pode-se ver que, em 2010, as Ilhas Canárias lideraram o mercado de "sol e praia" com um IDT de 0,9619, situando-se no intervalo 0,577 – 1,401 que corresponde à fase de Desenvolvimento. A ilha do Sal (0,0039) estava na fase de Exploração, bem como a República Dominicana (0,1798). Marrocos (0,2191) e Tunísia (0,2486) estavam na fase de Envolvimento. Em relação a 2018, as posições relativas permaneceram para três destinos: a ilha do Sal (0,0002), revelando uma perda de posição dentro de sua fase e em pior situação competitiva; a Tunísia melhora sua posição (0,5467), sendo o destino que mais cresceu em 2018, mantendo a mesma fase apesar de seu crescimento em torno de 120%; a República Dominicana (0,0099) também perde a posição. As Ilhas Canárias (0,0265) têm uma mudança da fase de Desenvolvimento para a fase de Exploração; Marrocos (0,0061) tem uma mudança negativa, passando da fase de Envolvimento para a fase de Exploração, traduzindo-se na possibilidade de um declínio.

Tabela 4.2: Determinação do IDT - ilha do Sal.

Destinos	Ano	PIB Líquido (€)	Receita Turística (€)	População (no.)	Turistas (no.)	IDT
Ilha do Sal	2010	143717	26013	25779	154115	0,0039
	2018	197197	196445	38243	379020	0,0002
Ilhas Canárias	2010	40811 000	8028000	2064876	10649727	0,9619
	2018	45720000	12895000	2207225	15559791	0,0265
República Dominicana	2010	41683 000	3859063514	9479000	4125000	0,1798
	2018	72589000	6954102921	10266000	6500000	0,0099
Marrocos	2010	70305 000	7580899397	32182000	9288000	0,2191
	2018	100414000	6616376776	35220000	12300000	0,0061
Tunísia	2010	33231 000	3223922114	10640000	7828000	0,2486
	2018	33760000	1261010663	11565204	140000000	0,5467
Total	2010	26545 395	14844581804	59296672	1434738811	1,6134
	2018	260545395	14844581805	59296672	1434738811	0,5893

Fonte: Neves, Nunes e Fernandes (2020, p.16)

4.2.1.3. Considerações Finais

A situação competitiva, considerando destinos semelhantes, de Cabo Verde e da ilha do Sal é o mesmo, o que mostra que permaneceram na fase de Exploração por um período de oito anos (2010-2018). Dada a sua perda relativa de posição, é uma indicação de que os dois destinos estão em uma fase Estagnada. Neste espaço competitivo (mercado de “sol e praia”), os concorrentes da ilha do Sal/Cabo Verde não evoluíram em termos de turismo, com exceção da Tunísia, que apesar de ter crescido mais de 100%, não mudou em termos relativos, o que indica uma certa estagnação. O destino da República Dominicana estagnou, e os destinos de Marrocos e das Ilhas Canárias encontram-se em declínio.

As escolhas estratégicas a serem adotadas tanto para a ilha do Sal como para Cabo Verde, para o mercado de ‘sol e praia’ devem levar em consideração a posição relativa e a dinâmica dos concorrentes diretos. Com referência a Cabo Verde constatou-se que a República Dominicana foi o país que mais cresceu com relação ao PIB e às Receitas Turísticas. A Tunísia foi quem cresceu mais em número de turistas e Cabo Verde cresceu mais a nível demográfico nos últimos oito anos. Com referência à ilha do Sal, esta foi a que mais cresceu a nível do PIB, da Receita Turística e da população.

Para que Cabo Verde e a ilha do Sal entre na fase de Envolvimento, neste mercado, o IDT deverá crescer 43% e 48%, respetivamente, considerando que seus concorrentes se mantêm sem grandes melhorias no seu produto e oferta turística. Este cenário, como é razoável supor, não é atingível a curto prazo e por isso a melhoria competitiva deve ser pensada a longo prazo e com medidas mais agressivas para se poderem afastar dos concorrentes mais diretos.

Em termos globais, há uma certa redução das desigualdades neste mercado, uma vez que o IDT total passou de 1.6133 em 2010 para 0,5893 em 2018, indicando que a ilha do Sal tende a melhorar em termos de competitividade. Nesse sentido, a ilha aproximou-se dos destinos em desenvolvimento, embora de maneira menor. O mesmo acontece com Cabo Verde.

A análise da competitividade com destinos com o mesmo mercado de ‘sol e praia’ pode ser feita utilizando tanto os dados de Cabo Verde como os da ilha do Sal, pois os resultados em termos de posicionamento relativo e mudança de fase são os mesmos, durante o período analisado. E, sendo a ilha do Sal a que tem mais peso turístico, principalmente nas receitas turísticas, pode-se dizer que ela é a referência do destino turístico Caboverdiano. Estas considerações encontram-se publicado em Neves, Nunes e Fernandes (2020) preparado no âmbito desta tese.

4.2.2. Análise da Série Temporal

A Figura 4.2 apresenta a série temporal referente ao número de dormidas mensais nos estabelecimentos turísticos da ilha do Sal, no período entre 2000 e 2018. É visível a existência de oscilações irregulares que sugerem uma não estabilização da média e a presença de uma sazonalidade (valores máximos nos meses de inverno: entre outubro e dezembro, e valores mínimos nos meses de verão: entre maio e setembro), indicando que os valores das dormidas em estabelecimentos turísticos dependem da época do ano. Apresenta também uma tendência de aumento após o ano de 2010. Embora, no final de 2007 e início de 2008, apresentasse valores elevados quando comparados com o ano anterior. Desde 2014, tem havido um aumento significativo devido à promoção da ilha e mesmo do arquipélago de Cabo Verde entre os potenciais visitantes da Europa. Visitantes estes descritos no Capítulo 2.

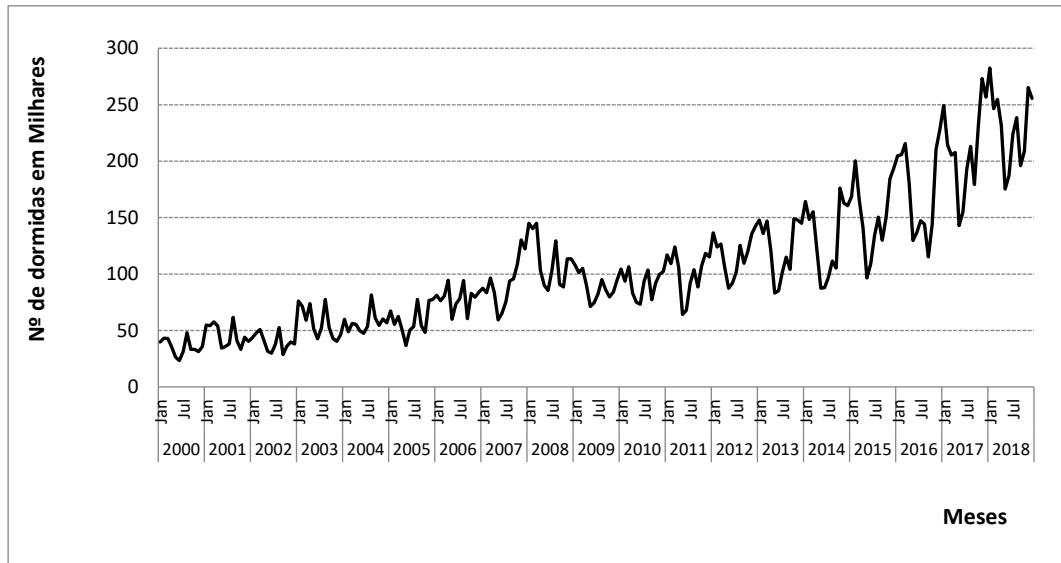


Figura 4.2: Número mensal de dormidas nos estabelecimentos turísticos da ilha do Sal [2000:01-2018:12].

Fonte: Elaboração própria, com dados cedidos pelo INE (Anexo III)

Como se pode ainda verificar, a sazonalidade vai sofrendo alguma mudança com o tempo se comparado com o início da série, o que é indicativo de que a série possui uma sazonalidade multiplicativa.

Em seguida apresenta-se (Figura 4.3) a decomposição multiplicativa da série nas três componentes: tendência, sazonalidade e aleatoriedade, utilizando a função `decompose()` que pode ser encontrado no pacote **forecast** do R (Hyndman *et al.*, 2021)

```
> autoplot(decompose(dados))
```

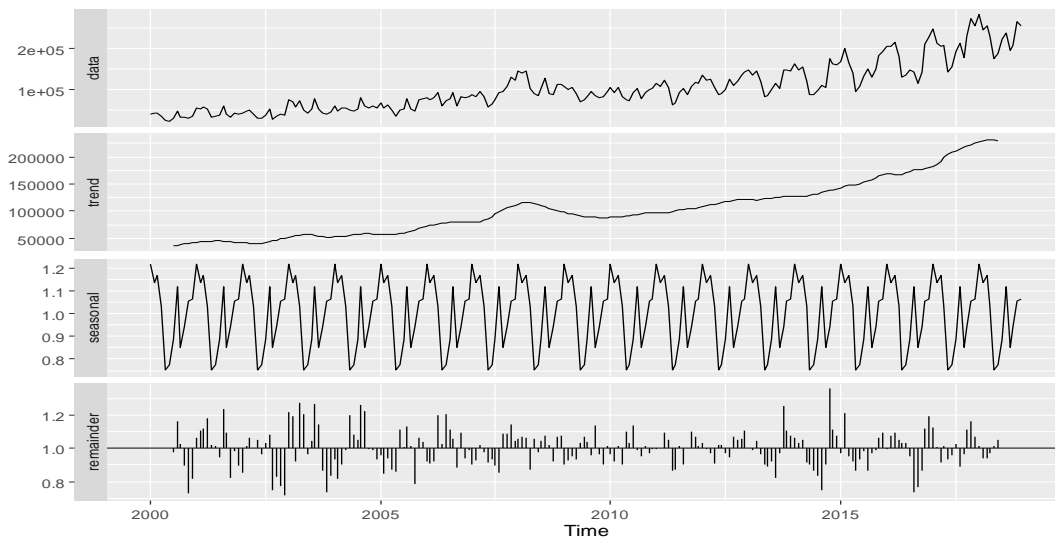


Figura 4.3: Decomposição multiplicativa da série de dormidas nos estabelecimentos turísticos da ilha do Sal [2000:01-2018:12].

Fonte: Elaboração própria

Utilizando medidas de dispersão como o desvio-padrão, a amplitude inter-quartil e o coeficiente de variação é possível compreender a variabilidade dos dados da série (Tabela 4.3). Quanto maior o valor destas medidas mais dispersos se encontram os dados em relação à média. Caso se considere apenas o desvio-padrão como medida para avaliar a dispersão, então o ano que apresenta uma maior dispersão de valores é o ano de 2016. Chega-se à mesma conclusão pelos valores da medida da amplitude inter-quartil. Interessante, ainda, verificar que a partir do ano de 2013 a dispersão aumenta. Quanto ao coeficiente de variação verifica-se que o ano de 2010 apresenta maior homogeneidade, com um valor aproximado de 12,6%. Já o ano de 2014 apresenta a maior dispersão em comparação com os outros anos.

A análise descritiva juntamente com a análise gráfica permite obter mais informações sobre a série e auxilia na escolha do modelo que melhor se irá adequar às características dos dados da série.

Tabela 4.3: Estatísticas descritivas do número mensal de dormidas nos estabelecimentos hoteleiros da ilha do Sal [2000:01-2018:12].

Anos	Soma	Média	Mediana	Desvio-padrão	Coefficiente de Variação (%)	Quartil inferior (Q1)	Quartil superior (Q3)	Amplitude inter-quartil
2000	424276	35356,3	34399	6901,3	19,5	31154,8	42166,8	11012,0
2001	548901	45741,8	42530,5	9621,2	21,0	36411,5	54615,0	18203,5
2002	478385	39865,4	38987,5	7486,1	18,8	32743,0	46757,8	14014,8
2003	685511	57126,0	52392,5	13389,0	23,4	43737,3	73111,0	29373,8
2004	685198	57099,8	55722,5	8470,8	14,8	50680,0	60096,0	9416,0
2005	709982	59165,2	54851,5	12584,5	21,3	50388,8	74262,0	23873,3
2006	945421	78785,1	80135,5	10238,5	13,0	74171,3	83815,8	9644,5
2007	1101642	91803,5	90691,0	20274,0	22,1	77198,5	105573,5	28375,0
2008	1347076	112256,3	108274,0	21638,6	19,3	90148,8	137376,3	47227,5
2009	1073300	89441,7	88061,0	11276,7	12,6	80495,8	99743,3	19247,5
2010	1104004	92000,3	93565,0	11582,0	12,6	78623,8	103298,5	24674,8
2011	1214066	101172,2	106758,0	18518,5	18,3	89507,0	116471,5	26964,5
2012	1406544	117212,0	122070,0	17121,6	14,6	102723,0	133195,0	30471,6
2013	1481980	123498,0	128279,0	24208,5	19,6	102449,4	147514,6	45065,3
2014	1577406	131450,5	134560,5	31548,5	24,0	99295,8	162312,8	63017,0
2015	1823603	151966,9	150267,0	30889,8	20,3	130993,3	180375,0	49381,8
2016	2062034	171836,2	164032,0	37880,3	22,0	138488,8	209277,5	70788,8
2017	2519487	209957,3	210328,5	37567,0	17,9	182543,5	244486,3	61942,8
2018	2765727	230477,3	235049,5	31696,0	13,8	199275,8	255258,0	55982,3

Fonte: Elaboração própria

4.2.3. Aplicação da Lei de Benford

Para aplicar a Lei de Benford, foram utilizados dados mensais do número de dormidas nos estabelecimentos turísticos da ilha do Sal, no período entre janeiro de 2000 e dezembro de 2018 (os dados se encontram em Anexo III), providenciados pelo Instituto Nacional de Estatística de Cabo Verde (INE). Assim, para comparar as frequências observadas do primeiro dígito do número de dormidas mensais com as frequências esperadas segundo a Lei de Benford, foi utilizado o teste estatístico de comparação de frequências Qui-Quadrado (χ^2). Este teste permite verificar a existência (ou não) de discrepâncias entre as frequências observadas e as esperadas, e é dado por Ross (2009):

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \quad (107)$$

onde,

o_i – Frequência relativa observada do i dígito, onde $i = 1, 2, \dots, 9$;

e_i – Frequência relativa esperada do i dígito, onde $i = 1, 2, \dots, 9$.

Na aplicação do teste é necessário a formulação da hipótese nula (H_0) e da hipótese alternativa (H_1):

H_0 : *O número de dormidas mensais registados nos estabelecimentos turísticos da ilha do Sal segue a Lei de Benford.*

H_1 : *O número de dormidas mensais registados nos estabelecimentos turísticos da ilha do Sal não segue a Lei de Benford.*

Para aceitar ou rejeitar a Hipótese Nula, existem duas regras:

- Se χ^2 calculado (equação (107)) for maior ou igual ao χ^2 crítico ($\chi^2_{(k-1);\alpha}$), então rejeitamos a hipótese nula, para um determinado nível de significância α .
- Se χ^2 calculado (equação (107)) for menor que χ^2 crítico, então não rejeitamos a hipótese nula, para um determinado nível de significância α .

Com base nos dados apresentados no Anexo III, na Figura 4.2 e na seção 2.2, pode-se afirmar que:

- O dígito 1 como o primeiro dígito aparece pela primeira vez quando o número de dormidas atinge seis dígitos, o que ocorre no último trimestre do ano de 2007. A partir de então e até ao ano de 2015, todos os números de dormidas com seis dígitos começam com o dígito 1, com a exceção do mês de fevereiro de 2015;

O dígito 2 como primeiro dígito aparece em 2000 algumas vezes quando o número de dormidas é de cinco dígitos, e só reaparece com maior frequência a partir de 2015, quando ocorre a reestruturação e abertura de grandes hotéis internacionais (ver secção 2.2.11). Em 2018, 2 é o primeiro dígito em praticamente todos os meses. Isso pode

ser devido aos seis eventos nacionais (finais dos campeonatos de Andebol, Basquetebol, Karaté, Atletismo, Boxe e Xadrez) e um evento internacional (*Kite-Surf World Tour*) que ocorreu em sete meses daquele ano;

- O número de dormidas tem vindo a aumentar com a abertura de hotéis de renome internacional, pontuais eventos desportivos e culturais internacionais e nacionais; sendo que a percentagem observada de números com o dígito 1 como primeiro dígito superou o esperado, uma vez que a falta de eventos significativos levou a um lento aumento do número de dormidas de 100.000 para 200.000;
- O intervalo de valores que começa com o dígito 1 ou com o dígito 2 é mais amplo do que o intervalo de valores que começa com os outros dígitos. Dado que se verifica uma tendência de aumento visível do número de dormidas; existem diferentes taxas de crescimento do número de dormidas, evidenciando uma possível alteração da procura turística após 2014; e, há um possível comportamento oscilatório/aleatório subjacente dos dados, que poderia ser parcialmente explicado pela lei de Benford.

Para este estudo, o cálculo do Qui-Quadrado χ^2 baseou-se na soma dos nove dígitos possíveis para ocupar a primeira posição no conjunto de dados (Anexo III). Para um nível de significância de 5% e com 8 graus de liberdade ($gl = 9 - 1 = 8$), temos, ao consultar a tabela Qui-Quadrado que o valor crítico de χ^2 é 15,507 (Ross, 2009).

A Tabela 4.4 apresenta as etapas realizadas no Excel para calcular o valor de χ^2 para o conjunto de dados. Nesse caso, o χ^2 calculado (= 28,940) é maior que o χ^2 crítico ($\chi^2_{(9-1);0,05} = 15,507$), portanto a hipótese nula (o número de dormidas mensais nos estabelecimentos turísticos da ilha do Sal segue a Lei de Benford) é rejeitada, ao nível de significância de 5%. A decisão é que o número de dormidas mensais em estabelecimentos turísticos na ilha do Sal no período entre 2000 e 2018 não obedece à Lei de Benford. Esses resultados foram confirmados pelo software IBM® SPSS® *Statistics versão 20*, com um valor de *p-value* <0,001.

Tabela 4.4: Cálculo do valor do χ^2 para o conjunto de dados do número de dormidas.

Dígito (d)	o_i	Valor Observado (%)	e_i	Valor Esperado (%)	$(o_i - e_i)$	$(o_i - e_i)^2$	$\frac{(o_i - e_i)^2}{e_i}$
1	79	34.6%	68,635	30.1%	10,365	107,437	1,565
2	27	11.8%	40,149	17.6%	-13,149	172,891	4,306
3	17	7.5%	28,486	12.5%	-11,486	131,929	4,631
4	17	7.5%	22,095	9.7%	-5,095	25,964	1,175
5	23	10.1%	18,053	7.9%	4,947	24,470	1,355
6	9	3.9%	15,264	6.7%	-6,264	39,236	2,571
7	18	7.9%	13,222	5.8%	4,778	22,828	1,726
8	21	9.2%	11,663	5.1%	9,337	87,184	7,475
9	17	7.5%	10,433	4.6%	6,567	43,129	4,134
Total	228				χ^2 calculado = 28,940		

Fonte: Elaboração própria

A Figura 4.4 mostra a discrepância entre as frequências relativas observadas e esperadas. Embora haja uma aproximação da forma da curva, as frequências observadas estão acima das frequências esperadas (de acordo com a Lei de Benford) para o número de dormidas começando com os dígitos 1, 7, 8 e 9. Para o número de dormidas começando com os dígitos 2 e 3 as frequências observadas estão abaixo do esperado, mas para os números que começam com os dígitos 4, 5 e 6 as frequências observadas estão próximas dos valores esperados. Isso mostra o quão sensível é o teste Qui-Quadrado (χ^2), mas também como pode haver distúrbios esporádicos nos dados e até mesmo a necessidade de aumentar o período de recolha de dados.

4.2.3.1. Considerações Finais

Neste estudo rejeitou-se a hipótese de que o número de dormidas mensais em estabelecimentos turísticos da ilha do Sal obedecia à Lei de Benford. As discrepâncias entre as frequências relativas observadas e esperadas foram estatisticamente significativas, embora haja semelhança no formato da curva.

A Lei de Benford foi uma abordagem preliminar para verificar a fiabilidade dos números mensais de dados de dormidas registados nos estabelecimentos turísticos na Ilha do Sal.

O atraso na diversificação da oferta em termos de infraestruturas e eventos culturais e desportivos pode fazer com que o número de dormidas fique fora da Lei de Benford,

levando a que o primeiro dígito fique, na maioria das vezes, abaixo e acima das frequências espradas.

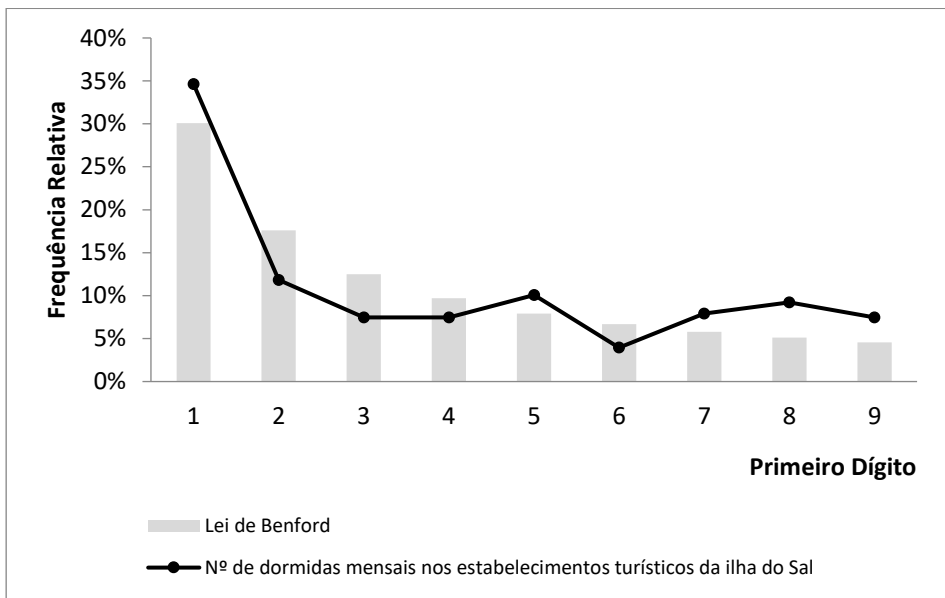


Figura 4.4: Comparação entre as frequências observadas e relativas do primeiro dígito do número de dormidas mensais nos estabelecimentos turísticos na ilha do Sal [2000:01-2018:12].

Fonte: Adaptado de Neves, Nunes e Fernandes (2021, p.595)

A principal conclusão deste estudo aponta que os resultados obtidos não só suportam a fiabilidade dos dados secundários (do INE), mas também sugerem que o número de referência de dormidas mensais em estabelecimentos turísticos pode ser simulado usando a Lei de Benford e os eventos podem ser considerados possíveis perturbações à lei.

Num estudo recente, levado a cabo por Matakovic (2021), a Lei de Benford foi aplicada ao número de dormidas nacionais e estrangeiras, numa série mensal ao longo de um período de três anos na Croácia. Nesse estudo, apenas as dormidas domésticas não seguiram a Lei de Benford. Porém, os autores não identificaram as causas da adaptação e não adaptação do número de dormidas à Lei, presumindo que isso ocorreu principalmente devido à falta de precisão e ausência de registros em acomodações

turísticas. Convém sublinhar que o trabalho de Matakovic (2021) é o primeiro estudo conhecido em que foi aplicado a Lei de Benford ao contexto turístico. O segundo estudo conhecido foi com os dados da ilha do Sal, no âmbito desta tese, e publicado em (Neves, Nunes & Fernandes, 2021).

4.2.4. Aplicação do Modelo SARIMA

Para esta aplicação também é feito o uso da série histórica do número mensal de dormidas em estabelecimentos de alojamento na ilha do Sal, entre janeiro de 2000 e dezembro de 2018 (Figura 4.2). Os pacotes, funções e comandos do *Software R* versão 4.0.5 são utilizados (Hyndman *et al.*, 2021; Trapletti *et al.*, 2021; Wickham *et al.*, 2021). Antes de aplicar qualquer modelo derivado do método de Box-Jenkins, os conhecidos modelos ARIMA e suas extensões, é realizado o teste de Cox-Stuart para verificar a tendência, o teste da raiz unitária de Dickey-Fuller para verificar a estacionariedade e o teste de Webel-Ollech para verificar a sazonalidade. Determinar se a série possui ou não o padrão sazonal é importante de modo a excluir à partida outros modelos ARIMA. Seguidamente, é feito um gráfico da série em estudo; tendência e sazonalidade removidas; identificou-se os parâmetros de ordem não sazonal p , d , q e de ordem sazonal P , D , Q ; modelos sazonais preliminares identificados e seus parâmetros estimados; o diagnóstico dos modelos foi realizado e o melhor modelo foi escolhido com base nos critérios de informação de Akaike (AIC) e erro de treino e teste MAPE (*Mean Absolute Percentage Error*).

4.2.4.1. Modelação da série

Para a construção do modelo SARIMA foi utilizado o *Software R* versão 4.0.5 (<https://www.r-project.org>), cujas funções utilizadas serão mencionadas ao longo da análise.

Passo 1: Análise e representação gráfica

```
>autoplot(dados) +  
geom_point(shape = 18) +
```

```
labs(x = "Anos - período 2000:2018", y = "Número de dormidas") +  
theme_bw()
```

A análise da Figura 4.5 apresenta uma tendência de crescimento bem como a presença de sazonalidade e ainda uma variação crescente ao longo do tempo. Estas características vão ser analisadas a seguir, começando-se pela análise da tendência.

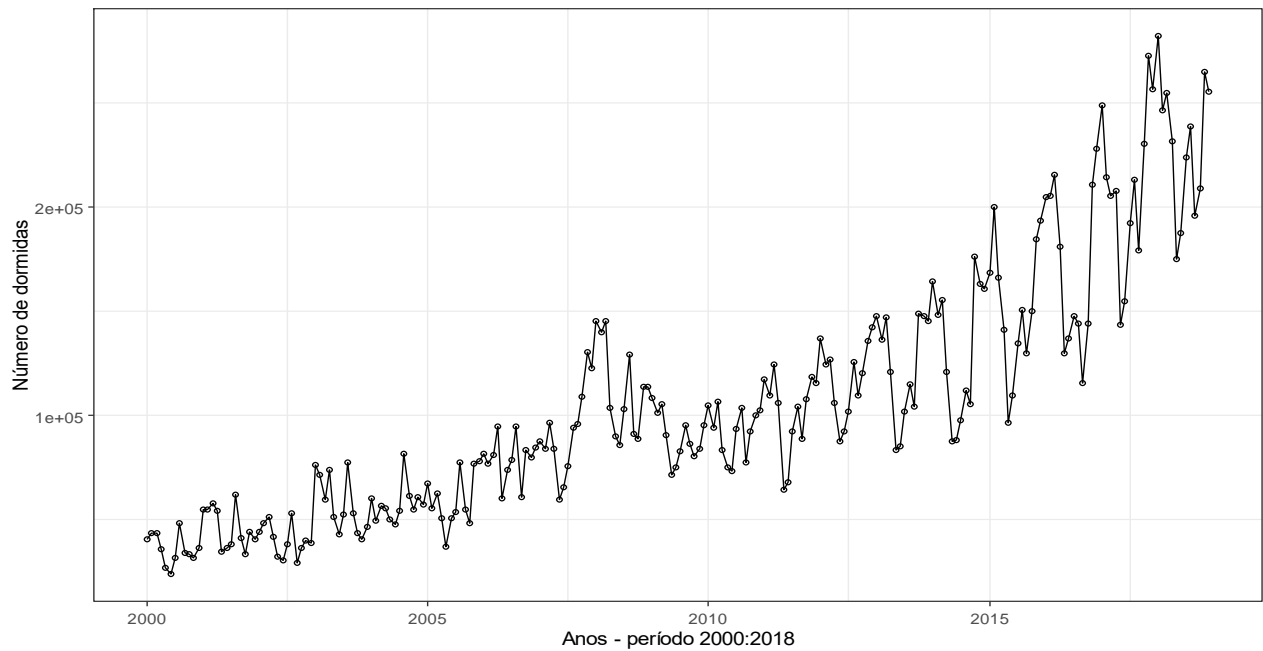


Figura 4.5: Número de dormidas mensais nos estabelecimentos turísticos na ilha do Sal, no período [2000:01-2018:12].

Fonte: Elaboração própria, com dados providos pelo INE

Tendência

Aplicando o teste de *Cox-Stuart* (`cox.stuart.test()`) confirma-se, pelo p -value $< 2,2e-16 < 0,05$, a existência da tendência percebida na série.

```
> cox.stuart.test(dados)  
Cox Stuart test  
data: dados  
statistic = 114, n = 114, p-value < 2.2e-16  
alternative hypothesis: non randomness
```

Tal tendência leva à especificação de um modelo não estacionário. Esta não estacionaridade é vista ainda no correlograma simples da série (Figura 4.6), uma vez que a ACF tende muito lentamente para zero. O correlograma é obtido com a função `acf2()`.

```
> acf2(dados)
```

Com a aplicação do teste da raiz unitária de Dickley-Fuller (`adf.test()`) chega-se à mesma conclusão quanto à não estacionaridade ($p\text{-value} = 0,05773 > 0,05$).

```
> adf.test(dados)
```

Augmented Dickey-Fuller Test

data: dados

Dickey-Fuller = -3.3867, Lag order = 6, p-value = 0.05773

alternative hypothesis: stationary

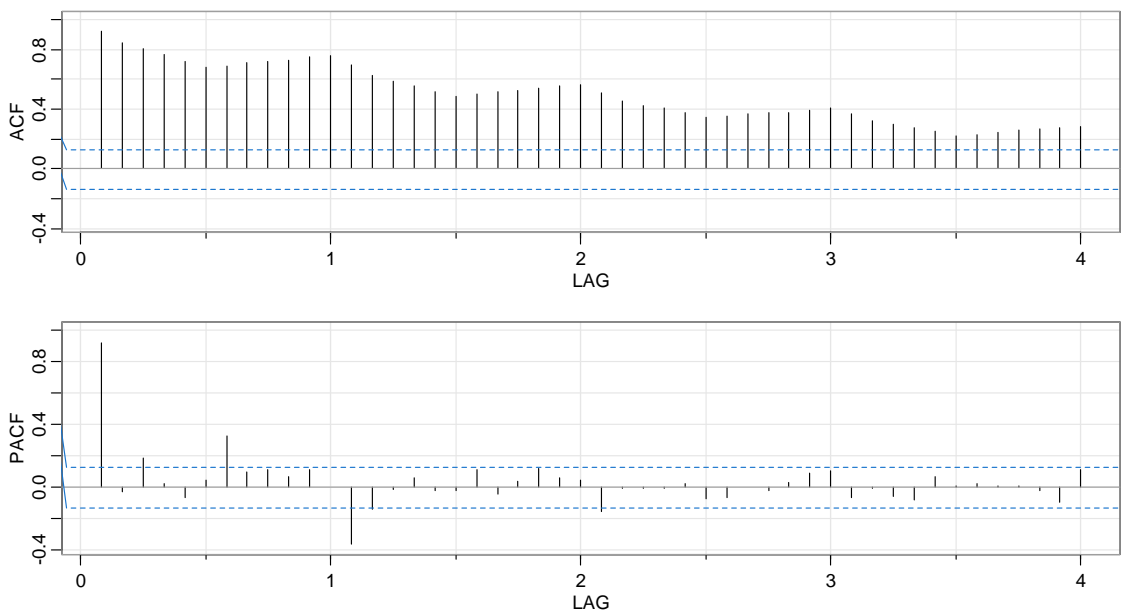


Figura 4.6: Correlograma simples da série.

Fonte: Elaboração própria

Estacionaridade

Em relação à variância é necessário estabilizá-la. Para estabilizá-la foi feita a transformação da série por meio do logaritmo $\log(X_t)$, como se pode constatar na Figura 4.7 e pela aplicação da função $\log()$.

```
> dadost<-log(dados)
> autoplot(dadost) +
  geom_point(shape = 18) + #shape é o estilo dos pontos no gráfico
  labs(x = "Anos - período 2000:2018", y = "Número de dormidas (log)") +
  theme_bw()
```

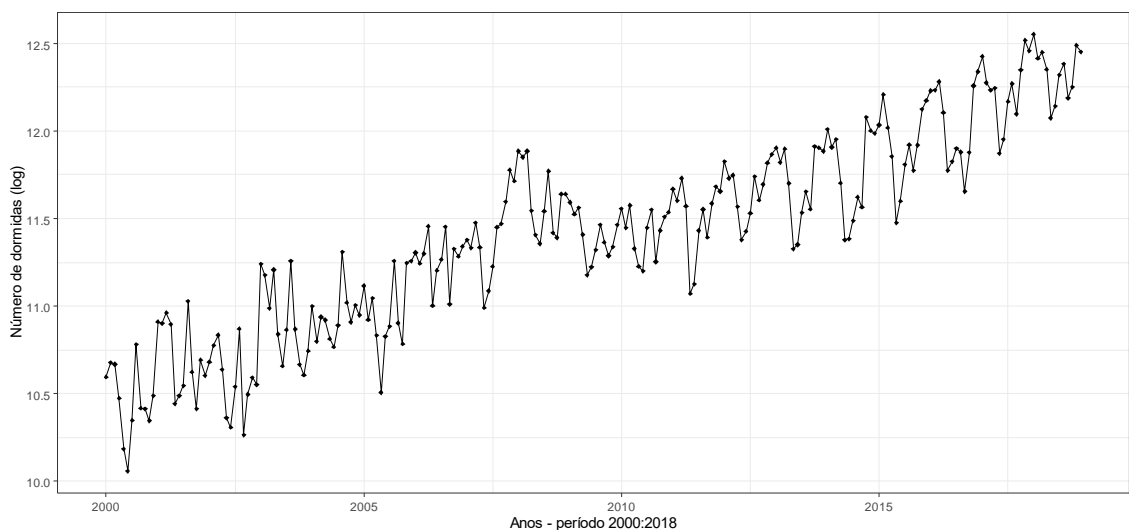


Figura 4.7: Transformação da série utilizando o logaritmo para estabilizar a variação.

Fonte: Elaboração própria

Mas, ainda se verifica uma tendência de crescimento da série, confirmado pelo teste de *Cox-Stuart*, `cox.stuart.test()`, ($p\text{-value}=2,2e-16 < 0,05$). Com relação a esta tendência e de forma a removê-la aplicam-se quantas diferenças forem necessárias. Mas, antes foi feita a divisão da série em duas amostras, uma para o conjunto de treino com 90% dos dados (2000:01 a 2017:12) e outra para o conjunto de teste com 10% dos dados (2018:01 a 2018:12).

```
> train <- window(dadost, end = c(2017, 12))
> test <- window(dadost, start = 2018)
```

Utilizando a função `ndiffs()` (que determina quantas diferenças não sazonais são necessárias), tem-se a indicação que deve ser feita apenas uma diferença simples do logaritmo da série, cujo resultado é visível na Figura 4.8.

```
> ndiffs(train)
[1] 1
```

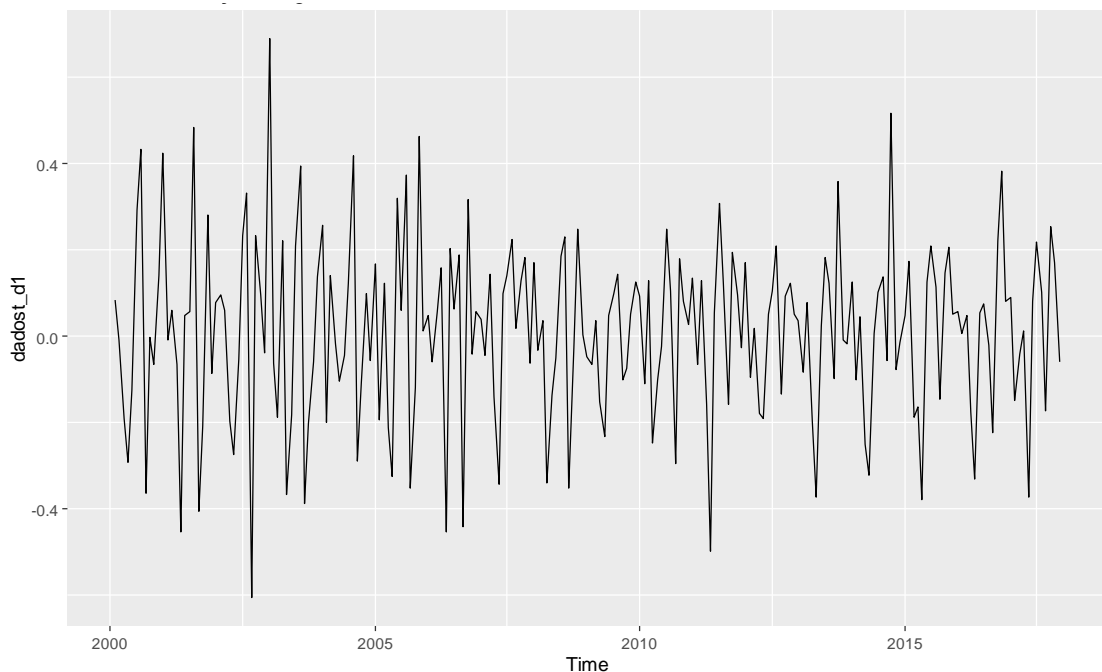


Figura 4.8: Resultado da primeira diferença não sazonal do logaritmo da série.

Fonte: Elaboração própria

A tendência de crescimento da série foi removida. Também foi utilizado o teste de Dickey-Fuller (`adf.test()`) cujo $p\text{-value} = 0,01 < 0,05$ confirmando que a série se tornou estacionária com essa primeira diferença não sazonal efetuada.

```
> adf.test(dadost_d1)
Augmented Dickey-Fuller Test
data: dadost_d1
Dickey-Fuller = -9.1951, Lag order = 5, p-value = 0.01
```

alternative hypothesis: stationary

Sazonalidade

Note-se que ainda parece haver sazonalidade na série. Isto torna-se mais evidente pela análise do correlograma da série diferenciada (Figura 4.9).

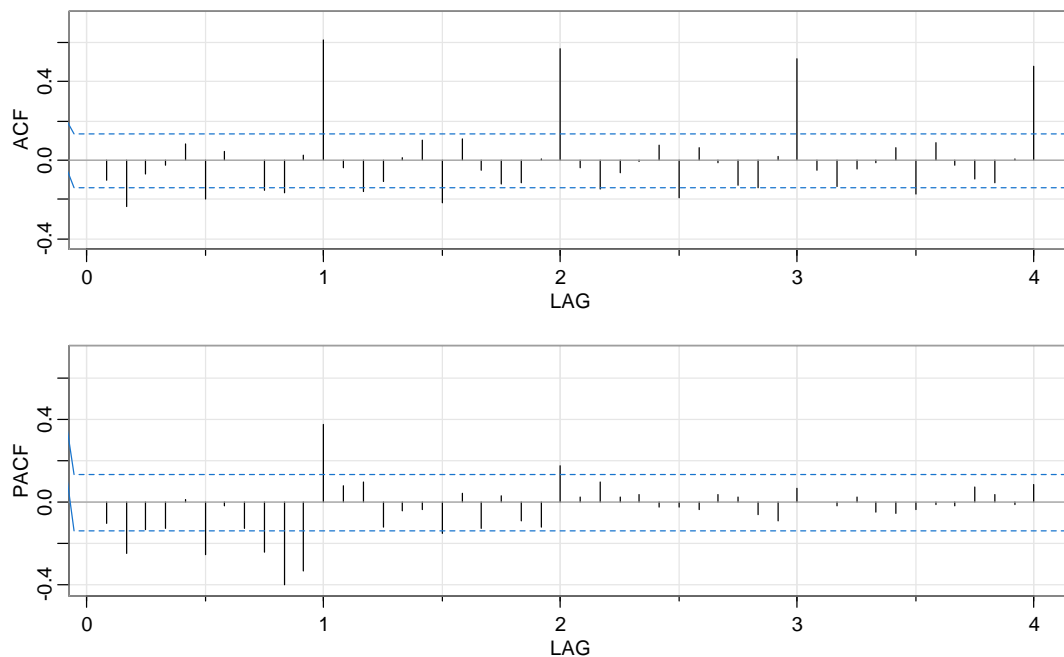


Figura 4.9: Correlograma (ACF and PACF) da série diferenciada.

Fonte: Elaboração própria

As relações de autocorrelação sazonal estão bem evidenciadas no gráfico ACF (Figura 4.9), em que uma forte correlação é vista nos *lags* 1, 2, 3 e 4 (no efeito sazonal). Os *lags* seguem a periodicidade da série, isto é, cada *lag* representa o equivalente a 12 meses.

Apesar de ser visível a presença da sazonalidade na série, e ainda pela análise do correlograma, convém que um teste seja aplicado. O teste de Webel-Ollech (`isSeasonal()`) cujo $p\text{-value} = 0,001 < 0,05$ confirma, portanto, a presença da sazonalidade.

Dada a conclusão sobre a existência da sazonalidade na série, procurou-se saber quantas diferenças sazonais é necessário aplicar. Com o uso da função `nsdiffs()` (que determina quantas diferenças sazonais são necessárias), chegou-se a uma diferença sazonal.

```
> nsdiffs(dadost_d1)
[1] 1
```

Com o comando seguinte é possível perceber que a questão da sazonalidade foi praticamente resolvida (Figura 4.10).

```
> dadost_sz <- diff(dadost_d1, 12)
> autoplot(dadost_sz) +
+ ggtitle("Primeira diferença sazonal do log da série")
```

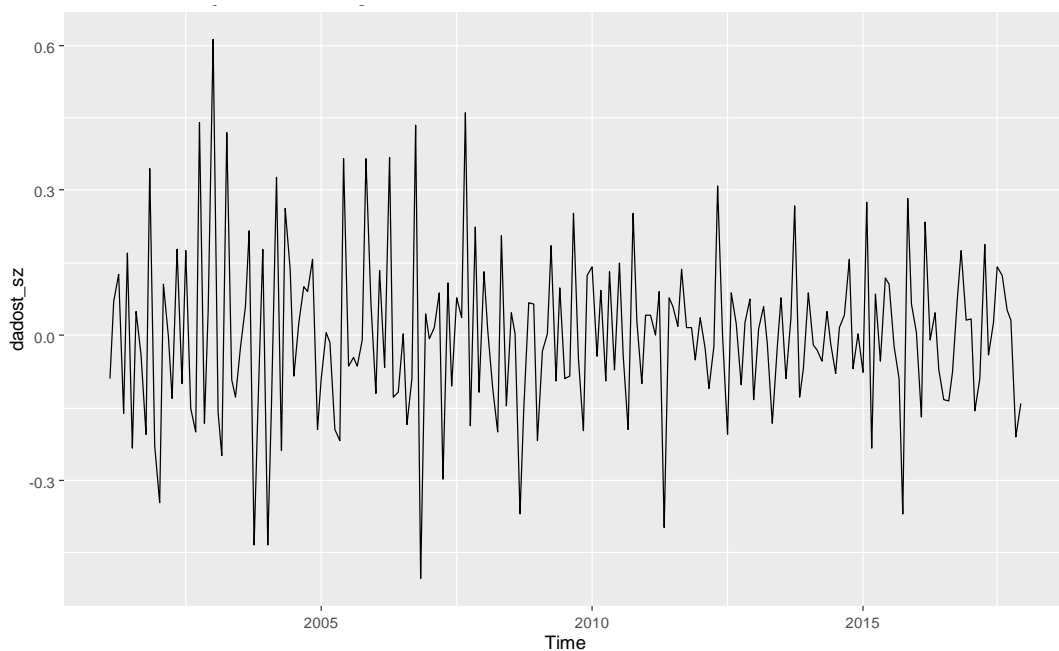


Figura 4.10: Resultado da primeira diferença sazonal da primeira diferença do logaritmo da série.

Fonte: Elaboração própria

Seguidamente retira-se uma diferença sazonal de 12 meses e constrói-se novamente o correlograma (Figura 4.11). Se confirma que muita pouca correlação permanece na série

após as duas diferenças tomadas anteriormente. Ainda, constata-se que provavelmente um modelo simples pode ser construído incorporando nele a autocorrelação *lag 1*.

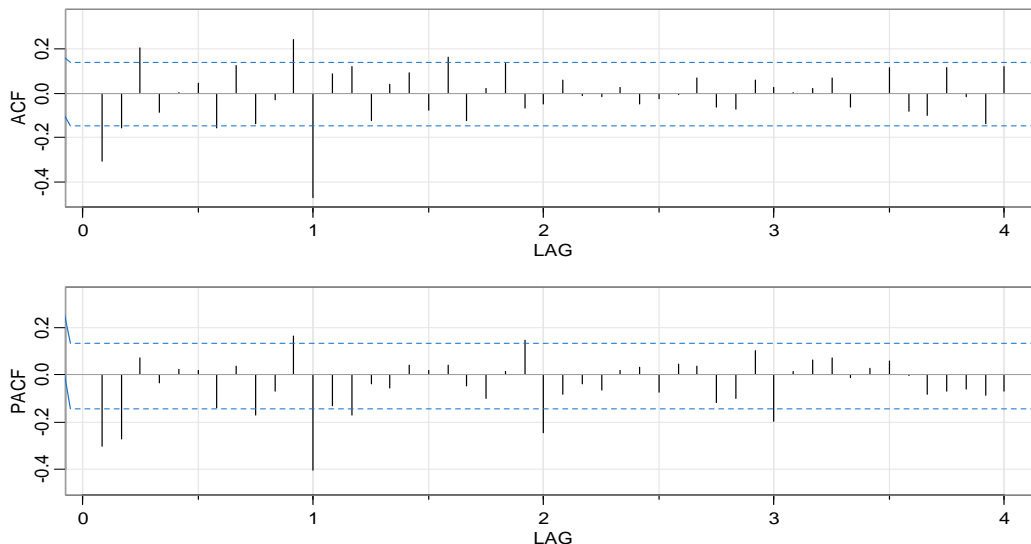


Figura 4.11: Correlograma (ACF and PACF) depois da primeira diferença sazonal do logaritmo da série.

Fonte: Elaboração própria

Passo 2 e 3: Construção do modelo provisório e identificação dos valores de ordem sazonal e não sazonal

O correlograma da Figura 4.11 mostra que o primeiro *lag* é significativo tanto na ACF como na PACF. O terceiro *lag* aparece significativo apenas na ACF. Ainda, o 12.º *lag* também é significativo em ambas as funções. No gráfico da PACF é evidente um decaimento exponencial nos *lags* 1, 2, 3 e 4 (no efeito sazonal), o que indicia o uso de um modelo de média móvel. Vê-se também que a ACF possui picos positivos (sinal da utilização da componente AR sazonal) e picos negativos (sinal da utilização da componente MA sazonal). Também, na PACF existe um *lag* de 2 (indicaria ordem 2 para a parte AR), e na ACF existe um *lag* de 1 (o 2 está no limite, indicaria ordem 1 para a parte MA – talvez ordem 2).

Dada esta análise, propõe-se o seguinte modelo sazonal inicial (**m1**):

SARIMA(2,1,1)(1,1,1)₁₂

constituído por um componente autoregressivo não sazonal de ordem dois, uma diferenciação, um componente de média móvel não sazonal de ordem um, um componente autoregressivo sazonal de ordem um, uma diferenciação sazonal e um componente de média móvel sazonal de ordem um.

Passo 4: Estimar os parâmetros do modelo provisório

O modelo inicial proposto (denominado de **m1**) contém cinco parâmetros, sendo o parâmetro de média móvel sazonal o com menor desvio-padrão.

```
>m1<-Arima(train,order=c(2,1,1),seasonal=list(order=c(1,1,1),period=12))
> m1
Series: train
ARIMA(2,1,1)(1,1,1)[12]

Coefficients:
      ar1   ar2   ma1   sar1   sma1
-0.3106 -0.2627 -0.0188 -0.0997 -0.6396
s.e. 0.3301 0.1081 0.3457 0.1002 0.0794

sigma^2 estimated as 0.01611: log likelihood=129.48
AIC=-246.96 AICc=-246.53 BIC=-227.08
```

Este modelo apresenta um AIC =-246,96, AICc=-246,53 e BIC=-227,08, cuja equação é expressa matematicamente, de acordo com a equação (25) do Capítulo III, da seguinte forma:

$$\begin{aligned} & (1 - \Phi_1 B^{12})(1 - \phi_1 B - \phi_2 B^2)(1 - B^{12})(1 - B)X_t \\ & = (1 + \Theta_1 B^{12})(1 + \theta_1 B)e_t \end{aligned} \quad (108)$$

Substituindo os parâmetros obtidos pelos seus valores respectivos, vem

$$\begin{aligned} & (1 + 0,0997B^{12})(1 + 0,3106B + 0,2627B^2)(1 - B^{12})(1 - B)X_t \\ & = (1 - 0,6396B^{12})(1 - 0,0188B)e_t \end{aligned} \quad (109)$$

Passo 5: Fazer o diagnóstico

Passo 5.1. Significância estatística dos parâmetros do modelo

Com a função `coefstest()` constata-se que alguns dos parâmetros do modelo são significativos, pois para um nível de significância de 5% o valor em módulo do T calculado é superior ao valor do T tabelado, $t_{(216-(2+1+1)),0.05/2} = 1,96$, o que leva à rejeição da hipótese nula que afirma que os parâmetros são nulos. A mesma conclusão pode ser obtida ainda pelo valor do $p\text{-value} < 0,05$.

```
> coefstest(m1)
t test of coefficients:

      Estimate Std. Error t value Pr(>|t|)
ar1 -0.310624  0.330094 -0.9410  0.34670
ar2 -0.262703  0.108148 -2.4291  0.01514 *
ma1 -0.018795  0.345734 -0.0544  0.95665
sar1 -0.099658  0.100162 -0.9950  0.31975
sma1 -0.639554  0.079390 -8.0558 7.894e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Passo 5.2. Estacionaridade do modelo

Utilizando o teste de Dickey-Fuller (`adf.test()`), concluiu-se que o modelo é estacionário, pois o $p\text{-value} = 0,01 < 0,05$, isto é rejeita-se a hipótese nula, podendo afirmar-se que a série é estacionária.

```
> adf.test(dadost_sz, alternative='stationary',k=12)

Augmented Dickey-Fuller Test

data: dadost_sz
Dickey-Fuller = -6.9973, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

Passo 5.3. Estabilidade do modelo

As correlações entre os parâmetros são, em valor absoluto, igual a 0,0008, inferior a 0,7 indicando, segundo Pires (2001), a inexistência de sobreparametrização do modelo e, portanto, não deverá haver necessidade, a princípio, de identificar novos modelos. Parece ser, de acordo com Lewis (1982) um modelo apropriado para descrever o comportamento futuro da série, pois possui um MAPE=0,77.

```
> summary(m1)
Series: train
ARIMA(2,1,1)(1,1,1)[12]

Coefficients:
      ar1   ar2   ma1   sar1   sma1
-0.3106 -0.2627 -0.0188 -0.0997 -0.6396
s.e.  0.3301  0.1081  0.3457  0.1002  0.0794

sigma^2 estimated as 0.01611: log likelihood=129.48
AIC=-246.96  AICc=-246.53  BIC=-227.08

Training set error measures:
              ME           RMSE          MAE           MPE           MAPE           MASE           ACF1
Training set -0.001141636  0.1215264  0.08734913  -0.01644501  0.7711757  0.4750835  -0.000778992
```

Passo 5.4. Redundância dos parâmetros do modelo

Ainda não se aplica, pois é o primeiro modelo em análise.

Passo 5.5. Avaliação da qualidade do modelo

Ruído branco

O ACF dos resíduos (Figura 4.12) do modelo mostra que todos os valores (à exceção de dois) estão dentro dos limites (linhas tracejadas), indicando que os resíduos se comportam como um ruído branco (resultado desejável).

```
> acf2(window(resid(m1),start=c(2000,1)), main="")
```

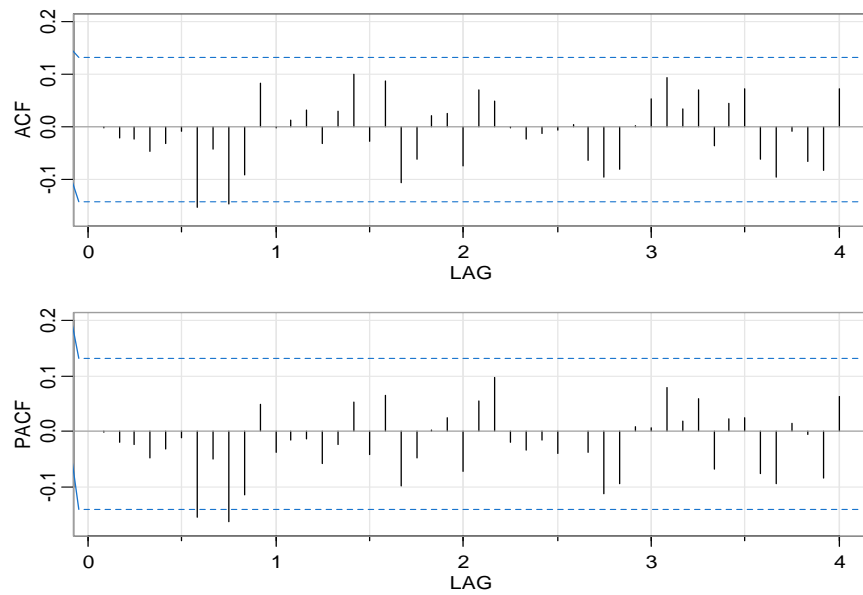


Figura 4.12: Correlograma (ACF and PACF) dos resíduos padronizados.

Fonte: Elaboração própria

Variância constante

Verifica-se, pela Figura 4.13, que os resíduos se distribuem aleatoriamente em torno do zero, o que é um indício de um “bom” ajuste do modelo.

```
>plot(window(resid(m1),start=c(2000,1)),ylab='Standardized Residuals',type='o')
>abline(h=0)
```

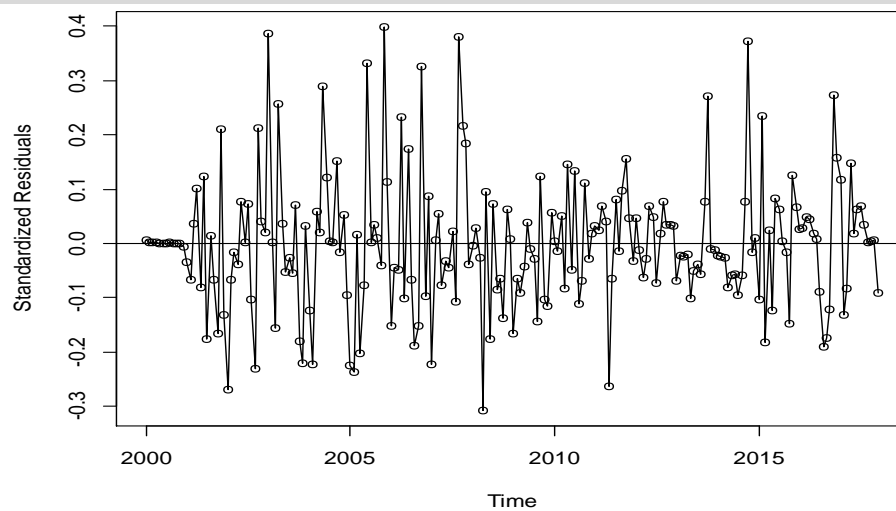


Figura 4.13: Distribuição dos resíduos padronizados.

Fonte: Elaboração própria

Distribuição normal

Para verificar a normalidade dos resíduos padronizados $\{\hat{\varepsilon}_t/S_{\hat{\varepsilon}}^2\}$ foram elaborados dois gráficos: o histograma e o Q-Q Plot (Figura 4.14), que revelaram indícios de não normalidade. Porém, ambos os gráficos ainda levavam a aceitar a normalidade, pois os resíduos não se afastam muito do modelo normal (isto é mais evidente no histograma). Mas no QQ-Plot, considerando o afastamento dos valores (observações) em relação à reta, verifica-se que apenas alguns valores situados à direita e no extremo superior do gráfico se encontram afastados, porém não muito.

```
> win.graph(width=3, height=3,pointsize=8)
> hist(window(resid(m1),start=c(2000,1)),xlab='Standardized Residuals')
> qqnorm(window(resid(m1),start=c(2000,1)))
> qqline(window(resid(m1),start=c(2000,1)))
```

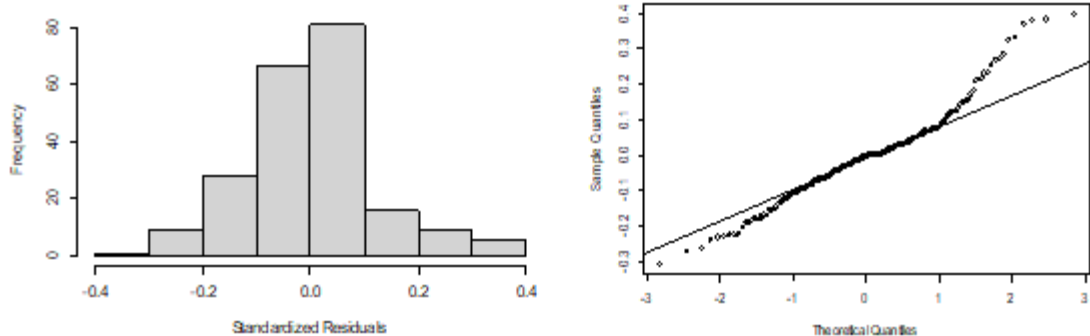


Figura 4.14: Análise da normalidade dos resíduos padronizados utilizando o histograma (à esquerda) e o Q-Q Plot (à direita).

Fonte: Elaboração própria

Para que não houvesse dúvidas quanto à normalidade, foi ainda realizado o teste de normalidade de Kolmogorov-Smirnov (`ks.test()`), obtendo assim um *p-value* < 0,05, confirmando a não normalidade.

```
> ks.test(resid(m1),"pnorm")
One-sample Kolmogorov-Smirnov test
data: resid(m1)
```

D = 0.39237, p-value < 2.2e-16
 alternative hypothesis: two-sided

Em síntese, o modelo **m1** segue todos os pressupostos exigidos no diagnóstico, com exceção da normalidade dos resíduos. Entretanto, dado que os resíduos padronizados não se afastam muito da normalidade, as análises do modelo podem ser continuadas, caso seja o modelo escolhido.

Para além do modelo **m1**, cuja proposta teve por base a análise do correlograma, outros modelos foram propostos e analisados, como apresentado na Tabela 4.5. Os modelos **m2**, **m3** e **m4** foram obtidos através da função `auto.arima()` cujos resultados também podem ser consultado no Anexo IV .

```
> auto.arima(train,d=1,D=1,stepwise=FALSE,approximation=FALSE,trace=TRUE)
```

Esta função apresentou praticamente todos os possíveis modelos (96 modelos no total) e escolheu automaticamente o melhor modelo para os dados do conjunto de treino tendo por base o AICc. Dentre todos os modelos, o **m2**, **m3** e **m4** (Tabela 4.5) são os que apresentam os melhores AICc, razão da necessidade de analisá-los com mais detalhes.

Tabela 4.5: Modelos escolhidos pelo melhor AICc da aplicação da função `auto.arima()`.

Modelo	Estrutura do Modelo	AICc
m1	<i>SARIMA</i> (2,1,1)(1,1,1) ₁₂	-246.53
m2	<i>SARIMA</i> (1,1,1)(0,1,1) ₁₂	-249.39
m3	<i>SARIMA</i> (2,1,0)(0,1,1) ₁₂	-249.78
m4	<i>SARIMA</i> (2,1,0)(0,1,2) ₁₂	-248.85

Fonte: Elaboração própria

À semelhança do modelo **m1**, todos os outros modelos (m2, m3 e m4) se comportaram de acordo com os pressupostos, pelo que a análise dos mesmos é resumida no Anexo VII, para comparação de cada um com os demais.

Passo 6: Escolher o modelo

As Tabelas 4.6 e 4.7 resumem os resultados dos quatro modelos.

Tabela 4.6: Comparação entre os critérios de informação para cada um dos modelos propostos.

Modelo	Estrutura do Modelo	AIC	AICc	BIC
m1	$SARIMA(2,1,1)(1,1,1)_{12}$	-246.96	-246.53	-227.08
m2	$SARIMA(1,1,1)(0,1,1)_{12}$	-249.59	-249.39	-236.34
m3	$SARIMA(2,1,0)(0,1,1)_{12}$	-249.98	-249.78	-236.73
m4	$SARIMA(2,1,0)(0,1,2)_{12}$	-249.15	-248.85	-232.59

Fonte: Elaboração própria

Tabela 4.7: Comparação entre os modelos propostos com base no seu desempenho nos dados do conjunto de treino e teste.

		Desempenho nos dados do conjunto de treino						
Modelo	Estrutura do Modelo	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
m1	$SARIMA(2,1,1)(1,1,1)_{12}$	0.001	0.122	0.087	0.005	0.773	-0.001	0.556
m2	$SARIMA(1,1,1)(0,1,1)_{12}$	0.003	0.121	0.090	0.020	0.794	-0.044	0.550
m3	$SARIMA(2,1,0)(0,1,1)_{12}$	0.001	0.122	0.087	0.006	0.774	0.001	0.556
m4	$SARIMA(2,1,0)(0,1,2)_{12}$	0.001	0.121	0.087	0.005	0.773	-0.002	0.556
		Desempenho nos dados do conjunto de teste						
Modelo	Estrutura do Modelo	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
m1	$SARIMA(2,1,1)(1,1,1)_{12}$	0.046	0.094	0.073	0.363	0.586	0.501	0.595
m2	$SARIMA(1,1,1)(0,1,1)_{12}$	0.005	0.075	0.065	0.037	0.526	0.518	0.487
m3	$SARIMA(2,1,0)(0,1,1)_{12}$	0.054	0.095	0.072	0.426	0.578	0.504	0.607
m4	$SARIMA(2,1,0)(0,1,2)_{12}$	0.043	0.094	0.074	0.343	0.595	0.501	0.594

Fonte: Elaboração própria

Pelo valor do AIC e do AICc, conclui-se que o melhor modelo a ajustar os dados da série e a prever dados futuros seria o modelo **m3**

$$\mathbf{m3: SARIMA(2,1,0)(0,1,1)_{12}}$$

Contudo, ao observar o desempenho com os dados do conjunto de treino, constata-se pelo valor do indicador de erro MAPE que os modelos **m1** e **m4** apresentam os menores erros.

$$\mathbf{m1: SARIMA(2,1,1)(1,1,1)_{12}}$$

$$\mathbf{m4: SARIMA(2,1,0)(0,1,2)_{12}}$$

Sendo que, entre os dois, o escolhido seja o modelo **m4**, pois possui o menor valor na maioria dos indicadores de erro. Todavia, tomando o MAPE como o nosso indicador de desempenho, a escolha recairia para o modelo **m1**, pois $0,7727594$ (modelo **m1**) < $0,7729454$ (modelo **m4**).

Mas ao observar a qualidade dos modelos na série de teste, constata-se que o modelo **m2** pode ser o escolhido pois apresenta o menor valor de erro MAPE, aproximadamente $0,526$.

Diante deste cenário envolvendo os indicadores de erro dos dois momentos de modelação da série (treino e teste), e pelo valor do AIC/AICc, caso se tivesse que escolher um modelo por via dos indicadores de erros estar-se-ia num dilema, pois o melhor modelo baseado no AIC/AICc não é o mesmo para a série de treino e nem tão pouco é o melhor para a série de teste.

Assim, optou-se por fazer uma escolha baseada no indicador de erro da série de teste, em específico o indicador MAPE. Verificou-se que os modelos **m2** e **m3** apresentaram os melhores valores para o MAPE, além do fato de todos os coeficientes de ambos os modelos serem significativos. Sendo assim, seguidamente vão ser analisados em detalhes ambos os modelos **m2** e **m3**, começando pelo diagnóstico.

Modelo m2: $SARIMA(1,1,1)(0,1,1)_{12}$.

Passo 5: O diagnóstico

Passo 5.1. Significância estatística dos parâmetros do modelo

Com a função `coefstest()` constata-se que todos os parâmetros do modelo são significativos, pois para um nível de significância de 5% (e mesmo para 1%) o valor em módulo do T calculado é superior ao valor de T tabelado, $t_{(216-(1+1+1)),0.05/2} = 1,96$, o que leva à rejeição da hipótese nula que afirma que os parâmetros são nulos. A mesma conclusão pode ser obtida ainda pelo valor do *p-value* < 0,05.

```
> coefstest(m2)
```

t test of coefficients:

```

Estimate Std. Error t value Pr(>|t|)
ar1  0.585892  0.107446  5.4529 4.956e-08 ***
ma1  -0.904472  0.069767 -12.9642 < 2.2e-16 ***
sma1 -0.723728  0.051110 -14.1603 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Passo 5.2. Estacionaridade do modelo

Sendo o modelo escolhido perante uma série estacionária, (teste de Dickey-Fuller com $p\text{-value} = 0,01 < 0,0$), conclui-se, portanto, que o modelo é também estacionário.

```

> adf.test(dadost_sz, alternative='stationary',k=12)
Augmented Dickey-Fuller Test
data: dadost_sz
Dickey-Fuller = -6.9973, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
    
```

Passo 5.3. Estabilidade do modelo

As correlações entre os parâmetros são, em valor absoluto, igual a 0,04430739, inferior a 0,7 indicando a inexistência de sobreparametrização do modelo e, portanto, não há necessidade de identificar novos modelos.

```

> summary(m2)
Series: train
ARIMA(1,1,1)(0,1,1)[12]
Coefficients:
    ar1    ma1    sma1
 0.5859 -0.9045 -0.7237
s.e. 0.1074 0.0698 0.0511

sigma^2 estimated as 0.01591: log likelihood=128.8
AIC=-249.59 AICc=-249.39 BIC=-236.34
    
```

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.002538441	0.121375	0.08974515	-0.03171314	0.7929086	0.4881153	-0.04430739

Passo 5.4. Redundância dos parâmetros do modelo

A redundância do parâmetro não se aplica, pois, apesar do modelo **m3** ser o mais próximo do modelo escolhido, mesmo assim ambos têm o mesmo número de parâmetros e todos significativos.

Passo 5.5. Avaliação da qualidade do modelo

Ruído branco

O ACF dos resíduos (Figura 4.15) do modelo mostra que todos os valores (à exceção de um) estão dentro dos limites (linhas tracejadas), indicando que os resíduos se comportam como um ruído branco (resultado desejável).

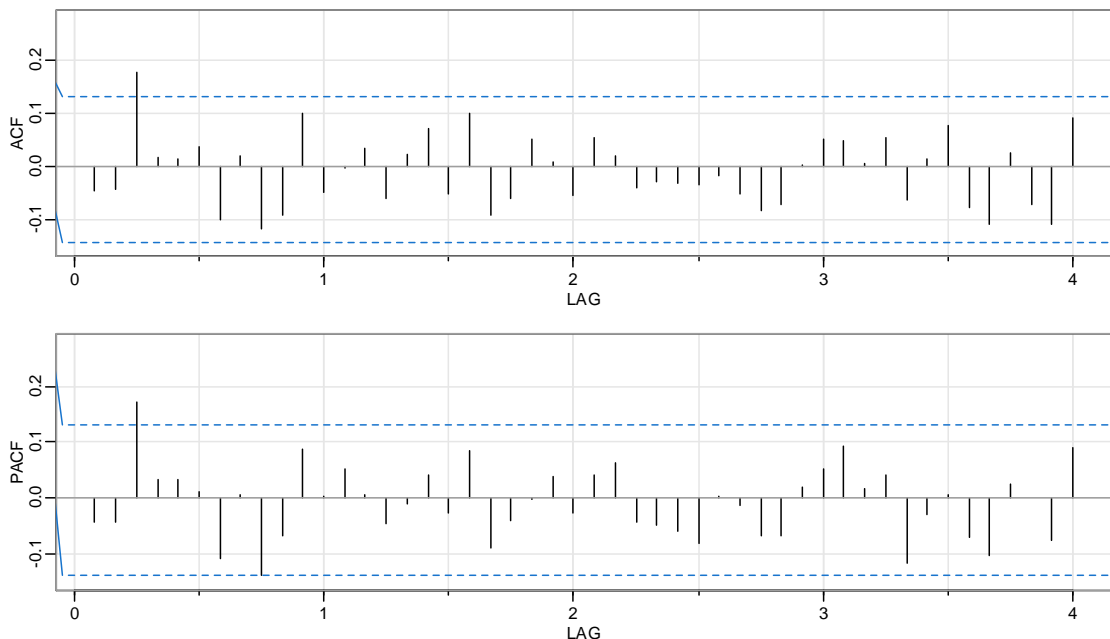


Figura 4.15: Correlograma (ACF and PACF) dos resíduos padronizados do modelo **m2** SARIMA(1,1,1) (0,1,1)[12].

Fonte: Elaboração própria

Variância constante

Verifica-se, pela Figura 4.16, que os resíduos padronizados se distribuem aleatoriamente em torno do zero, o que pode ser indício de um “bom” ajuste do modelo.

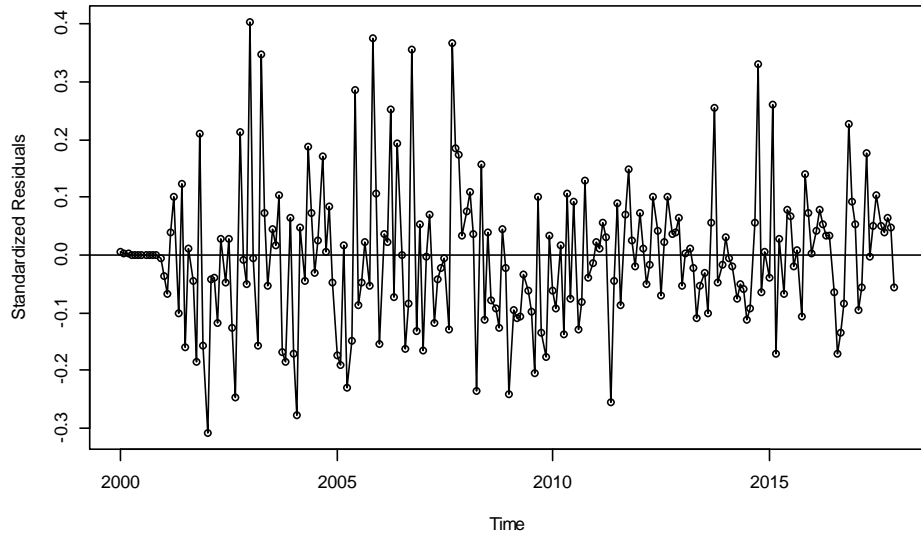


Figura 4.16: Distribuição dos resíduos padronizados do modelo m_2 SARIMA(1,1,1) (0,1,1)[12].

Fonte: Elaboração própria

Distribuição normal

Para verificar a normalidade dos resíduos padronizados $\{\hat{\varepsilon}_t/S_{\hat{\varepsilon}}^2\}$ foi elaborado dois gráficos: o histograma e o Q-Q Plot (Figura 4.17), que revelaram indícios de não normalidade. Porém ambos os gráficos ainda levariam a aceitar a normalidade, pois os resíduos não se afastam muito do modelo normal (isto é mais evidente no histograma). Mas no QQ-Plot, considerando o afastamento dos valores (observações) em relação à reta, verifica-se que apenas alguns valores situados à direita e no extremo superior do gráfico se encontram afastados e não muito afastados.

Para que não houvesse dúvidas quanto à normalidade, foi ainda realizado o teste de normalidade de Kolmogorov-Smirnov, obtendo assim um *p-value* < 0,05, confirmando a não normalidade.

Em síntese, o modelo **m2** segue todos os pressupostos exigidos no diagnóstico, com exceção da normalidade dos resíduos. Entretanto, dado que os resíduos padronizados não se afastam muito da normalidade, as análises do modelo continuam.

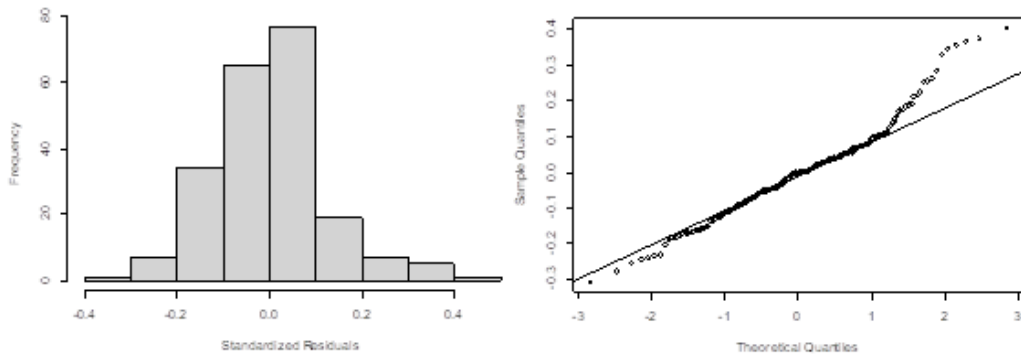


Figura 4.17: Análise da normalidade dos resíduos padronizados utilizando o histograma (à esquerda) e o Q-Q Plot (à direita) para o modelo $m2$ SARIMA(1,1,1) (0,1,1)[12].

Fonte: Elaboração própria

```
> ks.test(resid(m2), "pnorm")
One-sample Kolmogorov-Smirnov test
data: resid(m2)
D = 0.39, p-value < 2.2e-16
alternative hypothesis: two-sided
```

4.2.4.2. AJUSTAMENTO

O ajustamento do modelo será feito comparando-o com as observações do conjunto de treino, de Janeiro de 2000 a Dezembro de 2017 (Figura 4.18).

```
> plot(train, main=" ")
> lines(m2$fitted, col="red")
> legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
+       pch = c(NA, 1), col = c("red", "black"))
```

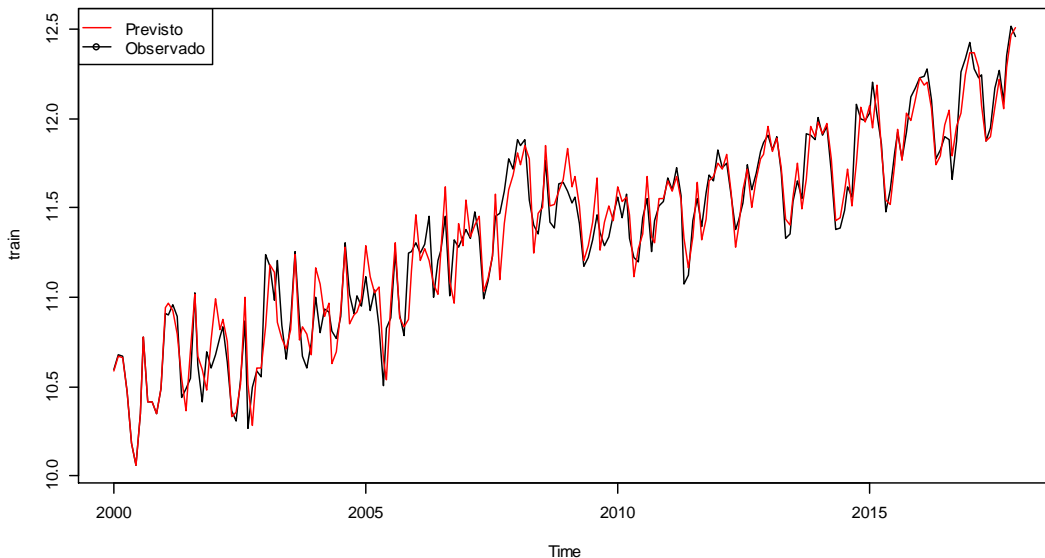


Figura 4.18: Ajustamento do modelo **m2** SARIMA(1,1,1) (0,1,1)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2017:12].

Fonte: Elaboração própria

Perderam-se 12 observações no início devido à diferenciação sazonal. Parece que os valores previstos se ajustaram bem aos valores observados captando bem a tendência. Para avaliar a qualidade desse ajustamento, vão ser obtidas algumas métricas utilizando a função `accuracy()`.

```
> accuracy(train,m2$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.002538441 0.121375 0.08974515 0.02004729 0.7940888 -0.04430739 0.5504231
```

Todos os indicadores de erros apresentam valores baixos, o que é um sinal de um bom ajustamento. O MAPE indica que o erro médio percentual cometido pelo modelo é de aproximadamente 0,79%. Uma outra medida que aparece é o U de Theil's cujo valor se encontra praticamente no centro do intervalo de 0 a 1. Neste caso o ajustamento é considerado bom. Mas, esta análise já tinha sido feita no âmbito da comparação com outros modelos antes propostos.

4.2.4.3. Previsão

Seguidamente vai realizar-se a previsão da série para o período considerado para teste, de Janeiro de 2018 a Dezembro de 2018, comparando com os valores previstos pelo modelo. São também apresentados os intervalos de confiança a 80% e 95% (zonas sombreadas) (Figura 4.19).

```
> plot(forecast(m2, h = length(test)),ylab="test",xlab="time", main="")
> lines(test, col = "black")
> legend("topleft", legend = c("Observado", "Previsto"),
+       lty = 1, col = c("black", "blue"))
```

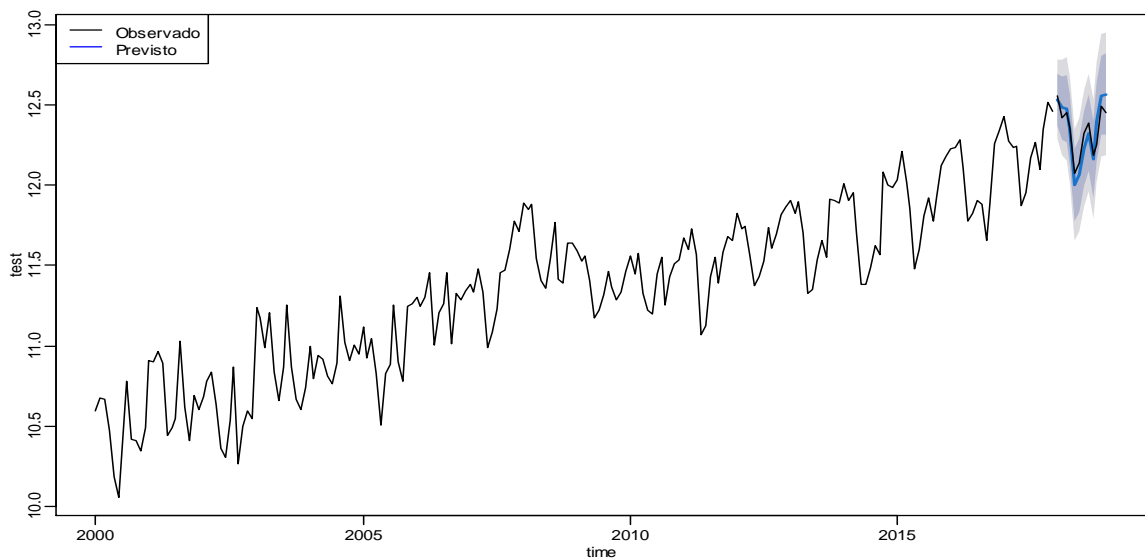


Figura 4.19: Ajustamento do modelo $m2$ SARIMA(1,1,1) (0,1,1)[12] comparando-o com os dados de teste, [2018:01, 2018:12], incluindo a visualização de todos os dados de treino.

Fonte: Elaboração própria

```
> accuracy(test,pred$pred)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.005332413	0.07549765	0.06486781	0.03653867	0.5256627	0.5180242	0.4873328

Verificando a qualidade de ajustamento, constata-se que o erro desta previsão nos dados para o conjunto de teste é de aproximadamente 0,53% (valor do MAPE), considerando assim uma boa previsão.

O valor do $ACF1 < 0,7$ elimina a hipótese de haver sobreparametrização do modelo. Algo a salientar aqui, é a visualização do intervalo de confiança. A parte sombreada mais escura representa um intervalo de confiança de 80% e a parte mais clara 95%.

A previsão da série a um passo à frente, ou seja, para os 12 meses do ano seguinte (2019), utilizando o modelo **m2**, é apresentada na Figura 4.20.

```
> m2<-Arima(dados,order=c(1,1,1),seasonal=list(order=c(0,1,1),period=12))
> plot(forecast(m2, h = 12),ylab="test",xlab="time", main="")
> lines(test, col = "black")
> legend("topleft", legend = c("Observado", "Previsto"),
+       lty = 1, col = c("black", "blue"))
```

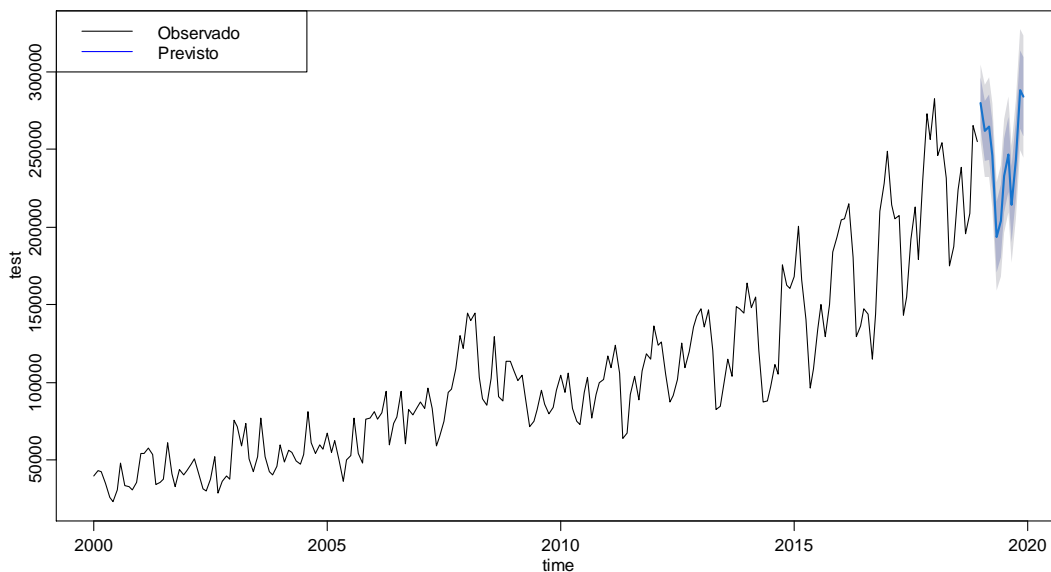


Figura 4.20: Valores previstos para os próximos 12 meses do ano de 2019 com o modelo **m2** SARIMA(1,1,1) (0,1,1)[12] e respetivos intervalos de confiança.

Fonte: Elaboração própria

Verificando a qualidade do modelo e as previsões numéricas:

```

> prev<-forecast(m2, h = 12)
> print(summary(prev))

Forecast method: ARIMA(1,1,1)(0,1,1)[12]

Model Information:
Series: dados
ARIMA(1,1,1)(0,1,1)[12]

Coefficients:
      ar1   ma1   sma1
      0.5734 -0.8846 -0.6074
s.e. 0.0976 0.0609 0.0586

sigma^2 estimated as 156276152: log likelihood=-2334.94
AIC=4677.89 AICc=4678.08 BIC=4691.37

Error measures:
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 516.5225 12054.44 8693.124 -0.9003931 8.842071 0.503759 -0.02097394

Forecasts:
      Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
Jan 2019  280255.9      264235.2  296276.7  255754.3  304757.5
Feb 2019  262150.6      242697.1  281604.0  232399.1  291902.1
Mar 2019  264536.0      243434.1  285637.8  232263.5  296808.4
Apr 2019  246118.3      224027.1  268209.4  212332.8  279903.7
May 2019  193793.0      171004.0  216582.1  158940.2  228645.9
Jun 2019  204042.5      180698.9  227386.0  168341.6  239743.4
Jul 2019  233153.2      209331.5  256974.8  196721.1  269585.2
Aug 2019  246934.5      222678.2  271190.8  209837.7  284031.4
Sep 2019  214689.4      190024.6  239354.1  176967.9  252410.8
Oct 2019  243675.9      218619.8  268732.0  205355.9  281995.9
Nov 2019  288538.2      263102.4  313973.9  249637.5  327438.8
Dec 2019  284120.6      258314.1  309927.2  244652.9  323588.3
    
```

Os mesmos procedimentos foram efetuados para o modelo **m3**

m3: $SARIMA(2,1,0)(0,1,1)_{12}$

A previsão da série a um passo à frente, ou seja para os 12 meses do ano seguinte (2019), utilizando o modelo **m3**, é apresentada na Figura 4.21.

```
> m3<-Arima(dados,order=c(2,1,0),seasonal=list(order=c(0,1,1),period=12))
> plot(forecast(m3, h = 12),ylab="test",xlab="time", main="")
> lines(test, col = "black")
> legend("topleft", legend = c("Observado", "Previsto"),
+       lty = 1, col = c("black", "blue"))
```

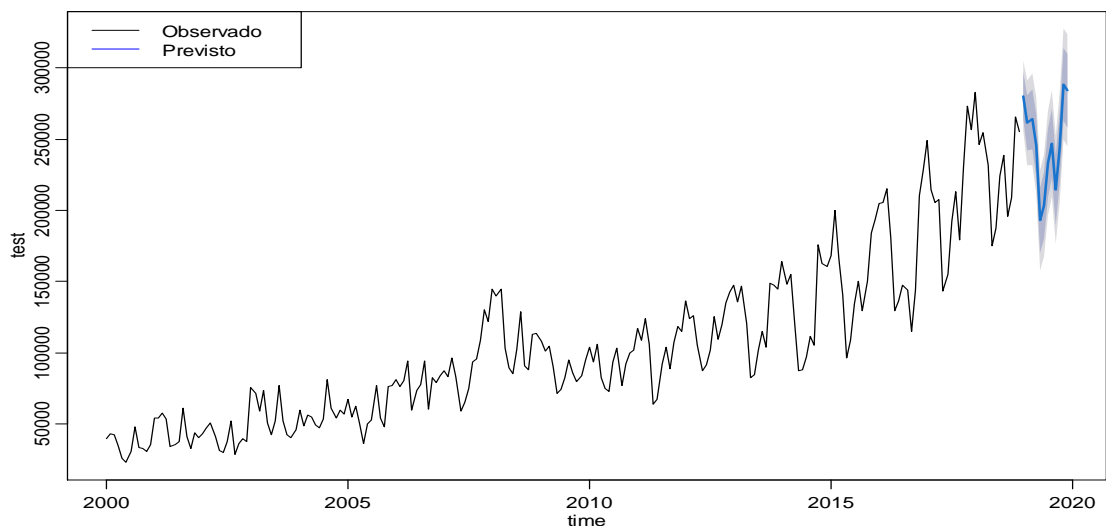


Figura 4.21: Valores previstos para os próximos 12 meses do ano de 2019 com o modelo **m3** $SARIMA(2,1,0)(0,1,1)[12]$ e respetivos intervalos de confiança.

Fonte: Elaboração própria

Verificando a qualidade do modelo e as previsões numéricas:

```
> prev<-forecast(m3, h = 12)
> print(summary(prev))
```

Forecast method: $ARIMA(2,1,0)(0,1,1)[12]$

Model Information:

Series: dados

ARIMA(2,1,0)(0,1,1)[12]

Coefficients:

ar1	ar2	sma1
-0.2672	-0.1976	-0.5948
s.e. 0.0675	0.0669	0.0593

sigma^2 estimated as 163559782: log likelihood=-2339.36

AIC=4686.72 AICc=4686.91 BIC=4700.2

Error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	84.99494	12332.15	8683.408	-0.8935421	8.784395	0.503196	-0.02670299

Forecasts:

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2019	273666.2	257276.4	290056.0	248600.1	298732.3
Feb 2019	253985.0	233665.6	274304.4	222909.1	285060.9
Mar 2019	254351.1	231729.6	276972.6	219754.5	288947.7
Apr 2019	234671.8	209358.0	259985.6	195957.7	273385.9
May 2019	181531.7	153773.4	209290.0	139079.1	223984.3
Jun 2019	191582.5	161686.4	221478.6	145860.4	237304.6
Jul 2019	220786.6	188870.9	252702.3	171975.7	269597.5
Aug 2019	234487.1	200659.6	268314.6	182752.4	286221.7
Sep 2019	201810.6	166181.4	237439.9	147320.4	256300.9
Oct 2019	230598.6	193254.7	267942.4	173486.0	287711.1
Nov 2019	276033.9	237049.3	315018.5	216412.1	335655.7
Dec 2019	271384.4	230825.6	311943.1	209355.2	333413.6

A Tabela 4.8 mostra a comparação entre o erro MAPE entre as previsões feitas por ambos os modelos (**m2** e **m3**) afim de ser escolhido aquele com o menor erro.

Tabela 4.8: Comparação entre os valores mensais observados com os valores previstos para os modelos **m2** e **m3**.

2019	Valores observados	Valores esperados com o modelo m2	Erro MAPE do modelo m2 (%)	Valores esperados com o modelo m3	Erro MAPE do modelo m3 (%)
Janeiro	267,940	280,256	4,60	273,666	2,14
Fevereiro	237,203	262,151	10,52	253,985	7,07
Março	233,508	264,536	13,29	254,351	8,93
Abril	207,956	246,118	18,35	234,672	12,85
Mai	152,691	193,793	26,92	181,532	18,89
Junho	192,645	204,043	5,92	191,583	0,55
Julho	217,145	233,153	7,37	220,787	1,68
Agosto	235,091	246,935	5,04	234,487	0,26
Setembro	193,380	214,689	11,02	201,811	4,36
Outubro	224,602	243,676	8,49	230,599	2,67
Novembro	257,670	288,538	11,98	276,033	7,13
Dezembro	236,604	284,121	20,08	271,384	14,70
Média do Erro MAPE (%)			11,96		6,77

Fonte: Elaboração própria

Ao prever os dados mensais de 2019, o modelo **m3** mostrou-se melhor do que o modelo **m2**, pois apresentou um menor valor MAPE=6,77%. De acordo com Lewis (1982), na avaliação da qualidade das previsões, um MAPE <10% indica alta precisão.

De acordo com a equação (25) do Capítulo III, o modelo **m3**, $SARIMA(2,1,0)(0,1,1)_{12}$, é expressa, matematicamente, da seguinte forma:

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B^{12})(1 - B)X_t = (1 + \theta_1 B^{12})e_t \quad (110)$$

Substituindo os parâmetros obtidos pelos seus respectivos valores, a equação do modelo é expressa assim

$$(1 + 0.2672B + 0.1976B^2)(1 - B^{12})(1 - B)X_t = (1 - 0.5948B^{12})e_t \quad (111)$$

4.2.4.4. Considerações Finais

Empiricamente, considerando a natureza do turismo e a visualização gráfica da série, se percebeu que um dos modelos do método Box-Jenkins que poderia ser aplicado seria o SARIMA, principalmente pela presença de sazonalidade no turismo na região da ilha do Sal. Os resultados confirmaram a presença de sazonalidade nos dados. Após realizar todas as análises na procura do melhor modelo SARIMA que pudesse descrever as séries de dormidas mensais, optou-se pelo modelo $SARIMA(2,1,0)(0,1,1)_{12}$ (modelo **m3**). Este

modelo apresentou o menor MAPE, além de atender a todas as premissas exigidas para sua identificação e diagnóstico.

Os resultados do modelo de previsão um passo à frente (dados de 2019) mostraram pouca diferença entre os valores previstos e o número de dormidas observadas em cada mês de 2019, com um erro médio de 6,77%, inferior a qualquer outro modelo SARIMA analisado nesta secção. Como a série de dados em estudo não é estacionária, o modelo SARIMA, com base nos resultados obtidos, demonstrou que pode lidar bem com este tipo de série, apresentando bom desempenho nas fases de treino e teste, bem como na previsão. Estes resultados foram publicados em Neves, Nunes e Fernandes (2021).

Enquanto a sazonalidade continuar a ser uma forte característica do turismo na Ilha do Sal, o modelo **m3**, $SARIMA(2,1,0)(0,1,1)_{12}$, poderá ser, dentro da metodologia de Box-Jenkins, a melhor escolha na previsão do número de futuras dormidas mensais em estabelecimentos turísticos. Note-se que o modelo **m3** foi o modelo escolhido pelo procedimento de escolha automática, a função `auto.arima()`, o que corrobora a análise do ACF e PACF.

Neste estudo, os valores de p, q, P, Q não ultrapassaram o valor 2, alinhando com a afirmação de Murteira (1993) que diz que raramente isso acontece.

4.2.5. Aplicação do Método de Holt-Winters (Suavização Exponencial Tripla)

Para a aplicação deste método à série de dormidas mensais em estudo, foi inicialmente utilizada a função `HoltWinters()` do R para verificar o ajustamento dos dados. Esta função solicita a introdução de dois parâmetros: a base de dados e o tipo de sazonalidade (aditiva ou multiplicativa). Neste caso obteve-se pelo tipo “multiplicativo”, pois é a característica da série em estudo.

4.2.5.1. Modelo m5: Holt Winters

Antes da aplicação, a série foi dividida em dois conjuntos, um de treino (2000:01 – 2017:12) com 90% dos dados e outro de teste (2018:01 – 2018:12) com 10% dos dados.

```
> train <- window(dadost, end = c(2017, 12))
> test <- window(dadost, start = 2018)
```

Seguidamente ajustam-se os dados do conjunto de treino (Figura 4.22). Infelizmente esta função não retorna o erro do ajustamento feito, e, portanto, foi calculado o erro utilizando o Microsoft Excel® utilizando os dados do conjunto de treino, as equações (62), (63), (64) e (65) do Capítulo III bem como os parâmetros e coeficientes gerados pelo *Software R*. Assim se verificou que os dados ajustados seguem o comportamento da série e com um nível de acurácia que corresponde a um MAPE = 12,5%.

```
> fit.mt<-HoltWinters(x=train,seasonal="multiplicative")
fit.mt
plot(fit.mt)
Holt-Winters exponential smoothing with trend and multiplicative seasonal component.

Call:
HoltWinters(x = train, seasonal = "multiplicative")

Smoothing parameters:
alpha: 0.4101928
beta : 0.008267106
gamma: 0.6297041

Coefficients:
  [,1]
a 12.36227311
b  0.01148224
s1  1.01505944
s2  1.00715027
s3  1.00361482
s4  0.99299366
s5  0.96053999
s6  0.96447835
s7  0.97709215
s8  0.98373575
s9  0.97185180
s10 0.99456943
s11 1.01144743
```

s12 1.00975825

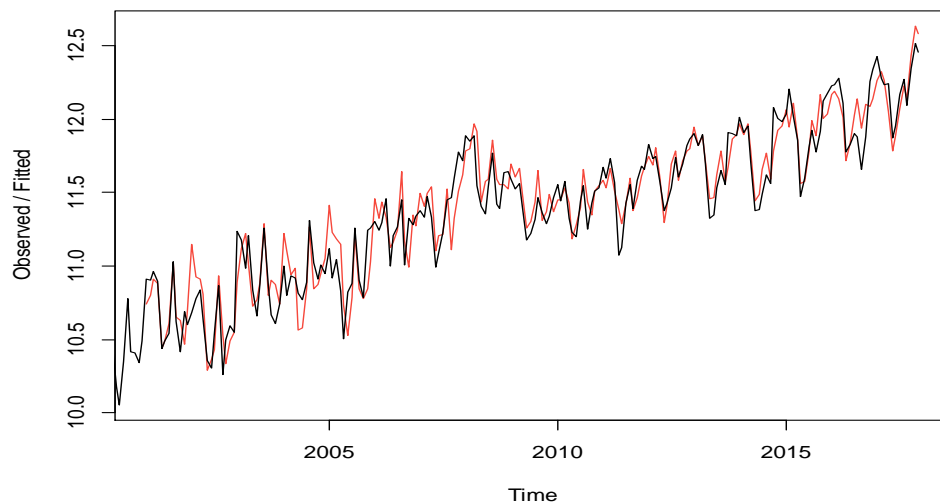


Figura 4.22: Ajustamento dos dados do conjunto de treino aos dados observados utilizando o modelo **m5** Holt-Winters.

Fonte: Elaboração própria

Tendo feito o ajustamento dos dados de treino, são analisados agora os resíduos, por meio das seguintes funções no *Software R versão 4.0.5*:

```
> resid<-fit.mt$residuals
> par(mfrow=c(2,2))
> plot(resid(fit.mt))
> acf(resid(fit.mt))
> pacf(resid(fit.mt))
> qqnorm(resid(fit.mt))
> qqline(resid(fit.mt))
```

Verifica-se, pela Figura 4.23, que os resíduos se distribuem aleatoriamente em torno do valor zero, o que é um indício de um “bom” ajuste do modelo. Quanto à normalidade dos resíduos padronizados $\{\hat{\varepsilon}_t/S_{\hat{\varepsilon}}^2\}$ a figura mostra o Q-Q Plot, que revela indícios de normalidade.

Entretanto a normalidade foi rejeitada pelo teste de normalidade de Kolmogorov-Smirnov, obtendo um *p-value* < 0,05.

```
> ks.test(resid(fit.mt),"pnorm")
```

One-sample Kolmogorov-Smirnov test

```
data: resid(fit.mt)
D = 0.36572, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Porém, olhando atentamente para o QQ-Plot e para o Histograma (Figura 4.24), ambos os gráficos ainda levam a aceitar a normalidade, pois os resíduos não se afastam muito do modelo normal (isto é mais evidente no histograma). Mas no QQ-Plot, considerando o afastamento dos valores (observações) em relação à reta, verifica-se que apenas alguns valores situados à direita e no extremo superior do gráfico se encontram afastados, porém não muito.

O ACF dos resíduos do modelo **m5** mostra (Figura 4.23) que alguns dos valores não estão dentro dos limites (linhas a tracejadas), indicando que os resíduos tendem a não se comportar como um ruído branco. O teste de Box-Pearce também foi utilizado, por meio da função `Box.test()`, e pelo valor do $p\text{-value} < 0,05$ chega-se à mesma conclusão encontrada na análise do correlograma.

```
> Box.test(x=resid(fit.mt), lag = 1, type = "Box-Pierce")
```

```
Box-Pierce test
```

```
data: resid(fit.mt)
```

```
X-squared = 17.014, df = 1, p-value = 3.711e-05
```

```
> plot(resid(fit.mt)) > win.graph(width=3, height=3,pointsize=8)
```

```
>hist(window(resid(fit.mt),start=c(2000,1)),xlab='Standardized Residuals')
```

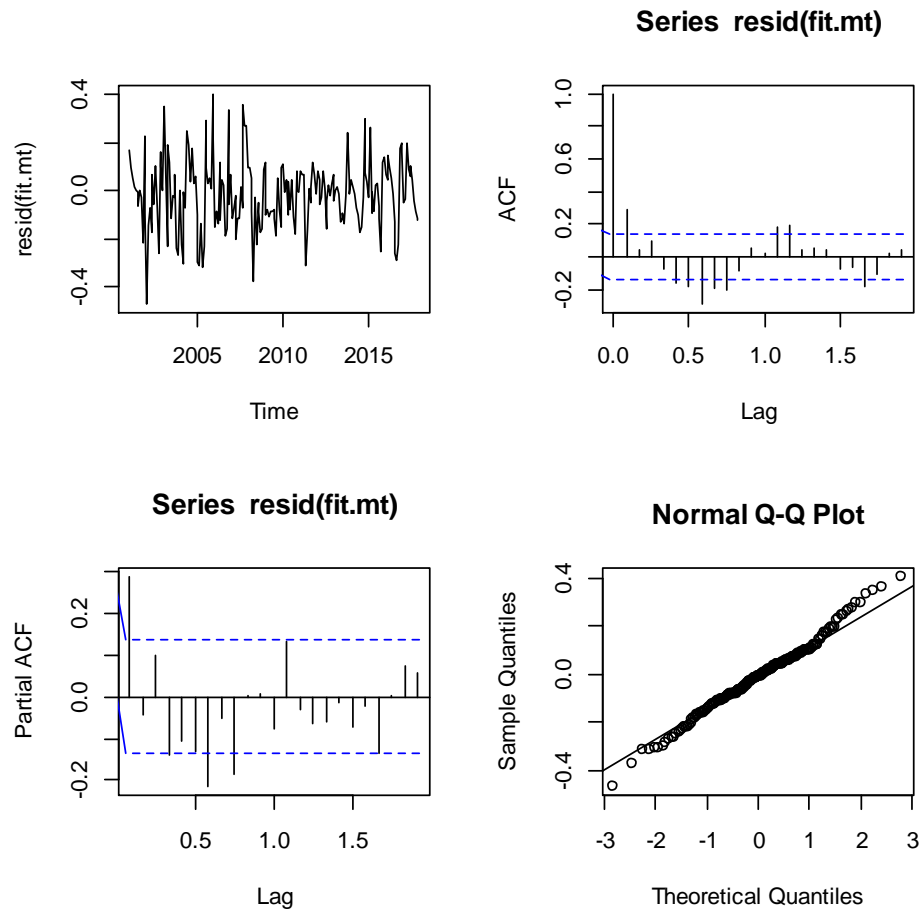


Figura 4.23: Análise dos resíduos obtidos utilizando o modelo *m5* Holt-Winters.

Fonte: Elaboração própria

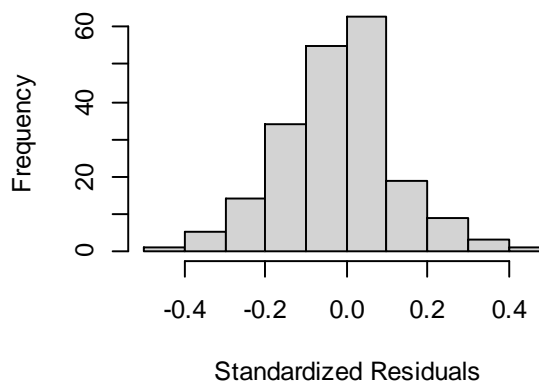


Figura 4.24: Histograma dos resíduos padronizados resultados do modelo *m5* Holt-Winters.

Fonte: Elaboração própria

Quanto à validação dos dados para o conjunto de teste:

```
> valid.t<-predict(fit.mt,n.ahead=12)
> plot.ts(dadost,ylim=c(10,13))
> lines(valid.t, col = 'red')
> legend("topleft", legend = c("Observado", "Teste"),
+       col = c("black", "red"), lty = 1)
```

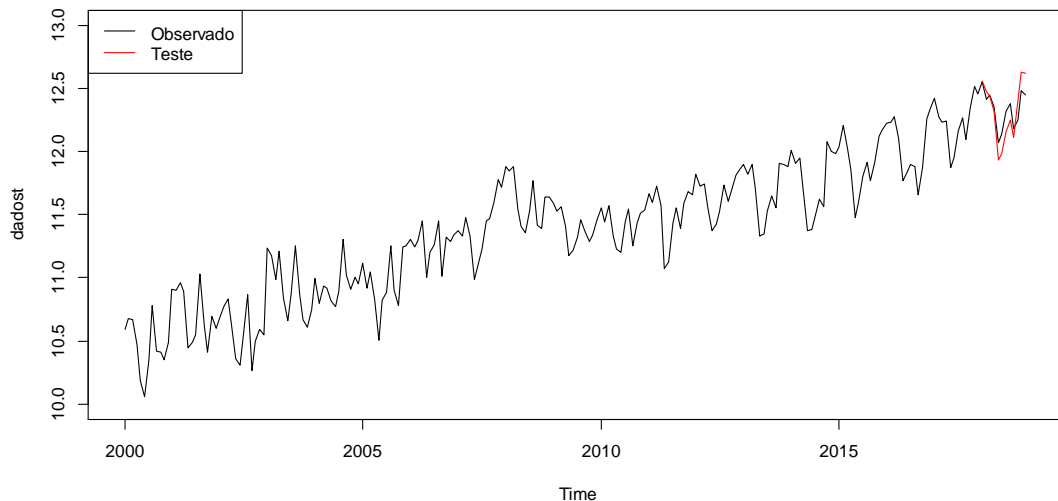


Figura 4.25: Validação dos dados para o conjunto de teste, no período de janeiro de 2018 a dezembro de 2018, para o modelo **m5** Holt-Winters.

Fonte: Elaboração própria

Para melhor visualizar o resultado da validação, destaca-se apenas o período de validação, ou seja, de janeiro de 2018 a dezembro de 2018.

```
valid.t<-predict(fit.mt,n.ahead=12)
plot.ts(test,ylim=c(11.5,13))
lines(valid.t, col = 'red')
legend("topleft", legend = c("Observado", "Teste"),
      col = c("black", "red"), lty = 1)
```

Os dados ajustados do conjunto de teste tende a acompanhar o comportamento da série e o nível de acurácia foi calculado com base no MAPE = 9,6%.

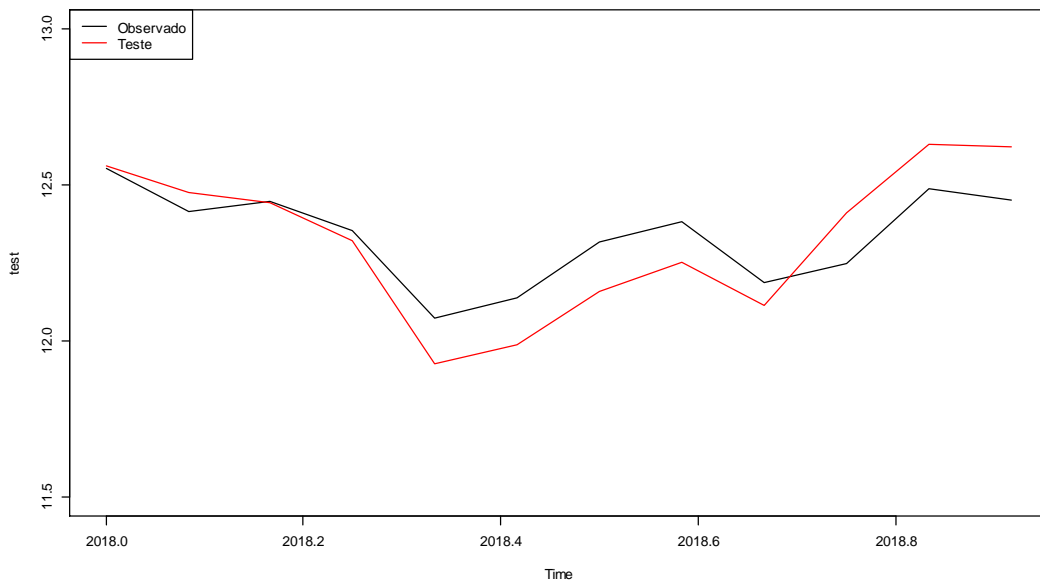


Figura 4.26: Validação dos dados para o conjunto de teste, no período de janeiro de 2018 a dezembro de 2018, utilizando o modelo *m5* Holt-Winters.

Fonte: Elaboração própria

Utilizando o pacote `forecast` do R vai prever-se então o número de dormidas mensais a um passo à frente, ou seja, para o ano de 2019; com os seguintes comandos do R:

```
> fit.m<-hw(dados,h=12,seasonal="multiplicative",initial="o")
> summary(fit.m)

Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
hw(y = dados, h = 12, seasonal = "multiplicative", initial = "o")

Smoothing parameters:
alpha = 0.5804
beta = 0.002
gamma = 1e-04
```

Como se viu no Capítulo 3, nas equações (62), (63) e (64), o *alfa* é o parâmetro do nível da série, o *beta* do nível da tendência e o *gamma* do nível da sazonalidade. Pelo resultado, fica claro que o *beta* é praticamente zero o que significa que a taxa de crescimento da

série é praticamente constante; o *gamma* também é praticamente zero, o que indica que todas as observações são atribuídas pesos e todas as sazonalidades são importantes para o modelo. A seguir apresenta-se a inicialização de todos os parâmetros do modelo, bem como a previsão a um passo à frente.

```
Initial states:
l = 33681.9791
b = 747.4704
s = 1.1123 1.1291 0.9532 0.8134 1.0096 0.9006
    0.7778 0.7371 1.0149 1.1696 1.1543 1.2282

sigma: 0.1399

AIC AICc BIC
5563.733 5566.647 5622.032

Error measures:
ME RMSE MAE MPE MAPE MASE ACF1
Training set 218.7933 11490.95 8281.075 -1.293296 9.419252 0.4798811 0.1264105

Forecasts:
Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
Jan 2019 285179.2 234044.6 336313.8 206975.59 363382.8
Feb 2019 268996.8 213094.7 324898.8 183501.98 354491.6
Mar 2019 273596.4 209847.5 337345.3 176100.91 371091.9
Apr 2019 238289.0 177322.4 299255.6 145048.68 331529.3
May 2019 173694.0 125590.9 221797.1 100126.67 247261.4
Jun 2019 183935.8 129371.1 238500.5 100486.37 267385.2
Jul 2019 213767.2 146379.5 281154.9 110706.68 316827.7
Aug 2019 240563.3 160481.1 320645.5 118088.15 363038.5
Sep 2019 194493.1 126467.0 262519.1 90456.22 298530.0
Oct 2019 228705.9 145011.5 312400.2 100706.39 356705.4
Nov 2019 271878.5 168145.3 375611.6 113232.26 430524.7
Dec 2019 268828.8 162206.9 375450.6 105764.73 431892.8
```

A previsão a um passo à frente também é representada na Figura 4.27.

Comparando os resultados previstos pelo modelo com os dados observados do número de dormidas mensais em 2019, verificamos pelo erro MAPE = 7,94%, que o modelo teve um alto desempenho (Tabela 4.9 e Figura 4.28).

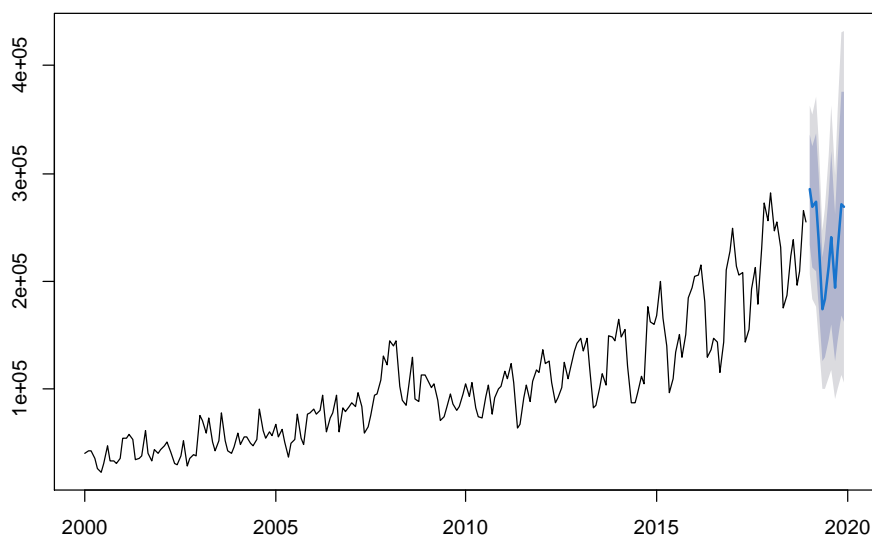


Figura 4.27: : Previsão a um passo à frente utilizando o modelo **m5** Holt-Winters, incluindo os intervalos de confiança a 80% e 95%.

Fonte: Elaboração própria

Tabela 4.9: Comparação entre os resultados previstos pelo modelo **m5** Holt-Winters com os dados observados do número de dormidas mensais em 2019.

Meses	Valores observados	Valores esperados	Erro MAPE (%)
Janeiro	267940	285179	6,43
Fevereiro	237203	268997	13,40
Março	233508	273596	17,17
Abril	207956	238289	14,59
Mai	152691	173694	13,76
Junho	192645	183936	4,52
Julho	217145	213767	1,56
Agosto	235091	240563	2,33
Setembro	193380	194493	0,58
Outubro	224602	228706	1,83
Novembro	257670	271879	5,51
Dezembro	236604	268829	13,62
Média do Erro MAPE (%)			7,94

Fonte: Elaboração própria

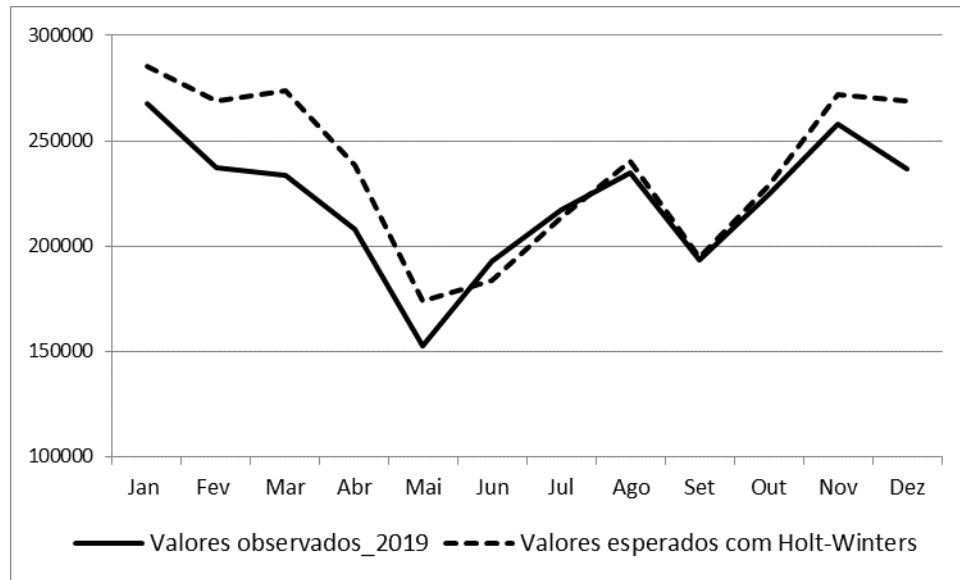


Figura 4.28: Comparação entre os resultados previstos pelo modelo **m5** Holt-Winters com os dados observados do número de dormidas mensais em 2019.

Fonte: Elaboração própria

As equações que compõe o modelo **m5** são dadas por:

$$L_t = 0,5804 \frac{X_t}{S_{t-s}} + (1 - 0,5804)(L_{t-1} + T_{t-1}) \quad (112)$$

$$T_t = 0,002(L_t - L_{t-1}) + (1 - 0,002)T_{t-1} \quad (113)$$

$$S_t = 0,0001 \frac{X_t}{L_t} + (1 - 0,0001)S_{t-s}, \quad (114)$$

Sendo os valores iniciais de $L_{t-1} = 33681,9791$ e $T_{t-1} = 747,4704$ e os coeficientes iniciais (S_{t-s}) dados por: $S_1 = 1,1123$; $S_1 = 1,1291$; $S_1 = 0,9535$; $S_1 = 0,8134$; $S_1 = 1,0096$; $S_1 = 0,9006$; $S_1 = 0,7778$; $S_1 = 0,7371$; $S_1 = 1,0149$; $S_1 = 1,1696$; $S_1 = 1,1543$; $S_1 = 1,2282$

Essas equações (112, 113 e 114), valores e coeficientes iniciais permitem prever os valores futuros por meio da equação (115), dado por.

$$\hat{X}_{t+h} = (L_t + hT_t)S_{t-12+h}, \quad h = 1,2, \dots s \quad (115)$$

4.2.5.2. Considerações Finais

Dos três principais métodos de suavização exponencial (simples, dupla e tripla), apenas foi necessário trabalhar com a suavização exponencial tripla de sazonalidade, dada a confirmação, na seção da análise da série, da existência da componente sazonalidade e de seu efeito multiplicativo. O modelo **m5** Holt-Winters conseguiu prever o número de dormidas mensais para o ano de 2019 com alta performance, gerando um erro MAPE de 7,94%, abaixo dos 10%, pelo que a sua aplicação na série temporal em estudo se mostrou apropriado e a sua inclusão neste trabalho se mostra muito útil para efeito de comparação com os outros modelos.

4.2.6. Aplicação da Rede Neuronal Artificial

Para esta aplicação também é feito o uso da série histórica do número mensal de dormidas em estabelecimentos de alojamento na ilha do Sal, entre janeiro de 2000 e dezembro de 2018 (Figura 4.2). Os pacotes do *Software R* versão 4.0.5 descritos na secção 3.4.15 e as funções `nnetar`, do pacote `forecast`, e `mlp()` do pacote `nnfor`, são utilizados.

Utilizando a função `mlp()`, descrito na secção 3.7.8, é criado diferentes modelos de rede neuronal *perceptron* de uma camada e de estrutura do tipo *feedforward* com o objetivo de compara-los e escolher o que apresenta menor erro MAPE. O comando da função `mlp()` é sempre a mesma para todos os modelos, alterando apenas os parâmetros **hd** e **lags** em cada modelo. Assim, para identificar cada modelo é utilizado uma notação semelhante à descrita na secção 3.7.8 do tipo RNAR (p,P,k) [m], onde **p** é o número de *lags* de entrada, **P** é a ordem de sazonalidade (para a série em estudo o **P=1**, como verificado na aplicação do modelo SARIMA), **k** é o número de neurónios na camada oculta dado por $\frac{p+P+1}{2}$ (arredondado para o inteiro mais próximo) e **m** o período sazonal.

4.2.6.1. Modelação com a função `mlp()`

Inicialmente os dados foram divididos em conjuntos de treino (2000:2015) com 80% dos dados e teste (2016:2018) com 20% dos dados.

```
> treino<- window(dadost, end = c(2015, 12))
> teste <- window(dadost, start = c(2016,1),end=c(2018,12))
```

Os resultados para cada modelo são apresentados da seguinte forma: os comandos da função `mlp()` para o ajustamento do modelo comparando-o com os dados do conjunto de treino [2000:01, 2015:12], o gráfico correspondente e os valores de acurácia; os comandos para a validação do modelo comparando-o com os dados do conjunto de teste [2016:01, 2018:12], o gráfico correspondente e os valores de acurácia. Os resultados de cada modelo se encontram no Anexo VIII.

A Tabela 4.10 apresenta um resumo dos erros de treino e teste de cada modelo RN pesquisado. A escolha recai sobre o modelo que apresentar o menor erro MAPE de teste. Neste caso é escolhido o modelo **m11** que possui 12 lags de entrada e 7 nós na camada oculta.

Tabela 4.10: Comparação dos erros de treino e teste de cada modelo de Redes Neurais (RN), com uso da função `mlp()`.

Modelo	Estrutura do Modelo RN	MAPE % (Treino)	MAPE % (Teste)
m6	RNAR(2,1,2)[12]	2,643	1,112
m7	RNAR(4,1,3)[12]	2,664	1,109
m8	RNAR(6,1,4)[12]	2,664	1,104
m9	RNAR(8,1,5)[12]	2,645	1,100
m10	RNAR(10,1,6)[12]	2,626	1,085
m11	RNAR(12,1,7)[12]	2,564	0,918
m12	RNAR(14,1,8)[12]	2,653	1,079
m13	RNAR(16,1,9)[12]	2,625	1,0752
m14	RNAR(18,1,10)[12]	2,624	1,0753
m15	RNAR(20,1,11)[12]	2,573	1,067
m16	RNAR(22,1,12)[12]	2,589	1,051
m17	RNAR(24,1,13)[12]	2,543	1,028
m18	RNAR(26,1,14)[12]	2,542	1,026
m19	RNAR(28,1,15)[12]	2,572	0,989
m20	RNAR(30,1,16)[12]	2,572	1,023
m21	RNAR(32,1,17)[12]	2,591	1,004
m22	RNAR(34,1,18)[12]	2,570	0,954

m23	RNAR(36,1,19)[12]	2,587	0,967
------------	-------------------	-------	-------

Fonte: Elaboração própria

Sendo assim, de seguida, vai apresentar-se os resultados mais detalhados do modelo escolhido (**m11**):

Modelo m11: RNAR(12,1,7)[12]:

Este modelo de rede neuronal de 12 lags de entrada e 7 neurónios na camada oculta, resultante da aplicação do comando da função *m1p()* dada anteriormente, permite a elaboração automática da Figura 4.29 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

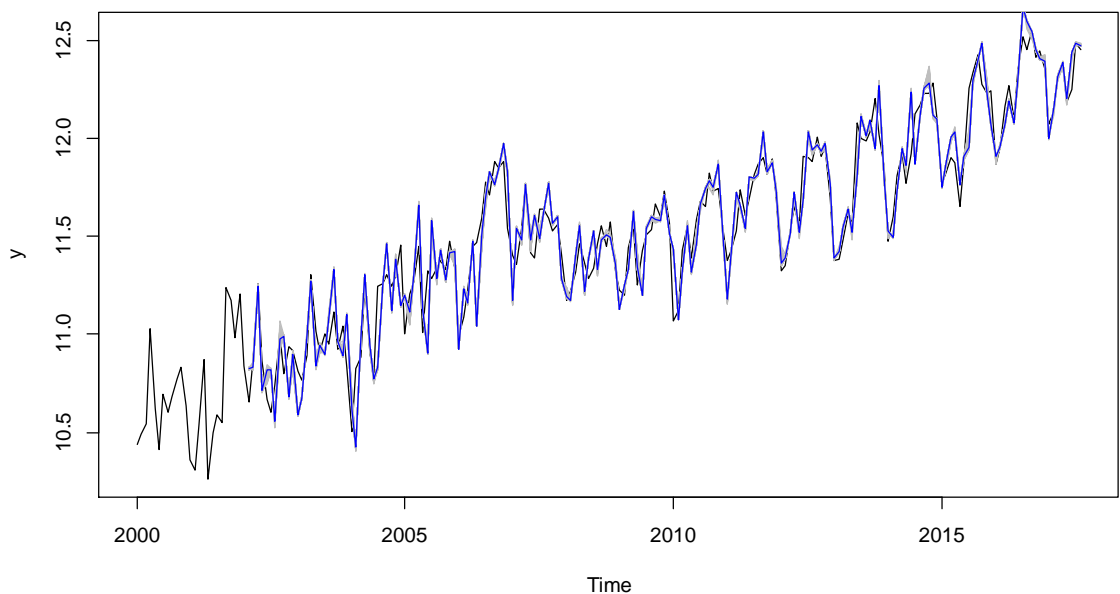


Figura 4.29: Ajustamento do modelo **m11** RNAR(12,1,7)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m11** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1380919	0.3617152	0.2979164	1.147435	2.563723	0.5884511	1.50372

Seguidamente é feita a validação do modelo **m11** comparando-o com os dados do conjunto de teste com apoio da Figura 4.30 e os valores da acurácia:



Figura 4.30: Validação do modelo **m11** RNAR(12,1,7)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -0.002874792  0.1428481  0.1042434 -0.04038602  0.9184138 -0.2544051  0.5818922
```

Com relação a previsão mensal a um passo à frente (ano de 2019) a Figura 4.31 mostra que a previsão segue o comportamento da série. Na Tabela 4.11 e na Figura 4.32 a comparação dos valores previstos com os valores observados.

```
> prev2019<-forecast(rede.teste,h=12,PI=TRUE)
> prev2019
      Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct
2019 279675.3 244989.8 243253.3 230017.2 170273.7 183303.6 221124.2 238576.7 200297.7 228292.6
      Nov      Dec
2019 277537.1 265701.9
```

```
>plot(prev2019,xlab="Anos",ylab="Número de Dormidas", main="")
>lines(dados,type="",lty=1,col="black")
>legend("topleft",c("Observado","Previsto RNA_mlp"),cex=0.8,col=c("black","blue"),lty=1)
```

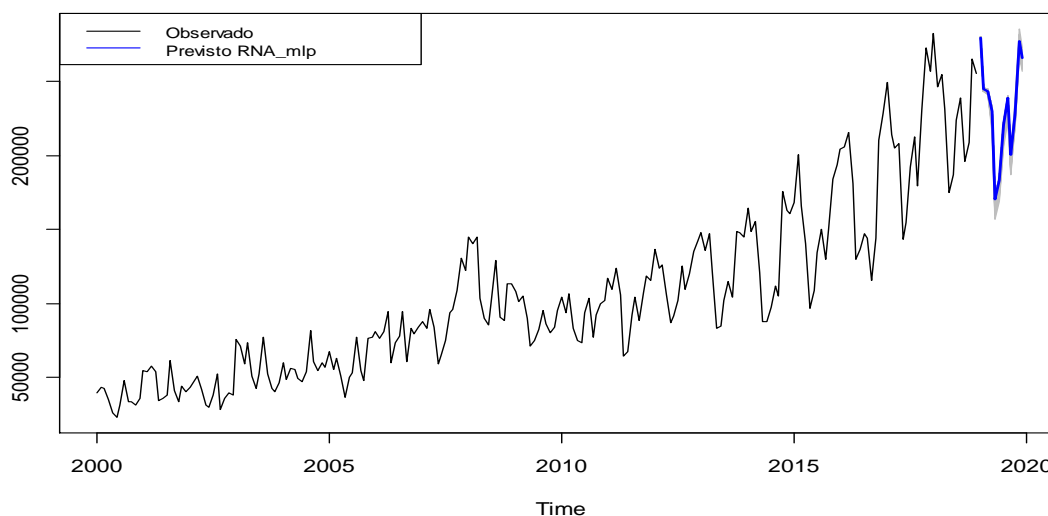


Figura 4.31: Resultado da previsão a um passo à frente usando a função *mlp()* do R, com o modelo **m11**.

Fonte: Elaboração própria

Tabela 4.11: Comparação dos valores observados com os previstos pela função *mlp()* para o modelo **m11**, ano de 2019.

Meses	Valores observados	Valores esperados	Erro MAPE do modelo (%)
Janeiro	267.940	279.675	4,38
Fevereiro	237.203	244.990	3,28
Março	233.508	243.253	4,17
Abril	207.956	230.017	10,61
Mai	152.691	170.274	11,52
Junho	192.645	183.304	4,85
Julho	217.145	221.124	1,83
Agosto	235.091	238.577	1,48
Setembro	193.380	200.298	3,58
Outubro	224.602	228.293	1,64
Novembro	257.670	277.537	7,71
Dezembro	236.604	265.702	12,30
Média do Erro MAPE (%)			5,61

Fonte: Elaboração própria

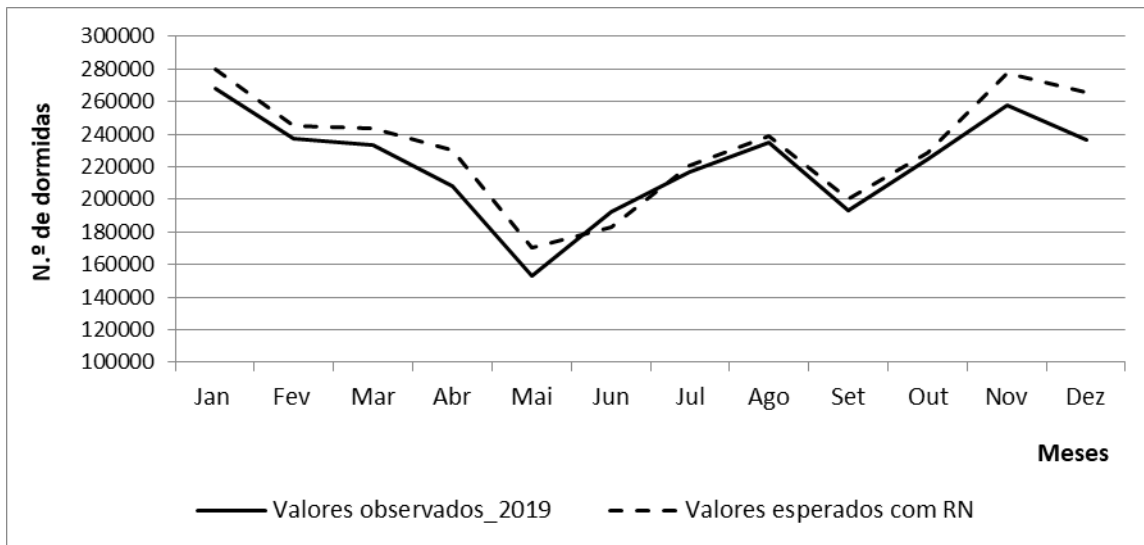


Figura 4.32: Comparação dos valores observados com os previstos pela função $mlp()$ com o modelo $m11$, ano de 2019.

Fonte: Elaboração própria

4.2.6.2. Modelação com a função $nnetar()$

Utilizando a função $nnetar()$, descrito na secção 3.7.8, é criado diferentes modelos de rede neuronal *perceptron* de uma camada e de estrutura do tipo *feedforward* com o objetivo de compara-los e escolher o que apresenta menor erro MAPE. O comando da função $nnetar()$ será aplicada sempre da mesma forma para todos os modelos, alterando apenas os parâmetros p e k em cada modelo.

Assim, para identificar cada modelo é utilizado uma notação semelhante à descrita na secção 3.7.8 do tipo $RNAR(p,P,k)[m]$, onde p é o número de *lags* de entrada, P é a ordem de sazonalidade (para a série em estudo o $P=1$, como verificado na aplicação do modelo SARIMA), k é o número de neurónios na camada oculta dado por $\frac{p+P+1}{2}$ (arredondado para o inteiro mais próximo) e m o período sazonal.

Inicialmente, à semelhança da implementação da função $mlp()$, foi dividido os dados em conjuntos de treino (2000:2015) e teste (2016:2018).

```
> treino<- window(dados, end = c(2015, 12))
> teste <- window(dados, start = c(2016,1),end=c(2018,12))
```

Seguidamente diferentes modelos foram analisados por meio da implementação da função `nnetar()` com o respectivo comando básico.

```
nnetar(dados,decay=, p=,P=,size=,lambda=NULL,repeat=, subset=NULL)
```

A Tabela 4.12 é um resumo dos erros de treino e teste de cada modelo pesquisado (Ver Anexo VI para visualizar todos os gráficos de cada modelo). A escolha recai sobre o modelo **m32** por apresentar o menor erro MAPE de teste.

Foram pesquisados doze modelos. Não se continuou com mais modelos, pois se verificou, pelos gráficos (Anexo VI), indícios de *overtraining* nos modelos **m33**, **m34** e **m35**, bem como o erro de teste em cada um estaria a ser superior ao do encontrado no modelo **m32**.

Tabela 4.12: Comparação dos erros de treino e teste de cada modelo de Redes Neurais (RN), com uso da função `nnetar()`.

Modelo	Estrutura do Modelo RN	MAPE (treino) (%)	MAPE (teste) (%)
m24	RNAR(2,1,2)[12]	12,217	12,149
m25	RNAR(4,1,3)[12]	10,435	10,550
m26	RNAR(6,1,4)[12]	8,120	10,249
m27	RNAR(8,1,5)[12]	5,828	8,724
m28	RNAR(10,1,6)[12]	3,980	6,517
m29	RNAR(12,1,7)[12]	2,891	5,966
m30	RNAR(14,1,8)[12]	1,705	4,128
m31	RNAR(16,1,9)[12]	0,852	3,971
m32	RNAR(18,1,10)[12]	0,452	3,293
m33	RNAR(20,1,11)[12]	0,238	4,006
m34	RNAR(22,1,12)[12]	0,092	3,925
m35	RNAR(24,1,13)[12]	0,064	4,246

Fonte: Elaboração própria

Sendo assim, de seguida, vai apresentar-se os resultados mais detalhados do modelo escolhido (**m32**):

Modelo m32: NNAR(18,1,10)[12]

O ajustamento do modelo:

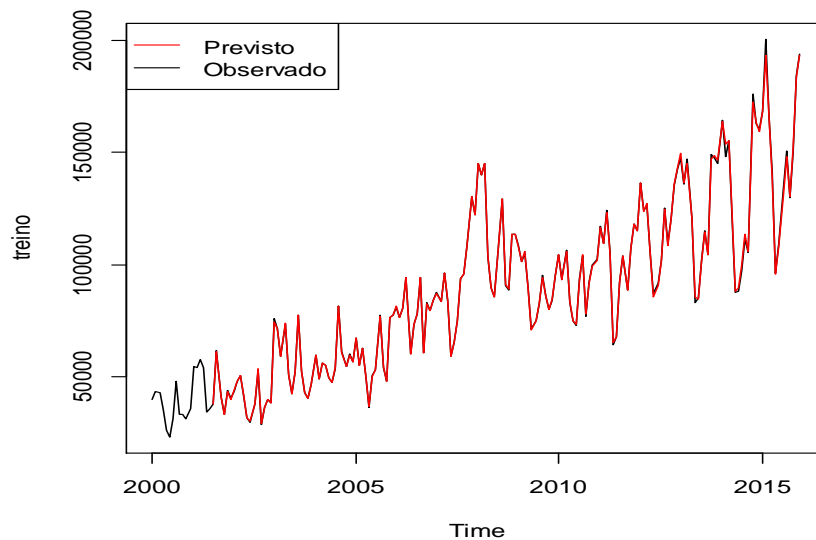


Figura 4.33: Ajustamento do modelo **m32** RNAR(18,1,10)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m32** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -43.48964 1003.196 480.1871 0.002142562 0.4517474 0.1290983 0.04064403
```

Seguidamente é feita a validação do modelo **m32** comparando-o com os dados do conjunto de teste com apoio da Figura 4.34 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -4300.887 14555.74 5445.761 -2.504617 3.292603 0.6674803 0.4121625
```

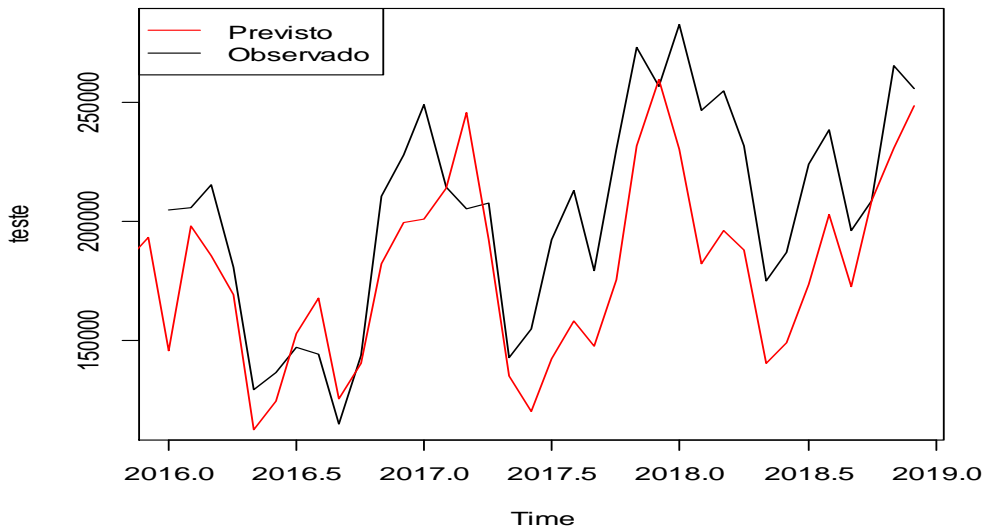


Figura 4.34: Ajustamento do modelo $m32$ RNAR(18,1,10)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12].

Fonte: Elaboração própria

Com relação a previsão mensal a um passo à frente (ano de 2019) a Figura 4.35 mostra que a previsão não segue o comportamento da série. Na Tabela 4.11 é feita a comparação dos valores previstos com os valores observados.

```
> prev2019<-forecast(rede_teste,h=12,PI=TRUE)
> prev2019
```

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2019	209501.0	179807.98	244402.0	166235.06	269485.1
Feb 2019	171149.6	141241.97	205321.9	128742.12	225760.7
Mar 2019	156707.4	126653.40	191302.1	111894.43	213320.2
Apr 2019	133608.5	108961.12	162974.6	98935.18	180656.9
May 2019	106159.1	87662.86	128349.2	80145.32	140213.5
Jun 2019	104229.1	85246.76	126569.4	77718.43	142941.8
Jul 2019	113875.3	95029.22	138393.1	86613.81	152365.3
Aug 2019	121918.3	99625.40	150035.0	86891.16	167309.9
Sep 2019	121969.2	97649.00	147656.0	86995.88	161395.9
Oct 2019	129775.9	102232.87	155347.3	94016.44	171656.6
Nov 2019	137228.7	108094.57	165878.0	98703.93	182705.5
Dec 2019	142645.8	113224.96	171084.4	100994.36	196036.1

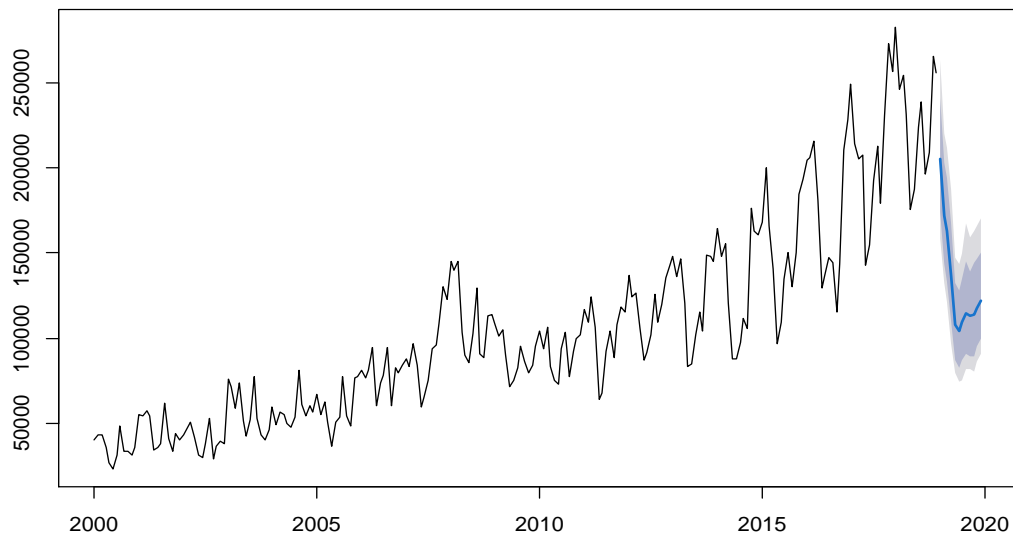


Figura 4.35: Resultado da previsão a um passo à frente usando a função `nnetar()` do R, com o modelo **m32**.

Fonte: Elaboração própria

Tabela 4.13: Comparação dos valores observados com os previstos pela função `nnetar()` para o modelo **m32**, ano de 2019.

Meses	Valores observados	Valores esperados	Erro MAPE do modelo (%)
Janeiro	267.940	209.501	21,81
Fevereiro	237.203	171.150	27,85
Março	233.508	156.707	32,89
Abril	207.956	133.609	35,75
Mai	152.691	106.159	30,47
Junho	192.645	104.229	45,90
Julho	217.145	113.875	47,56
Agosto	235.091	121.918	48,14
Setembro	193.380	121.969	36,93
Outubro	224.602	129.776	42,22
Novembro	257.670	137.229	46,74
Dezembro	236.604	142.646	39,71
Média do Erro MAPE (%)			38,00

Fonte: Elaboração própria

4.2.6.3. Considerações Finais

Das duas funções do R mais utilizadas e próprias para séries temporais univariadas (a $mlp()$ e a $nnetar()$), a função $mlp()$ com uma camada oculta apresentou melhores resultados (MAPE=5,61%) comparando-a com a função $nnetar()$ com uma camada oculta. Esta apresentou um erro MAPE=38% indicando uma precisão razoável, segundo Lewis (1982).

Em síntese, a função $mlp()$ é a que se mostrou ser mais adequada para modelar os dados da série de dormidas mensais e, por isso, na secção 4.10 será feita a comparação de seus resultados com os demais modelos e verificar em que aspetos é uma mais valia nas previsões mensais.

4.2.7. Aplicação do Algoritmo de Atribuição do Erro

Da aplicação do modelo SARIMA na secção 4.3.4 deste Capítulo, dois modelos se apresentaram como os melhores, $SARIMA(1,1,1)(0,1,1)_{12}$ e $SARIMA(2,1,0)(0,1,1)_{12}$ que se denominou por **m2** e **m3**, respetivamente. O modelo **m3** foi o modelo final escolhido. Os valores previstos a um passo à frente de ambos os modelos são apresentados na Tabela 4.14, onde também se apresentam todos os resultados derivados dos passos do algoritmo (Tabela 4.14).

Tabela 4.14: Resultados da aplicação do Algoritmo de Atribuição do Erro, na previsão do ano de 2019.

Meses	m2	m3	Média	Erro (%)	Iteração 1	Erro 1 (%)	Iteração 2	Erro 2 (%)	Iteração 3	Erro 3 (%)	Iteração 4	Erro 4 (%)
Janeiro	280256	273666	276961	1,20	272584	0,40	269936	1,36	268013	2,07	265586	2,95
Fevereiro	262151	253985	258068	1,61	253989	0,00	251522	0,97	249730	1,68	247469	2,57
Março	264536	254351	259443,5	2,00	255343	0,39	252862	0,59	251062	1,29	248788	2,19
Abril	246118	234672	240395	2,44	236596	0,82	234297	0,16	232628	0,87	230522	1,77
Mai	193793	181532	187662,5	3,38	184697	1,74	182902	0,75	181600	0,04	179955	0,87
Junho	204043	191583	197813	3,25	194687	1,62	192795	0,63	191422	0,08	189689	0,99
Julho	233153	220787	226970	2,80	223383	1,18	221213	0,19	219637	0,52	217648	1,42
Agosto	246935	234487	240711	2,65	236907	1,03	234605	0,05	232934	0,66	230825	1,56
Setembro	214689	201811	208250	3,19	204959	1,56	202968	0,57	201522	0,14	199697	1,05
Outubro	243676	230599	237137,5	2,84	233390	1,21	231122	0,23	229476	0,49	227398	1,39
Novembro	288538	276033	282285,5	2,27	277824	0,65	275125	0,33	273166	1,04	270692	1,93

Dezembro	284121	271384	277752,5	2,35	273363	0,73	270707	0,25	268779	0,96	266345	1,86
Raiz quadrada do Erro MAPE (%)				1,58		0,97		0,71		0,91		1,31

Fonte: Elaboração própria

(cont.)

Meses	Iteração 5	Erro 5 (%)	Iteração 6	Erro 6 (%)	Iteração 7	Erro 7 (%)	Iteração 8	Erro 8 (%)	Iteração 9	Erro 9 (%)	Iteração 10	Erro 10 (%)
Janeiro	262112	4,22	257574	5,88	252003	7,92	245461	10,31	238033	13,02	229818	16,02
Fevereiro	244232	3,84	240003	5,51	234813	7,55	228717	9,95	221795	12,67	214141	15,69
Março	245533	3,47	241282	5,14	236064	7,19	229936	9,60	222977	12,33	215283	15,36
Abril	227506	3,05	223567	4,73	218732	6,79	213054	9,21	206606	11,96	199476	15,00
Mai	177601	2,17	174526	3,86	170752	5,94	166319	8,38	161286	11,15	155720	14,22
Junho	187207	2,28	183966	3,98	179987	6,05	175315	8,49	170009	11,26	164142	14,32
Julho	214801	2,71	211082	4,40	206517	6,46	201156	8,89	195068	11,65	188337	14,70
Agosto	227805	2,85	223861	4,53	219020	6,60	213334	9,02	206878	11,77	199739	14,82
Setembro	197085	2,34	193672	4,03	189484	6,11	184565	8,55	178979	11,31	172803	14,37
Outubro	224423	2,68	220538	4,36	215768	6,43	210167	8,86	203807	11,62	196773	14,67
Novembro	267151	3,22	262525	4,89	256848	6,95	250180	9,37	242609	12,11	234237	15,14
Dezembro	262861	3,14	258310	4,82	252723	6,88	246163	9,29	238713	12,04	230475	15,07
		1,73		2,16		2,60		3,03		3,45		3,87

Fonte: Elaboração própria

Com a aplicação do algoritmo, feita no Microsoft Excel, versão 2010, foram realizadas dez iterações ($k=1,2, \dots, 10$) para exemplificar. Mas, pelo critério de paragem, um novo conjunto de valores previsto C é aquele que se encontra na sexta iteração ($k=6$), ou seja, C^6 , pois o erro da sexta iteração é maior que 2 (Figura 4.36). Com a escolha do novo conjunto de valores previstos, espera-se que o erro de previsão seja menor que o erro do modelo escolhido $m3$, isto é, $MAPE_{C^6} < MAPE_{M_3}$.

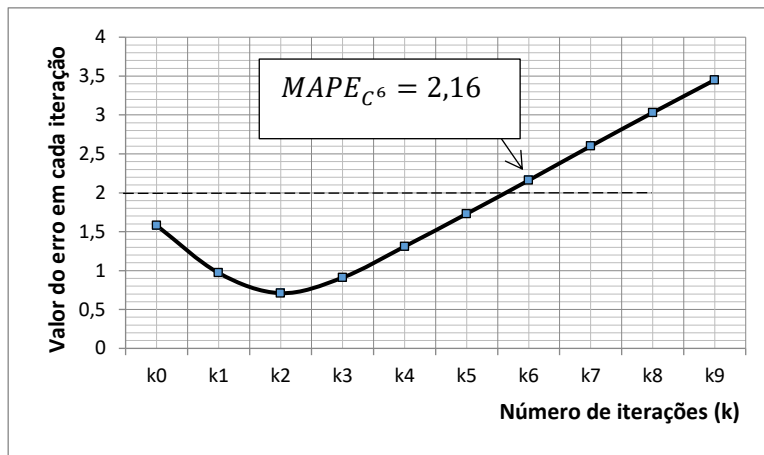


Figura 4.36: Escolha da iteração k=6.

Fonte: Elaboração própria

Os resultados comparativos entre as previsões obtidas pelo modelo **m3** $SARIMA(2,1,0)(0,1,1)_{12}$ e da aplicação do Algoritmo de Atribuição do Erro (AAE) mostram (Tabela 4.15 e Figura 4.37) que as previsões mensais melhoraram, com exceção dos meses de janeiro e agosto, reduzindo o MAPE total de 6,77% para 4,61%.

4.2.7.1. Considerações Finais

O Algoritmo de Atribuição do Erro (AAE) proposto mostrou, com relação à série temporal em estudo, ser possível encontrar novos valores previstos, a partir dos resultados de um modelo SARIMA, com um erro MAPE significativamente mais baixo quando comparado com os dados observados.

Tabela 4.15: Comparação das previsões obtidas do modelo SARIMA **m3** e da aplicação do Algoritmo de Atribuição do Erro com os valores observados do ano de 2019 e os respectivos erros.

Meses	Dados Observados 2019	Dados Previstos (SARIMA)	MAPE (%)	Dados Previstos (AAE)	MAPE (%)
Janeiro	267.940	273.666	2,14	262.112	2,18
Fevereiro	237.203	253.985	7,07	244.232	2,96
Março	233.508	254.351	8,93	245.533	5,15
Abril	207.956	234.672	12,85	227.506	9,40

Maio	152.691	181.532	18,89	177.601	16,31
Junho	192.645	191.583	0,55	187.207	2,82
Julho	217.145	220.787	1,68	214.801	1,08
Agosto	235.091	234.487	0,26	227.805	3,10
Setembro	193.380	201.811	4,36	197.085	1,92
Outubro	224.602	230.599	2,67	224.423	0,08
Novembro	257.670	276.033	7,13	267.151	3,68
Dezembro	236.604	271.384	14,70	262.861	11,10
Média do Erro MAPE (%)			6,77		4,61

Fonte: Elaboração própria

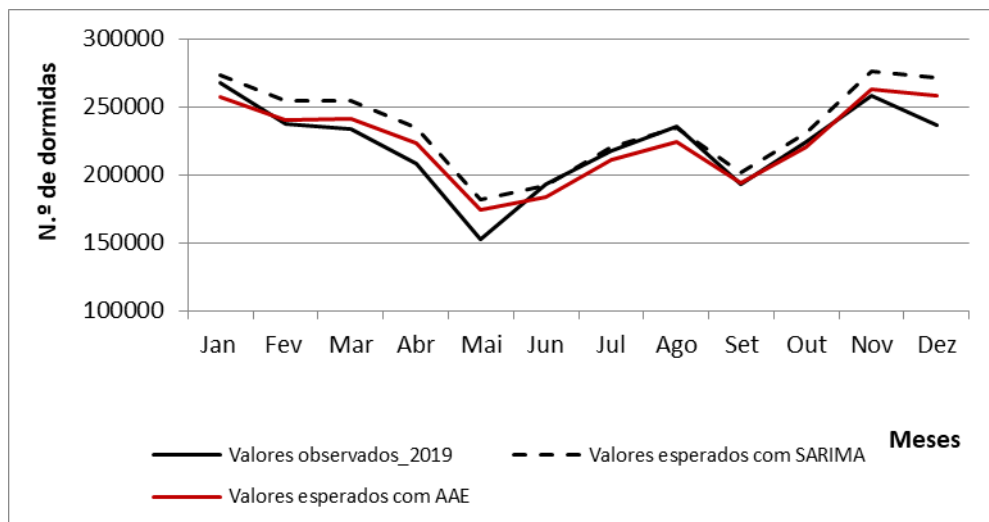


Figura 4.37: Comparação das previsões obtidas do modelo SARIMA *m3* e da aplicação do Algoritmo de Atribuição do Erro com os valores observados, ano de 2019.

Fonte: Elaboração própria

4.3. COMPARAÇÃO ENTRE OS MODELOS

Nesta secção será feita a comparação do modelo SARIMA com todos os demais analisados neste trabalho. A comparação é apoiada pela Tabela 4.16 e pelas Figuras 4.38 a 4.40.

4.3.1. SARIMA versus Holt-Winters

Como se pode constatar na Tabela 4.16, o modelo $SARIMA(2,1,0)(0,1,1)_{12}$ (Modelo **m3**, equação (111)), em termos gerais, superou o modelo de Holt-Winters (Modelo **m5**, equações (112-114), com um erro MAPE igual a 6,77% e 7,94% respetivamente. Entretanto, uma análise individual de cada mês, mostra que o modelo **m5** de Holt-Winters tende a ser mais eficiente nas previsões para os meses de maio (fim da época alta), julho (época baixa) e de setembro a dezembro (época alta), conforme a Figura 4.38. Ambos os modelos funcionam muito bem para os meses onde finda a época alta, percorrendo toda a época baixa e durante último trimestre do ano considerado também época alta. O modelo **m5** de Holt-Winters consegue captar melhor os picos que ocorrem no início e no fim da época alta. Ainda, mencionar que os dois modelos praticamente se confundem no sentido de ser quase impercetível nos resultados gráficos saber se é um ou outro; mas o modelo **m3** SARIMA confirma a sua posição histórica de melhor eficiência, com uma diferença de erro de 1,17%.

4.3.2. SARIMA versus Redes Neurais

Como se pode constatar na Tabela 4.16, o modelo $SARIMA(2,1,0)(0,1,1)_{12}$ (modelo **m3**), em termos gerais, foi superado pelo modelo de Redes Neurais (modelo **m11**, RNAR(18,1,10)[12]). Este com um erro MAPE igual a 5,61% contra os 6,77% do modelo SARIMA. Uma análise individual de cada mês, mostra que o modelo SARIMA **m3** tende a ser mais eficiente nas previsões para os meses de janeiro (época alta), junho, julho e agosto (épocas baixas) e novembro (época alta), conforme Figura 4.39.

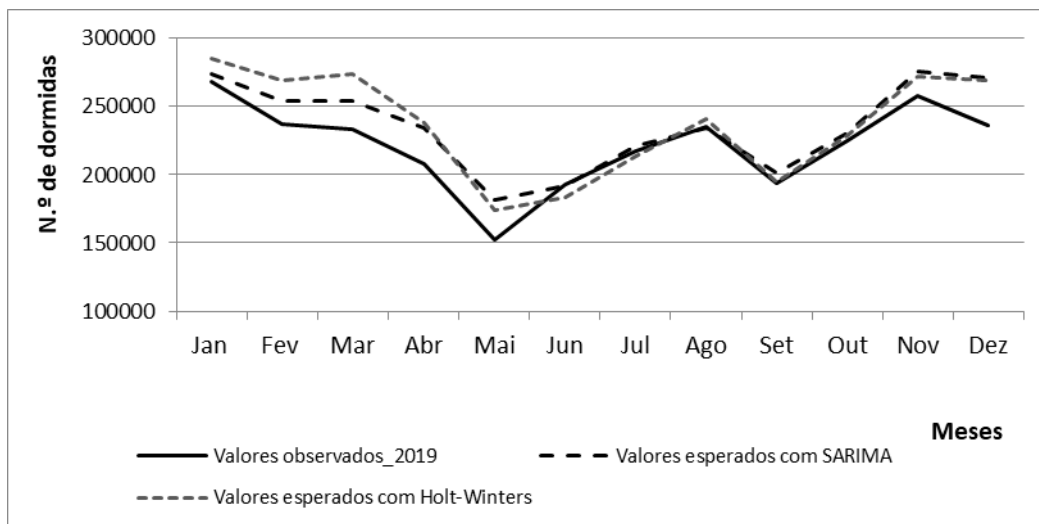


Figura 4.38: Comparação das previsões obtidas pelos modelos SARIMA (m3) versus Holt-Winters (m5).

Fonte: Elaboração própria

Ambos os modelos, praticamente, sobrepõem-se a meio da época baixa até a época alta, no terceiro trimestre do ano. Entretanto, o modelo de m11 NR consegue captar melhor os picos que ocorrem nos meses de maio e setembro. As previsões de ambos seguem o mesmo comportamento da série. E finalmente, estes resultados trazidos pelo modelo m11 RN, confirma o relatado no estado da arte, que os métodos de Inteligência Artificial têm vindo a revelar-se com melhores resultados, apesar de suas limitações teóricas e computacionais.

4.3.3. SARIMA versus Algoritmo de Atribuição do Erro

Como se pode constatar na Tabela 4.16, os resultados do modelo m3 SARIMA, em termos gerais, foram melhorados pelo Algoritmo de Atribuição do Erro (AAE). Este com um erro MAPE igual a 4,61% em melhoria dos 6,77% do modelo m3 SARIMA. Uma análise individual de cada mês, mostra que as melhorias não se verificaram nas previsões para os meses de janeiro (época alta), junho e agosto (épocas baixas), conforme Figura 4.40.

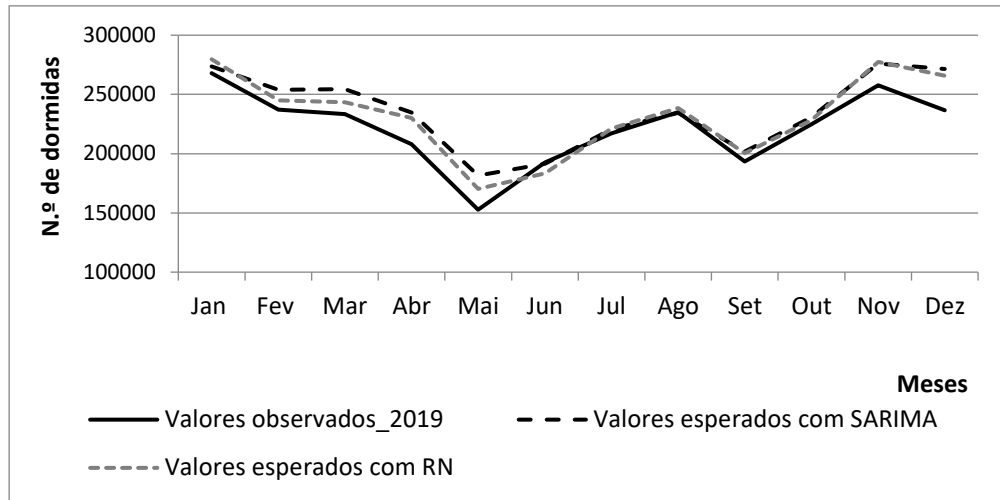


Figura 4.39: Comparação das previsões obtidas pelos modelos **m3** SARIMA versus **m11** Rede Neuronal.

Fonte: Elaboração própria

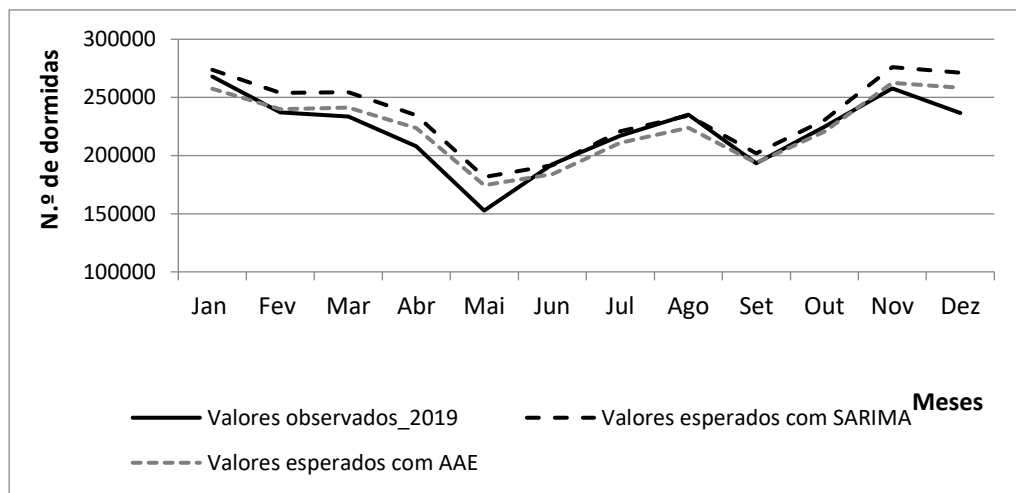


Figura 4.40: Comparação das previsões obtidas pelos modelos **m3** SARIMA versus a aplicação do Algoritmo de Atribuição do Erro.

Fonte: Elaboração própria

Tabela 4.16: Resultados comparativos entre os métodos de previsão.

Meses (2019)	Valores observados	Box-Jenkins, modelo m3 SARIMA		Holt-Winters modelo m5		Redes Neuronalis modelo m11		Algoritmo de Atribuição do Erro	
		Valores esperados com Sarima	Erro MAPE do modelo (%)	Valores esperados com Holt-Winters	Erro MAPE do modelo (%)	Valores esperados com RN	Erro MAPE do modelo (%)	Valores esperados com AAE	Erro MAPE do modelo (%)
Janeiro	267.940	273.666	2,14	285.179	6,43	279.675	4,38	257.574	3,87
Fevereiro	237.203	253.985	7,07	268.997	13,40	244.990	3,28	240.003	1,18
Março	233.508	254.351	8,93	273.596	17,17	243.253	4,17	241.282	3,33
Abril	207.956	234.672	12,85	238.289	14,59	230.017	10,61	223.567	7,51
Mai	152.691	181.532	18,89	173.694	13,76	170.274	11,52	174.526	14,30
Junho	192.645	191.583	0,55	183.936	4,52	183.304	4,85	183.966	4,51
Julho	217.145	220.787	1,68	213.767	1,56	221.124	1,83	211.082	2,79
Agosto	235.091	234.487	0,26	240.563	2,33	238.577	1,48	223.861	4,78
Setembro	193.380	201.811	4,36	194.493	0,58	200.298	3,58	193.672	0,15
Outubro	224.602	230.599	2,67	228.706	1,83	228.293	1,64	220.538	1,81
Novembro	257.670	276.033	7,13	271.879	5,51	277.537	7,71	262.525	1,88
Dezembro	236.604	271.384	14,70	268.829	13,62	265.702	12,30	258.310	9,17
Média do erro MAPE (%)			6,77		7,94		5,61		4,61

Fonte: Elaboração própria

4.3.4. Comparação dos modelos estatísticos e rede neuronal com a previsão feita com a Análise Lógica Intuitiva

Na secção 2.5 foi mencionada a previsão do número total de hóspedes para o ano de 2019 obtido com a Análise Lógica Intuitiva realizada no âmbito da elaboração do *Master Plan* do Turismo da ilha do Sal (MF, 2019). Foram feitas três previsões na qual cada uma se encontra relacionada com um tipo de cenário. Dado que neste trabalho, a série temporal em análise está composta pelo número de dormidas (e não pelo número de hóspedes), então será substituído o número de hóspedes previstos pelo número de dormidas previstos por forma a realizar as devidas comparações. Levando em consideração que a estadia média por

hospedagem¹¹ (fórmula 116), na ilha do Sal, em 2019, foi de 7,1 dias (Tabela 4.17), então pode-se fazer essa substituição (Tabela 4.18 , terceira coluna), utilizando a fórmula (117).

$$EMH = ND/NH \quad (116)$$

em que,

EMH – estadia média por hospedagem; ND – número de dormidas; NH – número de hóspedes.

$$NDP = EMH \times NHP \quad (117)$$

em que,

NDP – número de dormidas previstos; EMH – estadia média por hospedagem; NHP – número de hóspedes previstos.

Tabela 4.17: Dados mensais do número de hóspedes, dormidas e estadia média por hospedagem do ano de 2019.

Meses	Hóspedes	Dormidas	Estadia média por hospedagem
Janeiro	38.336	267.940	7,0
Fevereiro	34.124	237.203	7,0
Março	33.446	233.508	7,0
Abril	28.809	207.956	7,2
Maio	21.291	152.691	7,2
Junho	26.275	192.645	7,3
Julho	29.438	217.145	7,4
Agosto	30.786	235.091	7,6
Setembro	27.751	193.380	7,0
Outubro	32.362	224.602	6,9
Novembro	35.605	257.670	7,2
Dezembro	34.216	236.604	6,9
Total	372.439	2.656.435	
		Média total	7,1

¹¹ Segundo o INE, a estadia média por hospedagem é a relação entre o número de dormidas e o número de hóspedes que deram origem a essas dormidas (INE, 2020).

Fonte: Elaboração própria, com dados do INE (2022)

Tabela 4.18: Previsão, por cenários, do número de dormidas e do número de hóspedes para o ano de 2019.

Cenários	Número de hóspedes previstos	Número de dormidas previstos
Ganhos otimistas	424.576	3.014.490
Ganhos moderados	413.204	2.933.748
Poucos ganhos	398.040	2.826.084

Fonte: Elaboração própria

Tendo os valores previstos do número total de dormidas para o ano de 2019 (Tabela 4.18), seguidamente é realizada a comparação dessas previsões com as previsões feitas com o modelo SARIMA (2,1,0) (0,1,1)[12], a Rede Neuronal Artificial RNAR(12,1,7)[12], o Método Holt Winters e com a aplicação do Algoritmo de Atribuição do Erro (Tabela 4.19).

Da comparação é possível verificar que as previsões realizadas pelos modelos estatísticos e rede neuronal se aproximam todos da previsão do cenário de “Poucos Ganhos”. Pelos valores do erro MAPE, as previsões feitas com os cenários de “Ganhos Otimista” e “Ganhos Moderados” se mostraram pouco precisas quando comparado com os modelos. Entretanto a rede neuronal conseguiu prever com maior precisão o total de dormidas em 2019, com um erro de 4,77%. Porém o Algoritmo de Atribuição do Erro, que otimiza os resultados do modelo SARIMA, proporciona uma previsão mais próxima do total de dormidas observadas, com um erro de precisão de 1,30%.

Tabela 4.19: Comparação das previsões feitas com a Análise Lógica Intuitiva com os Modelos Estatísticos e Rede Neuronal, para o ano de 2019.

Número total de dormidas observado em 2019	2.656.435		Erro MAPE (%)
	Análise Lógica Intuitiva		
	Ganhos otimistas	3.014.490	13,48
	Ganhos moderados	2.933.748	10,44

Número total de dormidas previstos para 2019	Poucos Ganhos	2.826.084	6,39
	Modelos Estatísticos e Rede Neuronal		
	SARIMA	2.824.890	6,34
	Rede Neuronal	2.783.044	4,77
	Holt Winters	2.841.928	6,98
	AAE	2.690.906	1,30

Fonte: Elaboração própria

4.3.5. Considerações Finais

Na sequência das análises por pares, pode-se considerar o seguinte: o modelo SARIMA (modelo **m3**), comparativamente com os outros modelos conseguiu prever com mais eficiência os meses de junho e agosto (considerado épocas baixas); o método de Holt-Winters (modelo **m5**) os meses de julho e novembro; o modelo de Redes Neurais (modelo **m11**) os meses de outubro e dezembro.

Olhando apenas para o SARIMA, o Holt-Winters e a Rede Neuronal, considera-se o seguinte: o modelo SARIMA conseguiu prever com mais eficiência os meses de janeiro, junho e agosto; o método de Holt-Winters os meses de julho, Setembro e novembro; o modelo de Redes Neurais os restantes meses. Quando se aplica o Algoritmo de Atribuição do Erro (AAE) aos resultados do SARIMA, este modelo mostrou-se mais eficiente em prever o segundo semestre do ano, e a Rede Neuronal mais eficiente no primeiro semestre.

No cômputo geral, o modelo de Rede Neuronal (modelo **m11**) se mostrou mais eficiente na previsão das dormidas mensais para o ano de 2019, com um erro MAPE=5,61%. O modelo SARIMA (modelo **m3**) superou apenas o Método de Holt-Winters (modelo **m5**). Entretanto, quando se melhora os resultados do SARIMA por meio do Algoritmo de Atribuição do Erro proposto, torna-se então o mais eficiente com um erro MAPE=4,61%.

Verificou-se que os modelos de SARIMA, Rede Neuronal e Holt-Winters, ambos se mostraram com dificuldades (no sentido de existir um erro superior a 10%) em obter uma

previsão mais próxima do observado entre os meses de Outubro e Maio, donde se encontra incluído o período de época alta (Outubro a Março). Entretanto o modelo de Rede Neuronal é o que se adaptou melhor.

Quando é feito o somatório das previsões mensais de cada modelo, verifica-se que os modelos estatísticos e as redes neuronais, especificamente o modelo SARIMA e a Rede Neuronal, produzem melhores resultados a nível da precisão das previsões do que a técnica da Análise Lógica Intuitiva utilizada para prever o número total de dormidas para o ano de 2019.

CAPÍTULO 5: CONCLUSÕES

5.1. CONSIDERAÇÕES FINAIS

As conclusões a seguir apresentadas correspondem a cada objetivo proposto no presente trabalho de investigação.

Com relação ao Objetivo 1, foi determinado que o Ciclo de Vida do destino (ilha do Sal), de 2010 a 2018, se encontrava na fase de Desenvolvimento, corroborando com o que já havia sido avançado no Master Plan do Turismo da Ilha do Sal apresentado pelo Ministério das Finanças em 2019. Entretanto, em relação à competitividade, o destino permaneceu na fase de Exploração por um período de oito anos (2010-2018). Dada a sua perda relativa de posição, em relação ao mesmo espaço competitivo (mercado de 'sol e praia') é uma indicação de que a ilha do Sal, como destino turístico, está numa fase Estagnada devendo crescer seu Índice de Desenvolvimento Turístico em 48%, caso os seus concorrentes se mantiverem sem melhorias. Verificou-se, ainda, que os dados da ilha do Sal e os dados do país produzem praticamente os mesmos resultados a nível do posicionamento relativo e mudança de fase, pelo que os dados de um podem ser usados na análise do outro.

Para o Objetivo 2, empiricamente, pelo senso comum e pela visualização gráfica da série de dormidas mensais já se percebia que um dos modelos do método de Box-Jenkins que poderia ser aplicado seria o SARIMA, devido à presença da sazonalidade que é uma característica forte no turismo na ilha do Sal. Tal sazonalidade veio a ser confirmada com o uso do teste de Webel-Ollech cujo $p\text{-value} = 0,001 < 0,05$. Dada a componente de sazonalidade da série, foi escolhido também o Método de Suavização Exponencial Tripla (Holt-Winters). A série em estudo mostrou ter uma tendência crescente, com uma taxa de crescimento praticamente constante verificado no período entre 2000 e 2007 e também a partir de 2010, e uma certa variação na sazonalidade ao longo do tempo. Ainda foi concluído que a série é não estacionária.

Quanto ao Objetivo 3, dentre os modelos e métodos estudados e aplicados neste trabalho, sendo eles: o modelo SARIMA (do Método de Box-Jenkins), o Método de Suavização Exponencial Tripla (Holt-Winters), e as Redes Neurais Artificiais; o modelo de Rede Neuronal Artificial (RNAR(12,1,7)[12]), de 12 lags de entrada e 7 neurónios na camada oculta, mostrou-se superior aos demais em prever o número de dormidas mensais nos estabelecimentos turísticos na ilha do Sal, para o ano de 2019, com um erro de precisão de 5,61% (-1,16 p.p. com relação ao erro de precisão do modelo SARIMA, utilizando o indicador de erro MAPE). Esta rede foi implementada com a função *mlp()* em uso no *Software R* versão 4.0.5. Esta função se mostrou mais adaptada aos dados da série do que a função *nnetar()*.

O SARIMA (2,1,0) (0,1,1)[12] foi o segundo modelo que conseguiu prever os dados futuros da série com um menor erro de precisão, isto é um MAPE de 6,77%. Este modelo se mostrou adaptar muito bem à séries não estacionárias. Quanto ao método de Holt Winters, este obteve um erro de precisão de 7,94% em prever a um passo a frente da série, ou seja, em prever os próximos doze meses correspondentes ao ano de 2019. Todos os três resultados obtidos são considerados aceitáveis para preverem o número de dormidas mensais a curto prazo, porém a Rede Neuronal Artificial (RNAR(12,1,7)[12]) se mostrou mais eficiente corroborando com a literatura pesquisada no estado da arte.

Quando é feito o somatório das previsões mensais de cada modelo, verifica-se que os modelos estatísticos e as redes neuronais, especificamente o modelo SARIMA e a Rede Neuronal, produzem melhores resultados a nível da precisão das previsões (erro MAPE igual a 6,34% e 4,77% respetivamente) do que a técnica da Análise Lógica Intuitiva utilizada para prever o número total de dormidas para o ano de 2019. Isto é, o modelo SARIMA (2,1,0) (0,1,1)[12], e a Rede Neuronal RNAR(12,1,7)[12] superaram a técnica da Análise Lógica Intuitiva utilizada no *Master Plan* do Turismo da ilha do Sal. Porém o Algoritmo de Atribuição do Erro, que otimiza os resultados do modelo SARIMA,

proporciona uma previsão mais próxima do total de dormidas observadas, com um erro de precisão de 1,30%.

Do Objetivo 4, a principal conclusão da aplicação da Lei de Benford aponta que os resultados obtidos suportam a fiabilidade dos dados secundários (disponibilizados pelo Instituto Nacional de Estatística de Cabo Verde), apesar desses dados mensais não obedecerem à Lei de Benford, embora exista uma semelhança no formato da curva. Entretanto, as discrepâncias entre as frequências relativas observadas e esperadas foram estatisticamente significativas.

Ainda, os resultados sugerem que o número de referência de dormidas mensais em estabelecimentos de alojamento pode ser simulado usando a Lei de Benford e os eventos ocasionais podem ser considerados possíveis perturbações à lei e os investimentos em infraestruturas turísticas possíveis aliados da Lei.

Do estudo é possível inferir que a não adaptação dos dados à Lei de Benford não teve impacto negativo nos resultados dos modelos de previsão aplicados; até porque, todos os modelos selecionados em cada método foram considerados de alta precisão, pois os erros de previsão calculados mantiveram-se abaixo dos 10% considerados por Lewis (1982).

Com relação ao Objetivo 5, foi proposto um algoritmo, denominado Algoritmo de Atribuição do Erro (AAE), que usa os resultados iniciais de dois modelos SARIMA escolhidos e considerados de maior precisão, para melhorar as previsões iniciais. Com a sua aplicação o erro do modelo SARIMA (2,1,0) (0,1,1)[12] diminuiu de 6,77% para 4,98%; seguindo o mesmo comportamento dos dados observados. O AAE conseguiu otimizar o erro produzido pelos resultados do modelo SARIMA, isto é, as previsões feitas pelo modelo SARIMA podem ser atualizadas e otimizadas com o AAE.

Em síntese,

- ✓ O modelo de **Rede Neuronal Artificial (RNAR (12,1,7)[12])** obtido com a função *mlp()* se mostrou mais eficiente na previsão das dormidas mensais para o ano de

2019 (**MAPE=5,61%**), em comparação com o modelo SARIMA (2,1,0) (0,1,1)[12] (MAPE=6,77%) e o Holt-Winters (MAPE=7,94%).

- ✓ O modelo **SARIMA (2,1,0) (0,1,1)[12]** é o segundo modelo que se mostrou ser mais eficiente. Mas, quando se melhora os seus resultados por meio do **Algoritmo de Atribuição do Erro** proposto, o erro reduz-se de 6,77% para **4,61%**, isto é 2,16 p.p.
- ✓ Os modelos estatísticos e de redes neuronais se mostraram mais precisos em prever o número de dormidas anuais, no horizonte a curto prazo, quando comparado com a técnica da Análise Lógica Intuitiva; e portanto, tendem a ser as melhores opções.
- ✓ Tendo como facto de que o turismo é a maior atividade económica da ilha do Sal e com uma grande contribuição na economia Cabo-verdiana, então ter uma previsão precisa (com um erro de precisão aceitável) da procura turística a nível de dormidas mensais é uma ferramenta fundamental e necessária à gestão do destino a fim de tomar melhores decisões no que diz respeito ao investimento e à definição de políticas. Também é essencial à gestão de todos os negócios inerentes a atividade turística, nomeadamente o transporte, a hotelaria, a restauração, a saúde, a imobiliária e a recreação; principalmente pela questão da sazonalidade que afeta o destino o ano inteiro e também pela natural imprevisibilidade de eventos que possam perturbar a procura turística.

5.2. LIMITAÇÕES DA INVESTIGAÇÃO

Duas grandes dificuldades, que se destacam, condicionaram os objetivos inicialmente propostos da investigação, sendo elas:

- Dificuldade relacionada com a obtenção dos dados mensais no período considerado (2000 a 2018) para todas as variáveis consideradas inicialmente no Plano de Tese

como sendo exógenas para a aplicação dos modelos econométricos e de outras extensões do ARIMA.

- Com a pandemia resultante da COVID-19, não foi possível alargar as previsões para além do ano de 2019, deixando assim uma interrogação quanto à precisão dos métodos estudados para mais que um passo à frente da série.

Além das limitações apresentadas, anteriormente, uma outra limitação está relacionada com a obtenção dos dados para o cálculo do IDT. Por um lado, a qualidade dos mesmos ainda é questionável, pois diferentes instituições trabalham com critérios diferentes de recolha. Por outro lado, a indisponibilidade das instituições em ceder os dados a tempo leva a busca de fontes não oficiais cuja qualidade dos dados pode ser questionada. Isto dificulta a comparação entre destinos.

5.3. RECOMENDAÇÕES E NOVAS LINHAS DE PESQUISA

5.3.1. Recomendações

Atualmente a ilha do Sal (e Cabo Verde) encontra-se no meio de uma crise sanitária no contexto da pandemia COVID 19 impactando significativamente a procura turística. Esse obstáculo ao turismo ‘apanhou’ o destino turístico despreparado, pelo que é necessário começar a inserir no leque de planeamento a questão do desenvolvimento de modelos e métodos de previsão que possam integrar esses tipos de eventos inesperados e poder melhorar as estratégias de competitividade. Ademais, é um momento que a ilha do Sal (e Cabo Verde) tem a oportunidades de ser mais criativo e inovador no sentido de diversificar a oferta, criar novos segmentos e descobrir novos mercados emissores, principalmente caso vierem a ocorrer outras crises no mercado europeu e até mesmo mundial.

Os destinos concorrentes selecionados neste estudo, para o cálculo do IDT, são apenas alguns exemplos daqueles que concorrem no mercado de ‘sol e praia’, sem considerar as especificidades de cada um, com o objetivo de conquistar os mesmos mercados

emissores, no caso o mercado europeu. Assim, outros destinos concorrentes podem ser utilizados em outros estudos e para outros mercados emissores.

O IDT utilizado, em certa medida, foi ao encontro das percepções e ideias defendidas por alguns investigadores quanto ao ciclo de vida e competitividade turística de Cabo Verde e da Ilha do Sal no mercado de ‘sol e praia’. Crê-se que o seu uso deve ser continuado. Entretanto convém aplicar o IDT da expressão (1) ou (2) utilizando o conjunto de variáveis referente ao fator “Condições Turísticas”.

Apesar do sector turístico ter diminuído drasticamente as suas atividades devido à pandemia causada pela COVID 19, como consequência não tendo registado muitas dormidas no ano de 2020, comparando com anos anteriores, mesmo assim o modelo de Rede Neuronal proposto (**RNAR (12,1,7)[12]**) pode contribuir para simular e analisar o percentual de dormidas que não se efetivou devido à pandemia; e poderá ser a melhor escolha na previsão do número de dormidas mensais futuras nos estabelecimentos turísticos, logo que o mercado turístico se restabelecer e apresentar-se num cenário igual (ou melhor) do que se encontrava na pré-pandemia COVID 19.

Dado que os dados anuais do número de dormidas ainda se encontram em um horizonte temporal curto e com poucas observações disponíveis, as previsões anuais podem ser estimadas por meio do somatório das previsões mensais. Neste trabalho se mostrou ser um procedimento útil, apesar de não ser um dos objetivos do estudo, porém recomendável.

Aplicar a Lei de Benford em paralelo com a aplicação dos modelos de previsão, pois assim é possível ir avaliando as discrepâncias, caso existam, com a precisão dos modelos.

5.3.2. Novas Linhas de Pesquisa

A avaliação e atualização da série é um desafio contínuo, pelo que outras versões do SARIMA (nomeadamente o SARIMAX) podem ser estudadas e aplicadas à série de dormidas bem como outras estruturas para as Redes Neurais Artificiais. Ainda, mesmo que o algoritmo *Backpropagation* seja aquele mais utilizado e melhores resultados tem produzido, outros algoritmos podem ser aplicados, nomeadamente versões do próprio

Backpropagation ou o algoritmo de Levenberg-Marquardt usado em modelos econométricos. Também fazer uso de diferentes funções de ativação de forma a comparar diferentes tipos de redes.

Na aplicação da Lei de Benford os dados de dormidas de turistas estrangeiros e nacionais encontravam-se juntos. É necessário considerar os dados desagregados de forma a poder comparar os resultados com os produzidos por Matakovic (2021) no seu estudo sobre o turismo na Croácia.

É necessário quantificar a magnitude do impacto dos eventos (nomeadamente, as construções de infraestruturas turísticas e o início de funcionamento das mesmas, eventos desportivos e culturais) na obtenção de melhores resultados em relação à aplicação da Lei de Benford.

Para a aplicação da Lei de Benford foram utilizados 268 dados. É necessário considerar mais dados, incluindo dados das outras ilhas turísticas do arquipélago, para uma melhor comparação em termos da adaptação dos mesmos à Lei.

Investigar o efeito da não adaptação dos dados da série à Lei de Benford na implementação da função *nnetar()*, dado que os modelos da rede neuronal derivados desta função não produziram valores de acurácia satisfatórios.

Reaplicar, no contexto turístico, o Algoritmo de Atribuição do Erro em outras séries temporais mensais com características semelhantes à série em análise neste trabalho de investigação.

REFERÊNCIAS BIBLIOGRÁFICAS

ASA (2018). *Boletim Estatístico de Tráfego – Ano 2018. Movimentos de Aeronaves, Passageiros, Cargas e Correios nos Aeroportos de Cabo Verde – Por Aeroportos e Natureza dos Movimentos*. Aeroporto de Segurança Aérea. Disponível em <https://www.asa.cv/sobre-nos/publicacoes-e-artigos/>

Armstrong, J.S., & Collopy, F. (1992). *Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons*. Disponível em https://repository.upenn.edu/cgi/viewcontent.cgi?article=1075&context=marketing_papers

BM (2018). *Doing Business 2018 – Reforming to Create Jobs. Comparing Business Regulation for Domestic Firms in 190 Economies*. Banco Mundial. Disponível em <https://www.doingbusiness.org/content/dam/doingBusiness/media/Annual-Reports/English/DB2018-Full-Report.pdf>

BM (2019). Banco Internacional para reconstrução e desenvolvimento e sociedade financeira internacional e agência multilateral de garantia dos investimentos – Quadro de parceria do país para a república de cabo verde para o período fiscal 20-25. Banco Mundial. Disponível em <http://documents.worldbank.org/curated/en/852241572961894946/pdf/Cabo-Verde-Country-Partnership-Framework-for-the-Period-FY20-25.pdf>

Butler, R.W. (1980). The concept of a tourism area cycle of evolution: implications for management resources. *Canadian Geographer*, 24, pp. 5-12.

Buhalis, D. (2004). *e-Tourism, Information technology for strategic Tourism Management*, Financial Times, Prentice Hall.

Batista, F.L.A. (2009). *Modelos de Séries Temporais e Redes Neurais Artificiais na Previsão de Vazão* (Dissertação de Mestrado). Disponível em <http://prpg.ufla.br/ppg/esistemas/wp-content/uploads/2012/08/7-Andre-Luis-Franca-Batista-Dissertacao.pdf>

Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford University Press: Oxford.

Baum, T. (1999). Seasonality in tourism: understanding the challenges. *Tourism Economics*, 5(1), 5-8. Disponível em <https://doi.org/10.1177/135481669900500101>

Butler, R. W. (1994). *Seasonality in Tourism: Issues and Problems*. In A. V. Seaton (eds.), *Tourism: the State of the Art*, Chichester: Wiley & Sons.

Butler, R.W. (2001). Seasonality in Tourism: Issues and Implications. In Baum, T., and Lundtorp, S., (eds.). *Seasonality in Tourism*, Advances in Tourism Research Series,

Pergamon, UK: Elsevier Science. Disponível em <https://doi.org/10.1016/B978-0-08-043674-6.50005-2>

Biederman, P.S. (2008). *Travel and tourism: an industry example*. Pearson Education, Upper Saddle River.

Barros, B.M.T. (2014). *Previsão de carga – Comparação de técnicas* (Dissertação de Mestrado). Disponível em <https://repositorio-aberto.up.pt/bitstream/10216/74135/2/31975.pdf>

Barreto M.J. (2002). Introdução às Redes Neurais Artificiais. Laboratório de Conexionismo e Ciências Cognitivas, UFSC. Disponível em <https://06486feb-a-62cb3a1a-sites.googlegroups.com/site/professorrn6/computacao/TC-RedesNeuraisArtificiais.pdf>

Benford, F. (1938). The law of anomalous numbers. In *Proceedings of the American Philosophical Society* 78(4), pp.551-572.

Brida J.G, Garrido N. (2009). Tourism Forecasting using SARIMA Models in Chilean Regions. *International Journal of Leisure and Tourism Marketing*, 2(2), pp.176-190. Disponível em <http://ssrn.com/abstract=1457984>

Caeiro, F.; Mateus, A. (2014). *Testing randomness in R*. Tpe: Package, version 1.0. Disponível em <https://cran.r-project.org/web/packages/randtests/randtests.pdf>

Cannas, R. (2012). An Overview of Tourism Seasonality: Key Concepts and Policies. *Alma Tourism – Journal of Tourism, Culture and Territorial Development*, 3(5), pp.40-58.

Carvalho, C. (2005). *Análise de Séries Temporais*. Disponível em https://www.maxwell.vrac.puc-rio.br/7060/7060_3.PDF

Caudill, M. (1987). Neural Networks Primer, Part I. *AI Expert*, 2(12), pp.46–52, December 1987

Cortez, R.A.P. (1997). Algoritmos Genéticos e Redes Neurais na Previsão de Séries Temporais. (Tese de Mestrado). Disponível em <https://repositorium.sdum.uminho.pt/bitstream/1822/2827/1/mestrado.pdf>

Cho, W., & Gaines, B. (2007). Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance. *The American Statistician*, 61(3), pp.218-223.

CTCV (2019). *Grupo de Trabalho – Competitividade do Turismo CV*. Cedido em 2019.

CMS (2017a). *Situação Socioeconómico das Famílias*. Câmara Municipal do Sal.

CMS (2017b). *Relatório de Atividade*. Câmara Municipal do Sal.

CMS (2018b). *Relatório de Atividade*. Câmara Municipal do Sal.

CMS (2018a). *Inquérito de Levantamento das crianças fora do pré-escolar*. Câmara Municipal do Sal.

CMS (2021). *Plano Estratégico Municipal de Desenvolvimento Sustentável*. Câmara Municipal do Sal.

Coelho, J., & Butler, R. (2012). The Tourism Area Life Cycle: a quantitative approach of the of the tourism area life cycle. *European Journal of Tourism, Hospitality and Recreation*, 3 (1), pp.9-31

Coelho, J., & Lourenço, P. (2015). *A Identificação da Fase do Ciclo de Vida de um Destino Turístico para as escolhas Estratégicas: Uma Proposta de uma Ferramenta expedita*. Disponível em <https://core.ac.uk/download/pdf/72046663.pdf>

Cunha, L. (2013). *Economia e Política do Turismo* (2ª ed.). Lisboa: McGraw-Hill.

Cunha, L., & Abrantes, A. (2013). *Introdução ao Turismo*, (5ª ed.), Lisboa.

Cooper, C., Fletcher, J., Shepherd, R., Gilbert, D., & Wanhill, S. (2001). *Turismo – Princípios e Práticas* (2.ª ed.). Bookman; Porto Alegre.

Cooper, C. (2004). *World Trade Law, Culture, Heritage and Tourism. Towards a Holistic Conceptual Approach*. *Current Issues in Tourism*, 7 (4), pp. 383-398.

Cryer, J. D., & Chan, K.S. (2008). *Time Series Analysis with Application in R*. (2nd ed.). Springer, New York, ISBN-10: 0387759581, pp: 491.

Cunha, L. (2001). *Introdução ao Turismo*; Editorial Verbo; Lisboa.

Chen, L., G. Li, Wu, C., & Shen, S. (2017). Forecasting Seasonal Tourism Demand Using a Multi-Series Structural Time Series Method. *Journal of Travel Research*. Disponível em DOI: [10.1177/0047287517737191](https://doi.org/10.1177/0047287517737191)

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

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management*, 24(3): 323–330.

De Castro A., Bruni. (2000). *Redes Neurais: Uma nova perspectiva na interpretação de informações ambientais*. (Dissertação de Mestrado). Universidade de São Paulo.

Divisekera, S. (2013). *Tourism Demand Models: Concepts and Theories*. In C. A., Tisdell (Eds.), *Handbook of Tourism Economics: Analysis, New applications and Case studies*. (pp. 33-66). Hackensack, NJ: World Scientific Publishing Co.Pte.Ltd.

Deckert, J., Myagkov, M., & Ordeshook, P. (2011). *Benford's Law and the Detection of Election Fraud*. *Political Analysis*, 19(3), 245-268.

De Brito F. S. (2012). *Práticas de Responsabilidade Social no Sector do Turismo em Cabo Verde: O Caso da Ilha do Sal*. Disponível em <http://www.portaldoconhecimento.gov.cv/bitstream/10961/2294/1/Pr%c3%a1ticas%20de%20R.S%20no%20Sector%20do%20Turismo%20de%20fernanda%20de%20Brito.pdf>

De Barros, J.M.V. (2007). *Impacto do Turismo no Desenvolvimento Socioeconómico: o caso da ilha do Sal*. Disponível em <https://ria.ua.pt/bitstream/10773/4675/1/213443.pdf>

De Faria H.P. N. (2017). *Previsão de séries temporais utilizando uma combinação não-linear entre ARIMA e Redes Neurais Artificiais*. Disponível em <https://www.researchgate.net/publication/319099137> *Previsao de series temporais u tilizando uma combinacao nao-linear entre ARIMA e Redes Neurais Artificiais*

De Lacalle, J., & Boshnakov, G. (2020). *Unit Root Tests for Seasonal Time Series*. Type: Package, version 2.1-2. Disponível em <https://cran.r-project.org/web/packages/uroot/uroot.pdf>

De Jesus, C. G., & Da Silva, R. D. (2010). *Economia e Turismo*. Volume 1 / CI -Rio de Janeiro: Fundação CECIERJ.

Dias, R.. *Introdução ao Turismo*. São Paulo: Atlas, 2005.

Ehlers, R.S. (2005). *Análise de Séries Temporais*. Departamento de Estatística, UFPR. Disponível em <http://www.each.usp.br/rvicente/AnaliseDeSeriesTemporais.pdf>

Expresso das Ilhas (2014). *As lições das Seychelles*. Por Jorge Montezinhos. Consultado em 29 mar. 2020. Disponível em <https://expressodasilhas.cv/economia/2014/07/15/as-licoes-das-seychelles/42537>

Expressos das ilhas (2018). *O Desenvolvimento das Maurícias como exemplo para Cabo Verde*. Por Jorge Montezinhos. Consultado em 29 mar. 2020. Disponível em <https://expressodasilhas.cv/economia/2018/01/05/o-desenvolvimento-das-mauricias-como-exemplo-para-cabo-verde/56015>

Electra Norte (2017). *Relatório e Contas 2017*. Disponível em <https://electra.cv/frontend/web/index.php?r=site%2Frelatorio&type=2&inst=3>

Electra Norte (2018). *Relatório e Contas 2018*. Disponível em <https://electra.cv/frontend/web/index.php?r=site%2Frelatorio&type=2&inst=3>

Eurostat (2018a). Europeans like to spend their holidays in Europe. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/EDN-20181107-1>

Eurostat (2018b). Coastal regions: popular tourist destinations. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/EDN-20180927-1>

Eurostat (2018c). Who spends the most on their holidays? Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/DDN-20180820-1>

Eurostat (2018d). Tourism from China to the EU more than tripled. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/DDN-20180118-1>

Eurostat (2018e). Tourism in the EU. Number of nights spent in the EU up in 2017. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/documents/2995521/8609597/4-24012018-CP-EN.pdf/59bebc89-d5e5-4276-8402-028fab29d9a2>

Eurostat (2017a). Trouble finding holiday accommodation?. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/DDN-20170821-1>

Eurostat (2017b). Who spends the most on their holidays?. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/DDN-20170814-1>

Eurostat (2017c). Europeans travel mostly by car. Consultado em 12 mar. 2020. Disponível em <https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/DDN-20170630-1>

Etuk, E.H. (2012) A Multiplicative Seasonal ARIMA Model for Nigerian Unemployment rates. *The Bulletin of Society for Mathematical Services and Standards*, ISSN: 2277-8020, Vol.3, pp 46-53. DOI:10.18052/www.scipress.com/BSMaSS.3.46

Fernandes, P. O., Monte, A. P., & Teixeira J. P. (2009). *Previsão da Procura Turística utilizando um modelo não linear*. XII Congreso Internacional de Investigación en Ciencias Administrativas (ACACIA). Mesa de Trabajo: Finanzas y Economía.

Formica, S. & Uysal, M. (2006). "Destination Attractiveness Based on Supply and Demand Evaluations: An Analytical Framework", *Journal of Travel Research*, Vol. 44, Nº 4, pp. 418-430.

Fritsch, S., Guenther, F., Wright, M., Suling, M., & Mueller S. (2019). *Training of Neural Networks*. Package 'neuralnet'CRAN, version 1.44.2. Disponível em <https://github.com/bips-hb/neuralnet>

Fewster, R.M. (2009). A Simple Explanation of Benford's Law. *The American Statistician*, 63(1), 26-32.

Gardner E.S. (2006). Exponential Smoothing: The State of the Art - Part II. *International Journal of Forecasting*, 22(4), pp. 637–666.

Gauvrit, N., Houillon, J.C., & Delahaye, J.P. (2017). Generalized Benford's law as a lie detector. *Advances in Cognitive Psychology*, 13(2), 121-127.

Gabaix, X. (2016). Power Laws in Economics: An Introduction. *Journal of Economic Perspectives*, 30(1), 185-206.

Gunnel, S., Tödter, K. (2007). *Does Benford's Law Hold in Economic Research and Forecasting?*. Bundesbank Series 1 Discussion Paper No. 32.

GOPEDS (2019). *Grandes Opções do Plano Estratégico de Desenvolvimento do Turismo para 2018-2030*. Disponível em <http://www.caboverde-info.com/content/download/9926/85352/version/1/file/BO+N.%C2%BA+2%2C+I+S%C3%A9rie%2C+de+09-01-2019.pdf>

Greiner, L. E. (1998). *Evolution and revolution as organizations grow*. *Harvard Business Review*, may-jun, pp. 55-67.

Goh C., & Law, R. (2011). *The Methodological Progress of Tourism Demand Forecasting: A Review of Related Literature*, *Journal of Travel & Tourism Marketing*, 28:3, 296-317. Disponível em DOI: [10.1080/10548408.2011.562856](https://doi.org/10.1080/10548408.2011.562856)

Hovinen, R. (2002). *Revisiting the destination lifecycle model*. *Annals of Tourism Research*, 29 (1), 209-230.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*, Springer-Verlag, Berlin. Disponível em <https://link.springer.com/book/10.1007%2F978-3-540-71918-2>

Hyndman, R. J., & Koehler, A. B. (2005). *Another look at measures of forecast accuracy*. Disponível em <https://robjhyndman.com/papers/mase.pdf>

Hyndamn, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. 2. Ed. Melbourne: OTexts, 2018. Disponível em <https://otexts.org/fpp2>

Holt, C.E. (1957). *Forecasting Trends and Seasonal by Exponentially Weighted Averages*. ONR Memorandum No. 52, Carnegie Institute of Technology, Pittsburgh, USA.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., Team, R Core, Ihaka, R., Reid, D., Shaub, D., Tang, Y., & Zhou, Z. (2021). *Forecasting Function for Time Series and Linear Model*. Version 8.15. Disponível em <https://cran.r-project.org/web/packages/forecast/index.html>

INE: Instituto Nacional de Estatística (2013). *Projeções Demográficas de Cabo Verde 2010-2030*. Disponível em <http://ine.cv>

INE (2015a). *Inquérito Multi-Objetivo Contínuo-2014. Estatísticas das Migrações*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>

INE (2015b). *Anuário Estatístico Cabo Verde 2015*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>

INE (2016). *Anuário Estatístico Cabo Verde 2016*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>

- INE (2017a). *Estatísticas do Turismo - Movimentação dos Hóspedes 2016/2017*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2017b). *III Inquérito às despesas e receitas familiares – IDERF 2015*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2018a). *Inquérito aos Gastos e Satisfação dos Turistas 2017*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2018b). *Estatísticas das Famílias e Condições de Vida – Inquérito Multi-Objetivo Contínuo 2017*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2018c). *Brochura. 15 de Setembro, Dia do Município Sal. Dados Estatísticos – 2016/2017*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019a). *Estatísticas das Famílias e Condições de Vida – Inquérito Multi-Objetivo Contínuo 2018*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019b). *Quadros do PIB por Ilhas 2016. Cedido em 16 de Maio de 2019*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019c). *Estatísticas do Turismo, Inventário Anual dos Estabelecimentos Hoteleiros – 2018*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019d). *Estatísticas do Turismo, Movimentação de Hóspedes 2018*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019e). *Inquérito aos Gastos e Satisfação dos Turistas 2018*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019f). *Estatísticas da Imigração. Pobreza e Condições de Vida 2015 e 2018*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019g). *Produto Interno Bruto por Ilha, Ano 2017 – Informação à Comunicação Social, 20 Novembro de 2019*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- INE (2019h). *V Recenseamento Empresarial 2017*. Instituto Nacional de Estatística. Disponível em <http://ine.cv>
- IPK International (2018). *ITP World Travel Trends Report 2017/2018. Boom or Bust? Where is Tourism Heading*. Consultado em 11 mar. 2020. Disponível em https://www.itb-kongress.de/media/itb/itb_dl_all/itb_presse_all/ITB_WTTR_A4_2018_interaktiv.pdf
- Kopczewski, T., & Okhrimenko, I. (2019). Playing with Benford's Law. *e-mentor*, 80(3), pp. 34-44.

- Kourentzes, N. (2019). Time Series Forecasting with Neural Networks. Package “nnfor” CRAN, version 0.9.6. Disponível em <http://kourentzes.com/forecasting/2019/01/16/tutorial-for-the-nnfor-r-package/>
- Kao, J.J., & Huang, S.S. (2000). *Forecasting using Neural Networks versus Box-Jenkins Methodology for ambiente air quality monitoring data*. J. Air & Waste Manage. Assoc. 50(2), pp. 219-226.
- Kirchgassner, G., & Wolters, J. (2007). *Introduction to Modern Time Series Analysis*. Berlin: Springer. eBook.
- Kim, J. H., & Moosa, I. (2001). Seasonal Behaviour of Monthly International Tourist Flows: Specification and Implications for Forecasting Models. *Tourism Economics*, 7, pp.381-396.
- Larrañaga, P., Inza, I., & Moujahid A. (2015). Tema 8: Redes Neuronales. Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidade del País Vasco. Consultado em 23 mai. 2018. Disponível em https://www.researchgate.net/profile/Pedro_Larranaga2/publication/268291232_Tema_8_Red_Neuronales/links/55b7b5c408ae9289a08c0c68.pdf
- La Torre, G. M. V., Naranjo, L. M. P., & Cárdenas, R. M. (2012). Etapas del ciclo de vida en el desarrollo del turismo religioso: una comparación de estudios de caso. *Cuadernos de Turismo*, Murcia, 0(30), pp. 241-266.
- Lage, B. H.G., & Milone, C. (2009). *Economia do Turismo*, (7ª ed.). Atlas, 05/2009. Vital-Book file.
- Lewis, C.D. (1982). *Industrial and Business Forecasting Method*. Butterworth Scientific. London.
- Lemos, F. (2006). *Metodologia para Seleção de Métodos de Previsão de Demanda*. (Dissertação de Mestrado no Programa de Pós-Graduação em Engenharia de Produção). Escola de Engenharia da Universidade Federal do Rio Grande do Sul, Porto Alegre.
- Lim, C. (1997). Review of international tourism demand models. *Annals of Tourism Research*, 24(4), 835-849.
- Maddala, G.S., & Lahiri, K. (2009). *Introduction to Econometrics*. (4ª ed.), John Wiley & Sons Ltd, UK.
- Maddala, G.S. (1992). *Introduction to Econometrics*. (2ª ed.), MacMillan Publishing Company, New York.
- Morettin, P. A., & Toloi, C. M. (1981). *Modelos para Previsão de Séries Temporais*, Edgard Blucher.
- Morettin, P. A., & Toloi, C. M. (2006). *Análise de Séries Temporais*. São Paulo: Edgard Blucher, 2ª ed.

Matakovic, H. (2021). Tourism Data in Croatia Assessed by Benford's Law. *Journal of Public Administration, Finance and Law*, 19, pp.166-184.

Meyer, D. (2021). *Holt-Winters Filtering*. R Documentation. Consultado em 16 mar. 2022. Disponível em <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/HoltWinters.html>

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16, pp. 451–476. Disponível em <https://www.sciencedirect.com/science/article/abs/pii/S0169207000000571?via%3Dihub>

Maroco, J. (2003). *Análise Estatística com Utilização do SPSS*. (2ª ed.). Revista e corrigida. Lisboa. ISBN: 972-618-331-6

Middleton, V.T. (2002). *Marketing de Turismo: teoria e prática*. (3ª ed.). Editora Campus; Rio de Janeiro

Murteira, B., Muller, D., & Turkman, F. (1993). *Análise de sucessões cronológicas*. McGraw-Hill, Portugal.

Mishra P.K., Rout H.B., & Pradhan B.B. (2018). Seasonality in Tourism and Forecasting Foreign Tourist Arrivals in India. *Iranian Journal of Management Studies (IJMS)*, 11(4), pp. 629-658.

Muller, D. (2007). *Processos Estocásticos e Aplicações*. II Série, Nº 3 Coleção Económicas, Edições Almeida, SA.

MIOTH (2019). *Perfil do Setor de Habitação de Cabo Verde*. Consultado em 24 jun. 2019. Disponível em <https://unhabitat.org/books/perfil-do-setor-de-habitacao-cabo-verde/>

MF (2019). *Master Plan do Turismo 2020-2030 para a Ilha do Sal*. Ministério das Finanças de Cabo Verde.

MF (2018). *Estudo de Avaliação da Dinâmica Social e Ambiental do Sector Turístico de Cabo Verde*. Consultado em 23 mar. 2020. Ministério das Finanças de Cabo Verde. Disponível em <https://webcache.googleusercontent.com/search?q=cache:EbVcoAzLYoJ:https://www.governo.cv/documentos/estudo-de-avaliacao-ambiental-e-social-estrategica-do-setor-do-turismo-em-cabo-verde-pt/+&cd=1&hl=pt-PT&ct=clnk&gl=cv>

MF (2017). *Plano Estratégico de Desenvolvimento Sustentável (PEDS) 2017-2021*. Ministério das Finanças de Cabo Verde. Disponível em <https://peds.gov.cv/sites/default/files/2018-10/PEDS%202017-2021%20-%20Vers%C3%A3o%20Final.pdf>

MSSS (2018a). *Relatório Estatístico 2018*. Ministério da Saúde e Solidariedade Social de Cabo Verde. Disponível em <https://www.minsaude.gov.cv/>

MSSS (2018b). *Programa Nacional Integrado de Luta Contra Drogas e Crimes Conexos 2018-2023 (Draft 0)*. Ministério da Saúde e Solidariedade Social de Cabo Verde.

Matias, A. (2007). *Economia do turismo: teoria e prática*. Lisboa: Instituto Piaget.

Moreno, M.J.J. (2002). *Redes Neuronales Artificiales aplicadas al Análisis de Datos* (Tese de Doutoramento). Universitat de les Illes Balears. Disponível em http://dspace.uib.es/xmlui/bitstream/handle/11201/2511/Montano_Moreno_JuanJose.pdf?sequence=1&isAllowed=y

Matich, J.D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universida Tecnológica Nacional. Consultado em 23 mai. 2018. Disponível em <ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf>

Miguel, F.L.M. (2011). *Utilização de Redes Neuronais para Previsão de longo prazo de tráfego internet a partir de informações de fluxos de dados* (Dissertação de Mestrado). Pontificia Universidade Católica do Paraná. Disponível em <https://www.ppgia.pucpr.br/pt/arquivos/mestrado/dissertacoes/2011/marcio-miguel-v7-9.pdf>

Machado, M.N.T. (2010). *Modelação da Procura Turística em Portugal: Regressão Linear versus Redes Neuronais Artificiais* (Dissertação de Mestrado). Consultado em 23 mai. 2018. Disponível em https://bibliotecadigital.ipb.pt/bitstream/10198/1822/3/INVTUR2010_final.pdf

Manning, R. E., & Powers, L. A. (1984). Peak and off-peak use: redistributing the outdoor recreation/tourism load. *Journal of Travel Research*, 23(2), 25-31. Disponível em <https://doi.org/10.1177/004728758402300204>

MECC (2014). *Plano Estratégico para o Desenvolvimento do Turismo em Cabo Verde (2010-2013)*. Consultado em 11 mar. 2020. Disponível em https://www.academia.edu/36449932/Plano_estrategico_do_turismo

Nelwamond, T. (2009). *Tourism development through strategic planning for non-metropolitan small to medium size accommodation facilities in limpopo province, south Africa* (Tese de Doutoramento)- University of Pretoria. Disponível em <https://repository.up.ac.za/handle/2263/24684>

Neves G.A., Nunes C.S., & Fernandes P.O. (2021). Challenges and Opportunities for Island Tourist Destinations: The Case of the Island of Sal, Cape Verde. In: Abreu A., Liberato D., González E.A., Garcia Ojeda J.C. (eds) *Advances in Tourism, Technology and Systems*. ICOTTS 2020. Smart Innovation, Systems and Technologies, vol 209 (pp. 498-508). Springer, Singapore. https://doi.org/10.1007/978-981-33-4260-6_42 (ISBN 978-981-33-4260-6) <http://hdl.handle.net/10400.2/11413>

Neves G.A., Nunes C.S., & Fernandes P.O. (2021). Seasonal Autoregressive Integrated Moving Average Time Series Model for Tourism Demand: the Case of the Island of Sal,

Cape Verde. In the 3rd International Conference on Tourism Technology & Systems - ICOTTS'2021, 4 and 5 November 2021, Universidad de Cartagena, Cartagena de Indias, Colombia. https://doi.org/10.1007/978-981-16-9701-2_2

Neves G.A., Nunes C.S., & Fernandes P.O. (2022). Application of Benford's Law to the Tourism Demand: The Case of the Island of Sal, Cape Verde. In: Pereira A.I. et al. (eds) Optimization, Learning Algorithms and Applications. OL2A 2021. Communications in Computer and Information Science, vol 1488 (pp 587-598). Springer, Cham. https://doi.org/10.1007/978-3-030-91885-9_43 <http://hdl.handle.net/10400.2/11589> (ISBN 978-3-030-91884-2)

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Math*, 4(1), 39-40.

Oliveira, A. P. (2001). Turismo e Desenvolvimento, Planejamento e Organização, Editora Atlas S.A., São Paulo), para classificar visitante: turistas e excursionistas.

Ollech, D. (2021). *Seasonality Tests. Type: Package*, version 0.15.4. Disponível em <https://cran.r-project.org/web/packages/seastests/seastests.pdf>

OMT (2019). *International tourism continues to outpace the global economy*. consultado em 10 mar. 2020. Disponível em <https://www.e-unwto.org/doi/pdf/10.18111/9789284421152>

OMT (1995). *Compilación de las estadísticas del gasto turístico*. (Manual Técnico No. 2). OMT.

ONU/OMT. (1999). *Mise à jour des Recommandations sur les Statistiques du Tourisme ONU-WTO –Série M 83* (1994).

Poon, A. (1993). *Tourism, Technology and competitive strategies*. Wallingford CAB International

Poon, S. (2005). *A practical guide to forecasting financial market volatility*, Chichester, John Wiley and Sons.

Prosser, R. (1994). *Societal Change and the Growth in Alternative Tourism*, Chichester, John Wiley and Sons.

Pearce, D. G. (1989). *Tourism Development* (pp. 109). New York: John Wiley.

Pearce, J.A. & Robinson, R.B. (1997). *Strategic Management formulation, implementation and control*. Irwin, McGraw-Hill.

Passari, A. (2003). *Exploração de dados atomizados para previsão de vendas no varejo utilizando redes neurais índice* (Tese de Mestrado). Universidade De São Paulo, Brasil.

Palmer, A., & Montano, J.J. (1999). Qué son las redes neuronales artificiales? Aplicaciones realizadas en el âmbito de las adicciones. *Adicciones*, 11 (3), 243-255.

- Pires, A.P. (2001). *Notas de Séries Temporais 2000/01*. Instituto Superior Técnico, Lisboa.
- Pommerenzenbaum R.I. (2014). *Redes Neurais Artificiais na Predição das Principais Séries do Índice IBOVESPA e Suas Aplicações em Sistemas Automatizados de Negociação* (Dissertação de Mestrado) Universidade Federal do Rio de Janeiro. Disponível em <http://pee.ufrj.br/teses/textocompleto/2014033101.pdf>
- Reis, D.F. (2014). *Avaliação de Modelos de Previsão de Vendas a Partir da Exploração de Técnicas de Análise de Séries Temporais, Métodos Causais e de Redes Neurais Artificiais* (Dissertação de Mestrado). Universidade FUMEC – Faculdade de Ciências Empresariais, Belo Horizonte.
- Reis E., Melo P., Andrade R., & Calapez T. (2016). *Estatística Aplicada - Vol.2*. 5ª ed., Edições Sílabo, Lda. ISBN: 978-972-618-841-4
- Ross, S. (2009). *Probability and Statistics for Engineers and Scientists*. 4ª ed. Elsevier Academic Press, London, UK.
- Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1987). *A general framework for parallel distributed processing*. In *Parallel distributed processing*, vol 1: Foundations, D.E. Rumelhart, J. L. McClelland and the PDP group, Ed. MIT Press, Cambridge, Massachusetts, pp. 45–76.
- Saville, A. (2014). *Using Benford's Law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange*. *South African Journal of Economic and Management Sciences*, 9(3), pp.341-354.
- Sarle, W. (1994). *Neural networks and statistical models*. In: Annual SAS Users Group International Conference, 19. Proceedings..., Cary: SAS Institute, pp. 1538-1550
- Santana, C.J. (2006). *Predicción de series temporales con redes neuronales: una aplicación a la inflación colombiana*. Disponível em <http://www.scielo.org.co/pdf/rce/v29n1/v29n1a05.pdf>
- Schwanninger, M. (1989). *Trends in Leisure and tourism for 2000 – 2010*, London, Prentice Hall.
- Schopf, C.E. (2007). *Método Neuro-estatístico para Predição de Séries Temporais Ruidosas* (Dissertação de Mestrado). Universidade de Federal do Rio Grande do Sul. Disponível em <https://www.lume.ufrgs.br/bitstream/handle/10183/11476/000614584.pdf?sequence=1>
- Song, H., Li, & G. (2008). *Tourism demand modelling and forecasting. A review of recent research*. *Tourism Management*, 29, pp. 203–220.
- Song, H., Qiu, R.T.R., & Park, J. (2019). *A review of research on tourism demand forecasting*. *Annals of Tourism Research*, 75, pp. 338-362.

Song, H., & Witt, S. F. (2000). *Tourism demand modelling and forecasting: Modern econometric approaches*. Cambridge: Pergamon.

Song, H., Gao, B. Z. & Lin, V. S. (2013). “Combining statistical and judgmental forecasts via a web-based tourism demand forecasting system”, *International Journal of Forecasting*, Vol. 29, Nº 2, pp. 295-310.

Sutcliffe, C. M., & Sinclair, M. T. (1980). The measurement of seasonality within the tourist industry: an application to tourist arrivals in Spain. *Applied Economics*, 12(4), pp.429-441. Disponível em <https://doi.org/10.1080/00036848000000004>

Swinkels R., & Longmore R. (2019). Como continuar a ser um “campeão” na redução da pobreza: Os desafios de Cabo Verde. Consultado em 11 mar. 2019. Disponível em <https://blogs.worldbank.org/pt/african/how-to-remain-a-poverty-reduction-champion-overcoming-cabo-verdes-challenges>

Shumway, H.R., & Stoffer, S. D. (2016). *Time Series Analysis and Its Applications: With R Examples*. Fourth Edition, Springer.

Souza, G.P., Samohyl, R.W., & Miranda, R.G. (2008). *Métodos Simplificados de Previsão Empresarial*. Rio de Janeiro: Editora Ciência Moderna.

Stoffer, D. (2021). *Applied Statistical Time Series Analysis*. Type:Package, version 1.14. Disponível em <https://cran.r-project.org/web/packages/astsa/astsa.pdf>

Tepán, A.C.E. (2013). *Estudio de los principales tipos de redes neuronales y las herramientas para su aplicación*. Disponível em <https://dspace.ups.edu.ec/bitstream/123456789/4098/1/UPS-CT002584.pdf>

Torman, L.B. V., Coster, R., & Ribolti, J. (2012). Normalidade de variáveis: Métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista HCPA*, 2012; 32(2), pp. 227-234. Disponível em <http://www.seer.ufrgs.br/hcpa/article/viewFile/29874/19186>

Trapletti, A., Hornik, K., & LeBaron, B. (2021). *Time Series Analysis and Computacional Finance*. Type:Package, version 0.10-49. Disponível em <https://cran.r-project.org/web/packages/tseries/tseries.pdf>

UNWTO (2008a). *Handbook on Tourism Forecasting Methodologies*. Madrid: UNWTO.

UNWTO (2001). *Tourism 2020 Vision*. Volume 7 – Global Forecasts and Profites of Market Segments. Madrid: World Tourism Organization.

UNWTO (2018). *European Union Tourism Trends European*. UNWTO, Madrid, DOI: <https://doi.org/10.18111/9789284419470>

Uysal, M. (1998). The determinants of tourism demand: A theoretical perspective In D. Ioannides & K. G. Debbage (Eds.). *The economic geography of the tourist industry: A supply-side analysis* (pp. 79-95). New York: Routledge

- Webel, K., & Ollech, D. (2019). *An overall seasonality test. Deutsche Bundesbank's Discussion Paper series*. Disponível em <https://cran.r-project.org/web/packages/seastests/seastests.pdf>
- Witt, S.F., Witt, C.A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11, pp.447-475, SSDL 0169-2070(95)00591-9
- Wickham, H., Chang, W., Henry, L., Pedersen, T., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2021). *Create Elegant Data Visualisations Using the Grammar of Graphics*. Type:Package, version 3.3.5. Disponível em <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>
- Wu D., Song H., & Shen S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, 29(1), pp. 507-529.
- Wei, W.S.W. (2006). *Time Series Analysis: Univariate and Multivariate Methods*, (2ª ed.). ISBN 0-321-32216-9
- Wheelwright, S.C., Makridakis, S., & McGee, V.E. (1983). *Forecasting: Methods and Applications*, (2ª ed.). New York: John Wiley and Sons.
- Williams, A. M., & Shaw, G. (1991). *Tourism and Economic Development-Western European Experiences* (2ª ed.). Chichester: Wiley & Sons.
- Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), pp.324-342. Disponível em <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.6.3.324>
- Xavier, N.M.J. (2016). *Análise e Previsão de Séries Temporais com Modelos ARIMA e Análise Espectral Singular* (Dissertação de Mestrado). Universidade Aberta.
- Yacoumis, J. (1980). Tackling seasonality-The case of Sri Lanka. *International Journal of Tourism Management*, 1(2), pp.84-98. Disponível em [https://doi.org/10.1016/0143-2516\(80\)90031-6](https://doi.org/10.1016/0143-2516(80)90031-6)
- Zar, J.H. (1999). *Biostatistical analysis*. Upper Saddle River, N.J.: Prentice Hall
- Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, Elsevier Science B.V., 14, pp. 35-62.
- Zuben J.V.F, & Boccatto L. (2017). *Rede MLP: Perceptron de Múltiplas Camadas*. IA353, DCA/FEEC/Unicamp. Disponível em http://www.dca.fee.unicamp.br/~lboccatto/topico_4.1_redes_MLP.pdf

ANEXO I: CENÁRIOS DO CRESCIMENTO DO FLUXO TURÍSTICO

Tabela:A1.1: Cenário 1 – Muito Otimista (com crescimento anual de 15% no fluxo turístico).

Cenário Muito Otimista	Unid. Med.	2017	2021	2025	2030
Entradas turistas	Unidades	716.775	1.253.644	2.192.631	5.071.689
Dormidas	Unidades	4.597.477	8.041.016	14.063.787	32.530.395
Nº de quartos	Unidades	12.463	21.798	38.125	88.185
Nº de camas	Unidades	20.421	35.716	62.468	144.493
Empregos diretos	Unidades	8.825	15.435	26.996	62.443
Empregos indiretos	Unidades	26.475	46.305	80.988	187.329
Empregos total	Unidades	35.300	61.740	107.984	249.772
Receitas taxa turística	Euros (€)	9.194.954	16.082.032	28.127.575	65.060.789

Fonte: GOPEDS-Turismo, 2019

Tabela:A1.2: Cenário 2 – Otimista (com crescimento anual de 12% no fluxo turístico).

Cenário Otimista Moderado	Unid. Med.	2017	2021	2025	2030
Entradas	Unidades	716.775	1.12.789	1.774.709	3.502.960
Dormidas	Unidades	4.597.477	7.234.219	11.383.184	22.468.386
Nº de quartos	Unidades	12.463	19.611	30.858	60.908
Nº de camas	Unidades	20.421	32.133	50.562	99.800
Empregos diretos	Unidades	8.825	13.886	21.850	43.129
Empregos indiretos	Unidades	26.475	41.659	65.551	129.386
Empregos total	Unidades	35.300	55.545	87.402	172.515
Receitas taxa turística	Euros (€)	9.194.954	14.468.438	22.766.368	44.936.773

Fonte: GOPEDS-Turismo, 2019

Tabela:A1.3: Cenário 3 – Médio (com crescimento anual de 9% no fluxo turístico).

Cenário Médio	Unid. Med.	2017	2021	2025	2030
Entradas	Unidades	716.775	1.011.786	1.428.219	2.395.266
Dormidas	Unidades	4.597.477	6.489.714	9.160.761	15.363.513
Nº de quartos	Unidades	12.463	17.593	24.833	41.648
Nº de camas	Unidades	20.421	28.826	40.690	68.241
Empregos diretos	Unidades	8.825	12.457	17.584	29.491
Empregos indiretos	Unidades	26.475	37.372	52.753	88.472
Empregos total	Unidades	35.300	49.829	70.337	117.963
Receitas taxa turística	Euros (€)	9.194.954	12.979.428	18.321.522	30.727.026

Fonte: GOPEDS-Turismo, 2019

Tabela:A1.4: Cenário 4 – Pessimista (com crescimento anual de 5% no fluxo turístico).

Cenário Pessimista	Unid. Med.	2017	2021	2025	2030
Entradas	Unidades	716.775	871.244	1.059.003	1.419.165
Dormidas	Unidades	4.597.477	5.588.262	6.792.567	9.102.690
Nº de quartos	Unidades	12.463	15.149	18.414	24.676
Nº de camas	Unidades	20.421	24.822	30.171	40.432
Empregos diretos	Unidades	8.825	10.727	13.039	17.473
Empregos indiretos	Unidades	26.475	32.181	39.116	52.419
Empregos total	Unidades	35.300	42.907	52.154	69.892
Receitas taxa turística	Euros (€)	9.194.954	11.176.524	13.585.135	18.205.380

Fonte: GOPEDS-Turismo, 2019

**ANEXO II: VANTAGENS E DESVANTAGENS DA APLICAÇÃO DE DIFERENTES
MODELOS DE PREVISÃO DA PROCURA TURÍSTICA**

Tabela:A2.1: Comparação entre diferentes modelos de previsão da procura turística.

Grupos	Sub grupos	Modelos	Vantagens	Desvantagens
Séries temporais		NAÏVE 1 E 2	Em alguns estudos, geralmente supera outros modelos de séries temporais. Para horizontes curtos de previsão a Naïve 1 tem tido melhor desempenho. Podem ser utilizados como <i>benchmarks</i> na avaliação da precisão das previsões.	A sua precisão, quando comparado com outros modelos de séries temporais, pode depender do tipo de Dados, como, por exemplo, o uso de dados sazonais ao invés de anuais.
		Sazonal- NAÏVE	-----	-----
		SE HOLT-WINTERS	Se mostrou superior ao SE	-----
		AR	Tem tido geralmente um bom desempenho nas previsões, independentemente da frequência dos dados. Produz melhor desempenho em horizontes de dois anos.	-----
		AR sazonal	-----	-----
		MA	-----	Pobre a nível do desempenho das previsões.
		ARIMA	Tem sido considerado como superior em termos de previsão turística. Principalmente quando é avaliado pelo menos um destino. Muito flexíveis e bastante aplicado.	Desempenho duvidoso quando comparado com o modelo SARIMA. <i>O ARIMA multivariado não supera o ARIMA univariado.</i>
		SARIMA	Tem tido bom desempenho. Tem-se tornado popular devido ao componente sazonal que é do interesse de decisores.	-----
		SARIMA-In	-----	-----
		SARIMAX	Mostrou ser superior ao modelo padrão SARIMA e o SE Holt-Winters.	-----

		ARIMAX	Tem mostrado superior em previsões de longo prazo, ao menos comparando com o SARIMA. Quando combinado (ex. com o VP, com o MIDAS) traz bons resultados.	-----
		MARIMA	Considerado superior aos demais modelos na presença de eventos externos ao destino.	-----
		ARFIMA	Se mostrou superior ao SARIMA e ao ARARMA.	-----
		REG-ARIMA	Mostrou ser superior ao Naïve e ARIMA.	Pouco utilizado.
		ARARMA	-----	-----
		TRIGONOMÉTRICO	-----	Incapaz de controlar a sazonalidade estocástica e os ciclos indeterministas.
		GARCH	Muito aplicado devido a sua capacidade de capturar a volatilidade e a existência de heterocedasticidade.	-----
		ARIMA-ARCH	-----	-----
		SSA	Tem mostrado ser superior ao SE, SARIMA e STSM	-----
		STSM	Resulta em previsões melhores, quando é acrescentado ao modelo os chamados indicadores de sentimento de negócios.	Permite que cada equação lide com seu próprio termo de erro.
		M-STSM	-----	-----
		ISS	Tem demonstrado grande desempenho em vários estudos.	-----
		ARAR	-----	-----
		DECOMPOSIÇÃO	-----	-----
		MUDANÇA PROPORCIONAL	-----	-----
		GOMPERTZ	-----	-----
		FRACIONAL SAZONAL	-----	-----

		QUEBRA DE TENDÊNCIA SEQUENCIAL	----	----
		ESPAÇO DE ESTADOS	----	----
		POLINOMIAL CÚBICO	----	----
Modelos Combinados				
		STSM + TVP	Mostrou superior a outros modelos tanto de séries temporais como econométricos	----
		SARIMA + REGRESSÃO DE ONDA SENOIDAL	----	Resultados não indicam melhorias na previsão comparando com os modelos não combinados
Em geral não se mostraram superiores aos modelos não combinados.				
Modelos não lineares				
		LIMIAR AUTO-EXCITANTE AUTOREGRESSIVO	----	----
		COMUTAÇÃO DE MARKOV	Se mostrou superior ao anterior e ao modelo AR.	----
Econométricos				
Econométricos		SR	Foi muito utilizado nos primórdios e reutilizado recentemente.	Problemas obtidos devido a ausência de estacionaridade.
		DL	Tem sido muito utilizado.	Muito limitado na previsão turística.
		ECM	Supera problemas de regressão espúrios. Captura dinâmicas de curto prazo. Tem revelado com um extraordinário desempenho. Enfático em medir a relação de causa e efeito entre os fatores que influenciam a procura turística. Supera alguns modelos como por exemplo o SARIMA e o Naïve.	----
		TVP	Captura instabilidades estruturais.	-----

			Quando se trata de previsões a um passo a frente tem um desempenho melhor que, por ex. o ECM, VAR e ADLM. Superando ainda modelos de séries temporais (ARIMA e Naïve). É o 2º modelo econométrico mais preciso em se tratando de previsões a longo prazo.	
		MIDAS	Tende a fornecer melhores precisões de previsão.	-----
		Regressão	-----	Resultados espúrios se não considerar variáveis não estacionárias.
	Modelos estático de multi-equações	SEM	Grande capacidade de representar as relações causais entre as variáveis.	Seu desempenho não tem sido analisado
		AIDS	Mostrou ter bom desempenho quando inserido em sistemas dinâmicos. Tende a ser superior ao VAR padrão.	Não se mostrou muito eficiente quando inserido em sistemas estático.
	Modelos dinâmicos de multi-equações	ADLM	Tem revelado com um extraordinário desempenho. Enfático em medir a relação de causa e efeito entre os fatores que influenciam a procura turística.	-----
		GETS	-----	-----
		VAR	Ignora a necessidade de modelagem estrutural como é o caso do modelo SEM. Não impõe restrições a priori nas séries temporais inter-relacionadas	Mostrou-se inconclusivo quanto ao seu desempenho. Paira alguma incerteza nos resultados produzidos, mesmo quando supera outros modelos.
		VEC	São capazes de capturar a interdependência de	-----

			várias series temporais.	
		BVAR	Produziu melhores resultados comparando com o VAR.	-----
		GVAR		-----
		BGVAR		-----
		EC-LAID		-----
		TVP-EC-AIDS	Supera seu equivalente de parâmetro fixo.	-----
		TVP-LR-AIDS	Supera seu equivalente de parâmetro fixo.	-----
Modelos combinados				
		ECM + TVP	Revelou-se mais preciso dentre os demais. Bom desempenho em previsões a curto prazo.	Ainda é pouco utilizado.
		TVP + ADLM	Funcionam muito bem.	-----
		TVP + AIDS	Funcionam muito bem.	-----
Modelo não-linear				
		STR	Muito vantajoso em termos técnicos.	Encontrado em apenas 1 estudo. Sem menção de seu desempenho.
Dados Agrupados			As aplicações foram vistas como vantajosas e sem qualquer obstáculo apresentado.	Pode violar a suposição de variação de erro constante.
Dados em Painéis		ANÁLISE DE DADOS EM PAINÉIS	As aplicações foram vistas como vantajosas e sem qualquer obstáculo apresentado. Reduz a multicolinearidade e aumenta o grau de liberdade no modelo. É muito mais eficiente quando a frequência dos dados das séries temporais são curtos.	Pode não ser válido se não for levado em consideração as condições necessárias para a escolha de um modelo.
		PDR	Pode ser um recurso integrado a análise de dados em painéis	-----

Inteligência Artificial	Métodos causais baseados em IA	REDE NEURONAL FEED-FORWARD	Superou os modelos econométricos mais em uso, bem como os modelos de regressão múltipla, Naïve 1, MA e ES	-----
		RETROPROPAGAÇÃO	Se mostrou superior a outras abordagens	-----
		RBF	-----	-----
		GRNN	-----	-----
		Perceptron	É a mais utilizada.	Seu processo de aprendizagem é muito longo.
	Métodos das séries temporais baseados em IA	REDE NEURONAL FEED-FORWARD	Tem demonstrado com melhor desempenho comparando com outros modelos de séries temporais.	-----
		ELMAN	Seu desempenho é visto como superior aos demais.	-----
		SÉRIES TEMPORAIS NEBULOSAS	São mais vantajosos para séries temporais com dimensões curtas.	Não demonstrou ser superior a Teoria Cinzenta.
		TEORIA CINZENTA	Vantajoso quando existe uma certa incerteza do modelo e na insuficiência de informação.	Não demonstrou ser superior a séries temporais nebulosas.
De uma forma geral, as redes neuronais tem superado várias técnicas de previsões incluindo as de séries temporais (e mesmo quando os dados das séries são de qualidade duvidosa).				
Modelos Combinados				
		REDE NEURONAL DE RETROPROPAGAÇÃO + MODELOS DE SÉRIES TEMPORAIS	-----	Não resultou num bom desempenho do modelo da rede.
		SVR + NAÏVE	-----	-----
		SVR + SE	-----	-----
		SVR + ARIMA	-----	-----
		GFS	Mostrou ter bom desempenho.	-----
Modelos Não-Lineares				
		SVR	-----	-----
		GA-SVR	-----	-----
		SVM	Superou os modelos de redes neuronais, o ARIMA e o SARIMA.	-----
		GPR	-----	-----

		GPR -esparso	É visto como melhor com relação ao GPR padrão, quanto a complexidade computacional.	-----
Outras abordagens quantitativas		CONJUNTOS APROXIMADOS	Mostrou-se encorajadora quanto aos resultados produzidos, tanto com o uso de variáveis independentes socioeconómica como não económicas. Pode complementar os modelos econométricos na análise da procura numa perspectiva micro.	-----
Métodos de julgamento		DELPHI	Trazem resultados encorajadores quando integrados com as técnicas estatísticas.	Seu desempenho é questionável devido a subjetividade das opiniões envolvido num debate de previsão.
De uma forma geral, os métodos de julgamento trazem melhores resultados quando integrados com métodos quantitativos.				

Fonte: Elaboração própria

**ANEXO III: NÚMERO DE REGISTOS MENSIS DE DORMIDAS NOS ESTABELECIMENTOS
TURÍSTICOS NA ILHA DO SAL NO PERÍODO [2000:01 -2018:12]**

Tabela:A3.1: Número de Registos Mensais de Dormidas nos Estabelecimentos Turísticos na Ilha do Sal no Período [2000:01-2018:12].

Ano	Meses												Total
	01	02	03	04	05	06	07	08	09	10	11	12	
2000	39.922	43.318	42.915	35.405	26.436	23.331	31.187	48.101	33.393	33.274	31.144	35.850	424.276
2001	54.739	54.243	57.621	53.946	34.270	35.893	37.967	61.574	41.054	33.305	44.007	40.282	548.901
2002	43.475	47.852	50.716	41.622	31.606	29.952	37.752	52.602	28.679	36.154	39.775	38.200	478.385
2003	76.066	71.431	59.082	73.671	50.992	42.512	52.307	77.439	52.478	42.879	40.342	46.312	685.511
2004	59.823	48.949	56.204	55.241	49.687	47.462	53.659	81.435	61.035	54.653	60.187	56.863	685.198
2005	67.218	55.353	62.544	50.586	36.578	50.323	53.355	77.364	54.350	48.221	76.610	77.480	709.982
2006	81.181	76.419	80.718	94.386	59.979	73.422	78.073	94.167	60.510	82.888	79.553	84.125	945.421
2007	87.408	83.494	96.368	83.681	59.295	65.312	75.100	93.974	95.706	108.642	130.322	122.340	1 101.642
2008	145.022	140.073	144.959	103.156	89.902	85.462	102.912	129.286	90.889	88.494	113.392	113.529	1 347.076
2009	108.226	101.263	105.014	90.069	71.362	74.777	82.424	95.145	86.053	79.853	83.930	95.184	1 073.300
2010	104.440	93.575	106.354	83.054	75.022	73.158	93.555	103.653	77.147	92.095	99.716	102.235	1 104.004
2011	116.879	109.266	124.106	105.980	64.291	67.778	92.090	103.965	88.646	107.536	118.280	115.249	1 214.066
2012	136.565	124.112	126.403	105.639	87.333	91.836	101.751	125.410	109.590	120.029	135.459	142.418	1 406.544
2013	147.746	135.967	146.822	120.591	83.032	85.033	101.882	114.924	104.153	149.004	147.795	145.033	1 481.980
2014	164.238	148.293	155.164	120.828	87.399	87.914	97.299	111.508	105.286	176.098	162.936	160.443	1 577.406
2015	168.420	200.288	165.904	140.721	96.388	109.098	134.288	150.369	129.895	150.165	184.360	193.707	1 823.603
2016	204.661	205.643	215.444	180.732	129.660	136.703	147.332	144.155	115.291	143.846	210.489	228.078	2 062.034
2017	249.162	214.312	205.295	207.714	143.093	154.921	192.532	212.943	179.214	230.459	273.066	256.776	2 519.487
2018	282.466	246.397	254.670	231.471	175.247	187.246	223.923	238.628	196.074	208.881	265.270	255.454	2 765.727

Fonte: Elaboração própria, com dados disponibilizados pelo INE

ANEXO IV: RESULTADOS DA APLICAÇÃO DO COMANDO *AUTO.ARIMA* ()

> **auto.arima(train,d=1,D=1,stepwise=FALSE,approximation=FALSE,trace=TRUE)**

```

ARIMA(0,1,0)(0,1,0)[12]      : -138.0656
ARIMA(0,1,0)(0,1,1)[12]      : -225.0894
ARIMA(0,1,0)(0,1,2)[12]      : -224.3582
ARIMA(0,1,0)(1,1,0)[12]      : -189.9168
ARIMA(0,1,0)(1,1,1)[12]      : -224.0548
ARIMA(0,1,0)(1,1,2)[12]      : Inf
ARIMA(0,1,0)(2,1,0)[12]      : -214.7123
ARIMA(0,1,0)(2,1,1)[12]      : -223.7734
ARIMA(0,1,0)(2,1,2)[12]      : -221.9133
ARIMA(0,1,1)(0,1,0)[12]      : -165.4322
ARIMA(0,1,1)(0,1,1)[12]      : -246.2383
ARIMA(0,1,1)(0,1,2)[12]      : -245.3158
ARIMA(0,1,1)(1,1,0)[12]      : -216.4462
ARIMA(0,1,1)(1,1,1)[12]      : -245.0805
ARIMA(0,1,1)(1,1,2)[12]      : -244.2937
ARIMA(0,1,1)(2,1,0)[12]      : -239.0391
ARIMA(0,1,1)(2,1,1)[12]      : -244.3508
ARIMA(0,1,1)(2,1,2)[12]      : -242.2707
ARIMA(0,1,2)(0,1,0)[12]      : -163.8946
ARIMA(0,1,2)(0,1,1)[12]      : -248.3726
ARIMA(0,1,2)(0,1,2)[12]      : -247.9303
ARIMA(0,1,2)(1,1,0)[12]      : -217.535
ARIMA(0,1,2)(1,1,1)[12]      : -247.6559
ARIMA(0,1,2)(1,1,2)[12]      : -246.4571
ARIMA(0,1,2)(2,1,0)[12]      : -240.1328
ARIMA(0,1,2)(2,1,1)[12]      : -246.5047
ARIMA(0,1,3)(0,1,0)[12]      : -167.6485
ARIMA(0,1,3)(0,1,1)[12]      : -247.4391

```

ARIMA(0,1,3)(0,1,2)[12]	: -246.5989
ARIMA(0,1,3)(1,1,0)[12]	: -218.3811
ARIMA(0,1,3)(1,1,1)[12]	: -246.3803
ARIMA(0,1,3)(2,1,0)[12]	: -239.6644
ARIMA(0,1,4)(0,1,0)[12]	: -166.5865
ARIMA(0,1,4)(0,1,1)[12]	: -247.2087
ARIMA(0,1,4)(1,1,0)[12]	: -216.826
ARIMA(0,1,5)(0,1,0)[12]	: Inf
ARIMA(1,1,0)(0,1,0)[12]	: -155.955
ARIMA(1,1,0)(0,1,1)[12]	: -237.0156
ARIMA(1,1,0)(0,1,2)[12]	: -235.9389
ARIMA(1,1,0)(1,1,0)[12]	: -204.7294
ARIMA(1,1,0)(1,1,1)[12]	: -235.6999
ARIMA(1,1,0)(1,1,2)[12]	: Inf
ARIMA(1,1,0)(2,1,0)[12]	: -228.7779
ARIMA(1,1,0)(2,1,1)[12]	: -235.4654
ARIMA(1,1,0)(2,1,2)[12]	: -233.5036
ARIMA(1,1,1)(0,1,0)[12]	: -163.6108
ARIMA(1,1,1)(0,1,1)[12]	: -249.3906
ARIMA(1,1,1)(0,1,2)[12]	: -248.5423
ARIMA(1,1,1)(1,1,0)[12]	: Inf
ARIMA(1,1,1)(1,1,1)[12]	: -248.3217
ARIMA(1,1,1)(1,1,2)[12]	: -247.1123
ARIMA(1,1,1)(2,1,0)[12]	: -239.107
ARIMA(1,1,1)(2,1,1)[12]	: -247.115
ARIMA(1,1,2)(0,1,0)[12]	: Inf
ARIMA(1,1,2)(0,1,1)[12]	: -247.6767
ARIMA(1,1,2)(0,1,2)[12]	: -246.8322
ARIMA(1,1,2)(1,1,0)[12]	: -217.4899
ARIMA(1,1,2)(1,1,1)[12]	: -246.6171

ARIMA(1,1,2)(2,1,0)[12]	: Inf
ARIMA(1,1,3)(0,1,0)[12]	: -166.1631
ARIMA(1,1,3)(0,1,1)[12]	: -245.705
ARIMA(1,1,3)(1,1,0)[12]	: -216.4261
ARIMA(1,1,4)(0,1,0)[12]	: Inf
ARIMA(2,1,0)(0,1,0)[12]	: -170.072
ARIMA(2,1,0)(0,1,1)[12]	: -249.7829
ARIMA(2,1,0)(0,1,2)[12]	: -248.8494
ARIMA(2,1,0)(1,1,0)[12]	: -221.4998
ARIMA(2,1,0)(1,1,1)[12]	: -248.6497
ARIMA(2,1,0)(1,1,2)[12]	: Inf
ARIMA(2,1,0)(2,1,0)[12]	: -242.5
ARIMA(2,1,0)(2,1,1)[12]	: -247.5024
ARIMA(2,1,1)(0,1,0)[12]	: -169.1289
ARIMA(2,1,1)(0,1,1)[12]	: -247.692
ARIMA(2,1,1)(0,1,2)[12]	: -246.7323
ARIMA(2,1,1)(1,1,0)[12]	: -219.4061
ARIMA(2,1,1)(1,1,1)[12]	: -246.5286
ARIMA(2,1,1)(2,1,0)[12]	: -240.3836
ARIMA(2,1,2)(0,1,0)[12]	: -167.0323
ARIMA(2,1,2)(0,1,1)[12]	: -246.0326
ARIMA(2,1,2)(1,1,0)[12]	: -217.3766
ARIMA(2,1,3)(0,1,0)[12]	: -180.4648
ARIMA(3,1,0)(0,1,0)[12]	: -168.9346
ARIMA(3,1,0)(0,1,1)[12]	: -247.687
ARIMA(3,1,0)(0,1,2)[12]	: -246.7295
ARIMA(3,1,0)(1,1,0)[12]	: -219.4044
ARIMA(3,1,0)(1,1,1)[12]	: -246.5274
ARIMA(3,1,0)(2,1,0)[12]	: Inf
ARIMA(3,1,1)(0,1,0)[12]	: -167.0302

ARIMA(3,1,1)(0,1,1)[12]	: Inf
ARIMA(3,1,1)(1,1,0)[12]	: -217.4495
ARIMA(3,1,2)(0,1,0)[12]	: -181.345
ARIMA(4,1,0)(0,1,0)[12]	: -167.0479
ARIMA(4,1,0)(0,1,1)[12]	: -246.2501
ARIMA(4,1,0)(1,1,0)[12]	: -217.4294
ARIMA(4,1,1)(0,1,0)[12]	: Inf
ARIMA(5,1,0)(0,1,0)[12]	: -165.0963

Best model: ARIMA(2,1,0)(0,1,1)[12]

Series: train

ARIMA(2,1,0)(0,1,1)[12]

Coefficients:

ar1	ar2	sma1
-0.3307	-0.2653	-0.6919
s.e. 0.0685	0.0675	0.0528

sigma² estimated as 0.01603: log likelihood=128.99

AIC=-249.98 AICc=-249.78 BIC=-236.73

ANEXO V: SCRIPTS

Série Temporal – Modelo SARIMA

```

#####Instalar pacotes#####
install.packages ("astsa") # Applied Statistical Time Series Analysis: acf2() e sarima()
install.packages ("forecast") # chamar as funções forecast() e autoplot()
install.packages ("ggplot2") # carregar temas para a função autoplot()
install.packages ("tseries") # teste de Dikey-fuller
install.packages ("uroot") # permite saber a significancia dos parametros
install.packages ("randtests") #permite fazer testes para verificar a tendencia
install.packages ("seastests") #permite fazer testes para verificar a sazonalidade
#####carregar pacotes instalados#####
library(astsa)
library(forecast)
library(ggplot2)
library(tseries)
library(uroot)
library(randtests)
library(seastests)
#####Importar a base de dados#####
file.choose()
[1] "C:\\Users\\Cliente\\Desktop\\TeseTurismo\\CAPITULOS\\BDados\\TSunivariado_Sal.txt"
> bdsal<-
read.table("C:\\Users\\Cliente\\Desktop\\TeseTurismo\\CAPITULOS\\BDados\\TSunivariado_Sal.txt",heade
r=TRUE)
> bdsal
#####Apresentar a tabela no formato de serie temporal#####
ts.bdsal<-ts(bdsal,start=2000,freq=12)
dados<-ts.bdsal[,2] #aqui fiz uma transformação pois não estava no formato pretendido (meses na coluna e
anos na linha)
#####Apresentar graficamente a série#####
autoplot(dados) +
  geom_point(shape = 18) + #shape é o estilo dos pontos no gráfico
  labs(x = "Anos - período 2000:2018", y = "Número de dormidas") +
  theme_bw()
#####verificar a estacionaridade#####
#1. por meio do correlograma
acf2(dados)
#2. por meio do teste de raiz unitária Dickey-fuller
adf.test(dados)
#####por meio da existencia ou não da tendência###
cox.stuart.test(dados.ts)
#####Testar se existe ou não sazonalidade na série###
isSeasonal(dados, test = "wo", freq = 12)
summary(wo(dados))
#####Usando a transformação logaritmica para estabilizar a variação#####
dadost<-log(dados)
autoplot(dadost) +
  geom_point(shape = 18) + #shape é o estilo dos pontos no gráfico
  labs(x = "Anos - período 2000:2018", y = "Número de dormidas (log)") +
  theme_bw()
#####divisão da serie em treino e teste #####

```

```

train <- window(dadost, end = c(2017, 12))
test <- window(dadost, start = 2018)
***ver quantas diferenciações são necessárias na parte não sazonal da série
ndiffs(train)
***remover a tendencia de crescimento, aplicando a primeira diferença***
dadost_d1 <- diff(train)
autoplot(dadost_d1) +
  ggtitle("Primeira diferença do log da série")
***confirmar se com esta diferença feita a serie se tornou estacionária***
adf.test(dadost_d1)
***fazer o correlograma da serie diferenciada para visualizar o que se tem***
acf2(dadost_d1)
***ver quantas diferenciações são necessárias na parte sazonal da série***
nsdiffs(dadost_d1)
***aplicar a primeira diferenciação sazonal***
dadost_sz <- diff(dadost_d1, 12)
autoplot(dadost_sz) +
  ggtitle("Primeira diferença sazonal do log da série")
***Fazer novamente o corelograma***
acf2(dadost_sz)
***Modelar***
m1<-Arima(train,order=c(2,1,0),seasonal=list(order=c(0,1,1),period=12))
***verificar a significancia dos parametros***
coefstest(m1)
***Para apresentar os indicadores de erro***
summary(m1)
***construir o gráfico dos residuos***
plot(window(resid(m1),start=c(2000,1)),ylab='Standardized Residuals',type='o')
abline(h=0)
qqnorm(window(resid(m1),start=c(2000,1)))
qqline(window(resid(m1),start=c(2000,1)))
***Confirmar se os residuos seguem a normalidade aplicando os testes***
shapiro.test(resid(m1))
ks.test(resid(m1),"pnorm")
***verificar as autocorrelações***
acf2(window(resid(m1),start=c(2000,1)), main="")
***Ajustamento do modelo com a serie de treino (train)***
plot(train,
  main="")
lines(m1$fitted,col="red")
legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
  pch = c(NA, 1), col = c("red", "black"))
***verificar a qualidade da previsão no treino***
accuracy(train,m1$fitted)
***validação do modelo para o período considerado para teste***
teste <- sarima.for(train, n.ahead = length(test),
  p = 2, d = 1, q = 0, P = 0, D = 1, Q = 1, S = 12)
***visualização gráfica***
plot.ts(teste$pred, col = "red", ylim = c(11, 13),
  ylab = "",
  main = "")

```

```

lines(test, type = "o")
legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
      pch = c(NA, 1), col = c("red", "black"))
***verificar a qualidade da validação para o período de teste (test)***
accuracy(test, teste$pred)
***verificar a validação tudo em um único gráfico
plot(forecast(m1, h = length(test)), ylab="test", xlab="time", main="")
lines(test, col = "black")
legend("topleft", legend = c("Observado", "Previsto"),
      lty = 1, col = c("black", "blue"))
***Previsão gráfica a um passo a frente***
plot(forecast(dados, h = 12), ylab="Dormidas", xlab="tempo", main="")
lines(test, col = "black")
legend("topleft", legend = c("Observado", "Previsto"),
      lty = 1, col = c("black", "blue"))
***resultado numérico da previsão a um passo a frente***
prev <- forecast(m1, h = 12)
print(summary(prev))

```

Série Temporal – Método Holt-Winters

```

***subdivisão da série em treino e teste***
train <- window(dadost, end = c(2017, 12))
test <- window(dadost, start = 2018)

***Ajustar os dados de treino***
fit.mt <- HoltWinters(x=train, seasonal="multiplicative")
plot(fit.mt)

#análise dos resíduos***
resid <- fit.mt$residuals
par(mfrow=c(2,2))
plot(resid(fit.mt))
acf(resid(fit.mt))
pacf(resid(fit.mt))
qqnorm(resid(fit.mt))
qqline(resid(fit.mt))

***aplicar testes de normalidade***
shapiro.test(resid(fit.mt))
ks.test(resid(fit.mt), "pnorm")

***fazer o histograma***
win.graph(width=3, height=3, pointsize=8)
hist(window(resid(fit.mt), start=c(2000,1)), xlab='Standardized Residuals')

***validação dos dados para o conjunto de teste***
valid.t <- predict(fit.mt, n.ahead=12)

```

```

plot.ts(dadost,ylim=c(10,13))
lines(pred.t, col = 'red')
legend("topleft", legend = c("Observado", "Teste"),
+   col = c("black", "red"), lty = 1)

***visualização apenas para o periodo de validação***
valid.t<-predict(fit.mt,n.ahead=12)
plot.ts(test,ylim=c(11.5,13))
lines(valid.t, col = 'red')
legend("topleft", legend = c("Observado", "Teste"),
      col = c("black", "red"), lty = 1)

***previsão a um passo a frente***
fit.m<-hw(dados,h=12,seasonal="multiplicative",initial="o")
summary(fit.m)
plot(fit.m)

```

Série Temporal – Rede Neuronal

```

***instalar o pacote nnfor e carrega-lo para utilizar a função mlp()***
install.packages ("nnfor")

library(nnfor)

***subdivisão da série em treino e teste***

treino<- window(dadost, end = c(2015, 12))

teste <- window(dadost, start = c(2016,1),end=c(2018,12))

***Ajustamento dos dados de treino***

rede.treino<-mlp(ts(dados[-c(1:16)]), start=2000,frequency = 12), lags = 12, hd = 7, reps = 100, comb
= "median", keep = NULL, difforder = c(1,12),outplot =TRUE, sel.lag = FALSE,allow.det.season =
TRUE, det.type = "auto",hd.auto.type = FALSE, hd.max = NULL, retrain = FALSE)

****verificar a acuracia do treino***

accuracy(dadost,rede.treino$fitted)

***Validação dos dados de teste***

rede.teste<-mlp(dados,teste,model = rede.treino)

plot(teste,ylim=c(11,13),
      main="")

lines(rede.teste$fitted,col="red")

legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
      pch = c(NA, NA), col = c("red", "black"))

```

```
##### verificar a acuracia do teste#####  
accuracy(dadost,rede.teste$fitted)  
##### previsão do modelo para um passo a frente (ano de 2019)#####  
prev2019<-forecast(rede.teste,h=12,PI=TRUE)  
prev2019  
plot(prev2019,xlab="Anos",ylab="Numero de Dormidas", main="")  
lines(dados,type="",lty=1,col="black")  
legend("topleft",c("Observado","PrevistoRNA_mlp"),cex=0.8,col=c("black","blue"),lty=1)
```

ANEXO VI: RESULTADOS DA IMPLEMENTAÇÃO DA FUNÇÃO *NNETAR* ()

Aqui se encontram todos os resultados da implementação da função `nnetar()` do R na construção de cada modelo.

Modelo m24: RNAR(2,1,2)[12]

O ajustamento do modelo:

```
rede_treino<-nnetar(treino,decay=0, p=2,P=1,size=(2,lambda="auto",repeats=100)
plot(treino, main="")
lines(rede_treino$fitted,col="red")
legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
      pch = c(NA, NA), col = c("red", "black"))
```

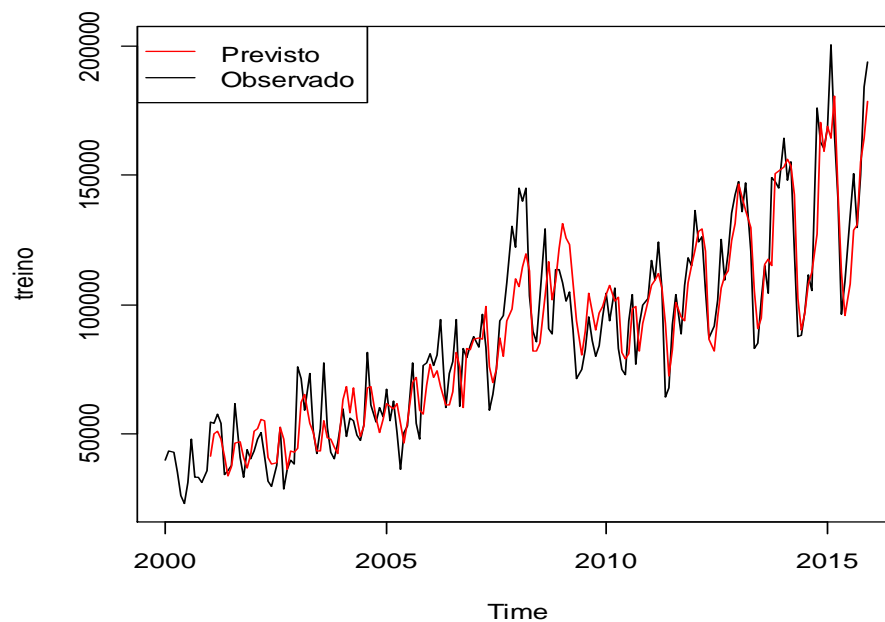


Figura:A6. 1: Ajustamento do modelo **m24** RNAR(2,1,2)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m24** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -992.7679 13269.74 10112.55 -1.227975 12.21729 0.1897617 1.151097
```

Seguidamente é feita a validação do modelo **m24** comparando-o com os dados do conjunto de teste com apoio da Figura A6.2 e os valores da acurácia:

```
rede_teste<-nnetar(dados,start=c(2016,1),model = rede_treino)
plot(teste,
     main="")
lines(rede_teste$fitted,col="red")
legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
     pch = c(NA, NA), col = c("red", "black"))
```

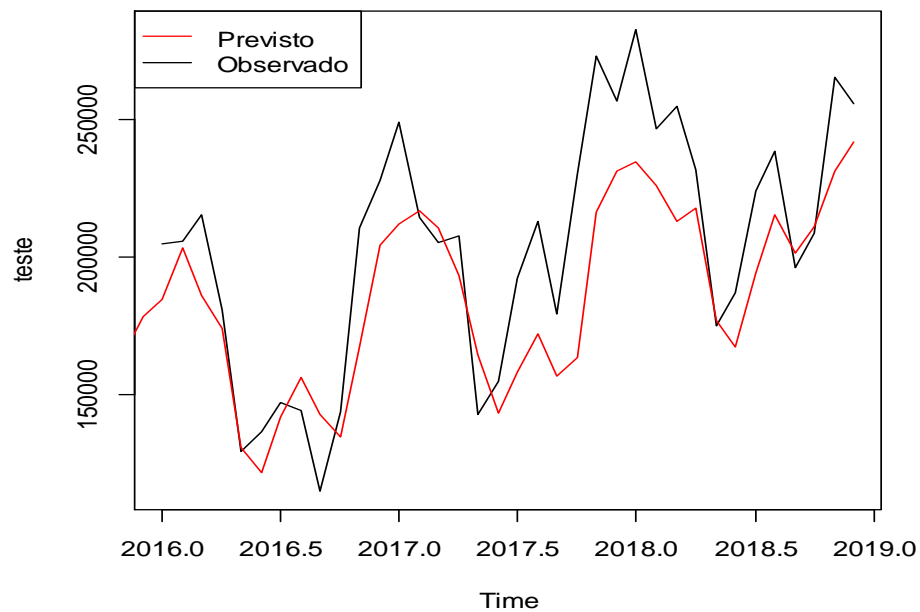


Figura:A6. 2: Ajustamento do modelo **m24** RNAR(2,1,2)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

```
>accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -3734.675 16443.65 12072.72 -2.533129 12.14898 0.3613798 1.172526
```

Modelo m25: NNAR(4,1,3)[12]

O ajustamento do modelo:

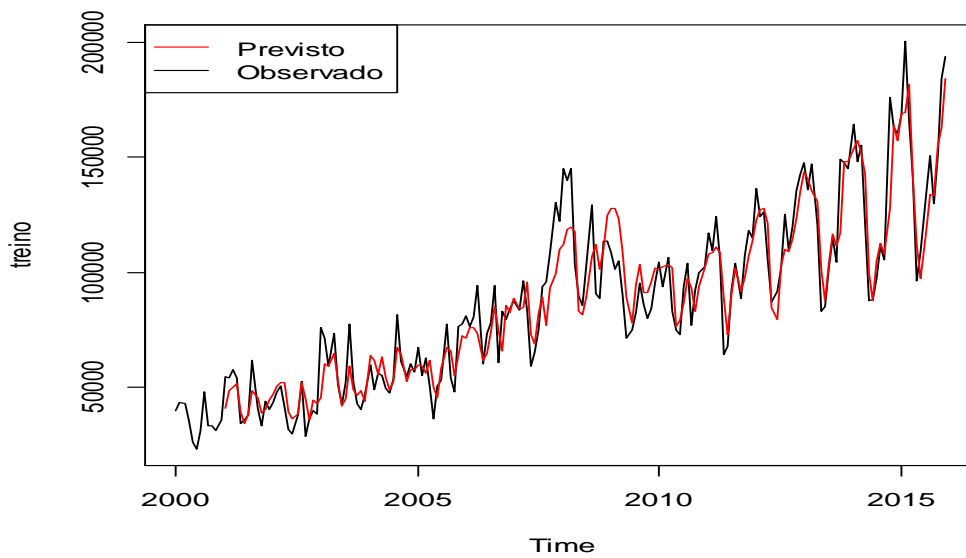


Figura:A6. 3: Ajustamento do modelo **m25** RNAR(4,1,3)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m25** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -928.2047 11791.43 8781.476 -0.9027131 10.43466 0.1893974 1.017617
```

Seguidamente é feita a validação do modelo **m25** comparando-o com os dados do conjunto de teste com apoio da Figura A6.4 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -3409.834 15121.21 10753.01 -2.111932 10.55028 0.3324466 1.019042
```

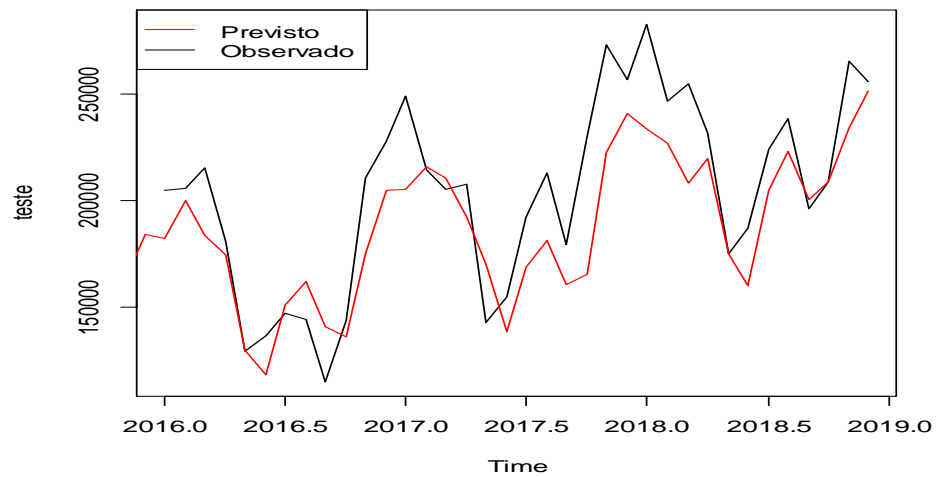


Figura:A6. 4: Ajustamento do modelo **m25** RNAR(4,1,3)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo **m26**: NNAR(6,1,4)[12]

O ajustamento do modelo:

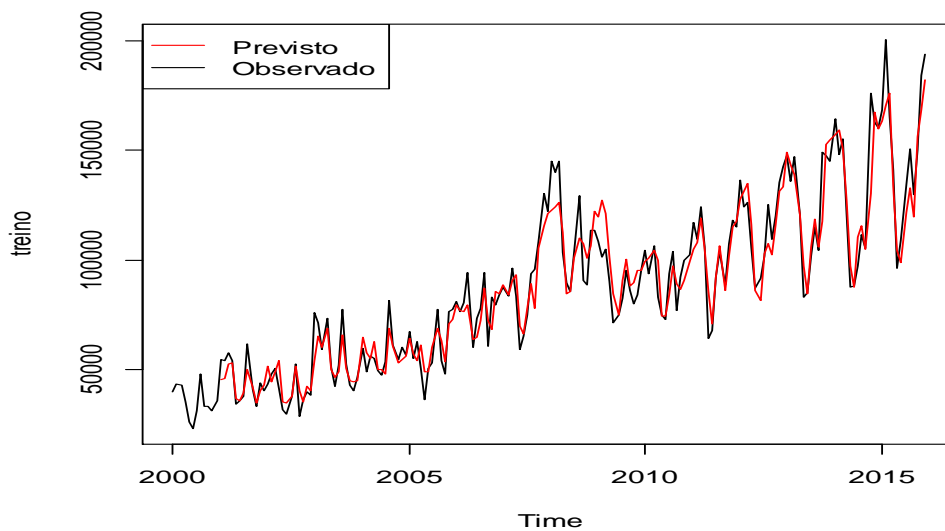


Figura:A6. 5: Ajustamento do modelo **m26** RNAR(6,1,4)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m26** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -786.7458 9620.697 7018.549 -0.5544889 8.120391 0.09641377 0.7342249
```

Seguidamente é feita a validação do modelo **m26** comparando-o com os dados do conjunto de teste com apoio da Figura A6.6 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -6343.547 18893.41 11637.59 -3.865062 10.24902 0.6192672 0.9657233
```

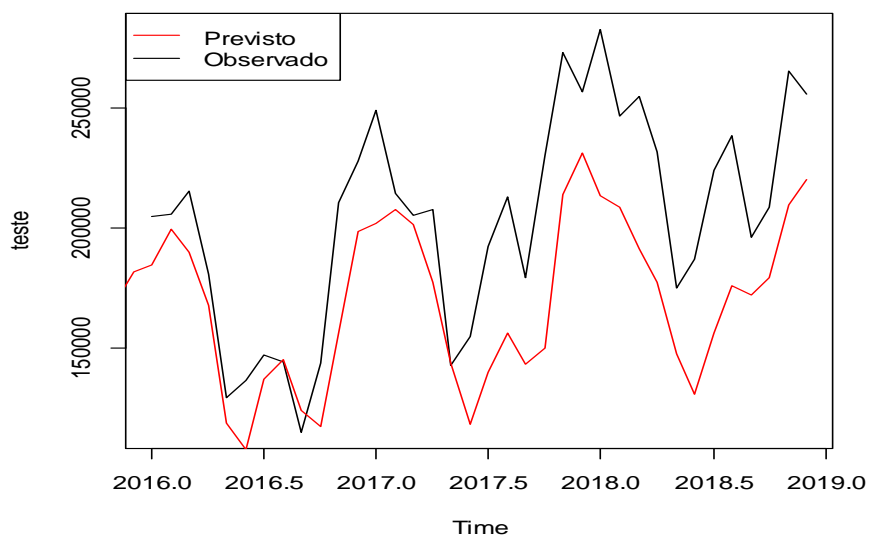


Figura:A6. 6: Ajustamento do modelo **m26** RNAR(6,1,4)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m27: NNAR(8,1,5)[12]

O ajustamento do modelo:

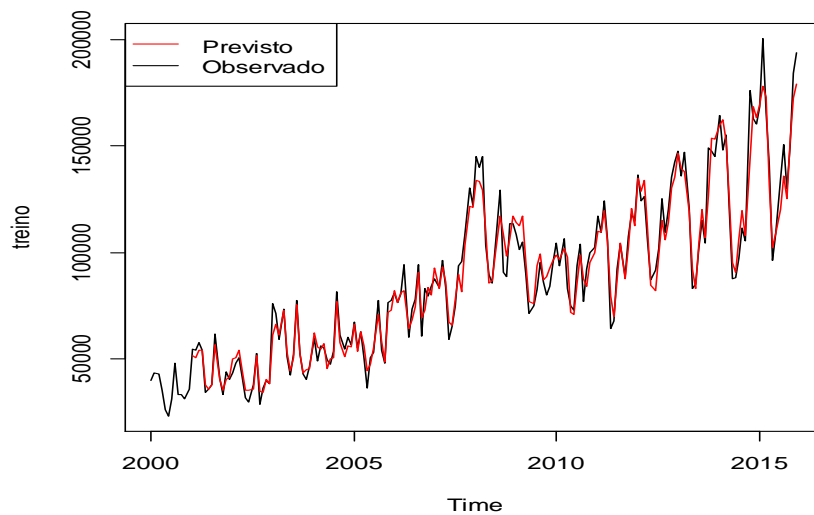


Figura:A6. 7: Ajustamento do modelo **m27** RNAR(8,1,5)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m27** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -590.7174 6976.41 5158.46 -0.2552228 5.828446 0.03545415 0.4601297
```

Seguidamente é feita a validação do modelo **m27** comparando-o com os dados do conjunto de teste com apoio da Figura A6.8 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -6336.7 19302.2 10693.56 -3.787302 8.723798 0.7111869 0.8219317
```

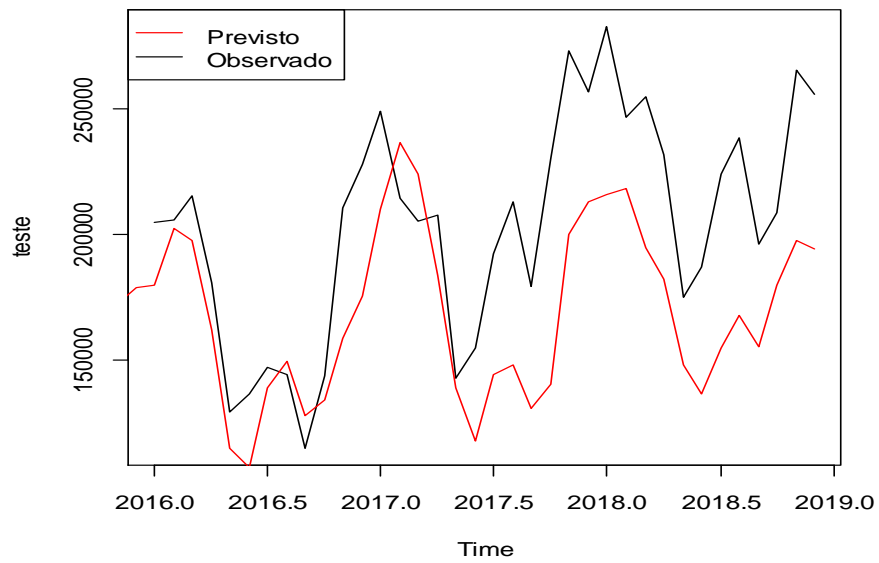


Figura:A6. 8: Ajustamento do modelo **m27** RNAR(8,1,5)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m28: NNAR(10,1,6)[12]

O ajustamento do modelo:

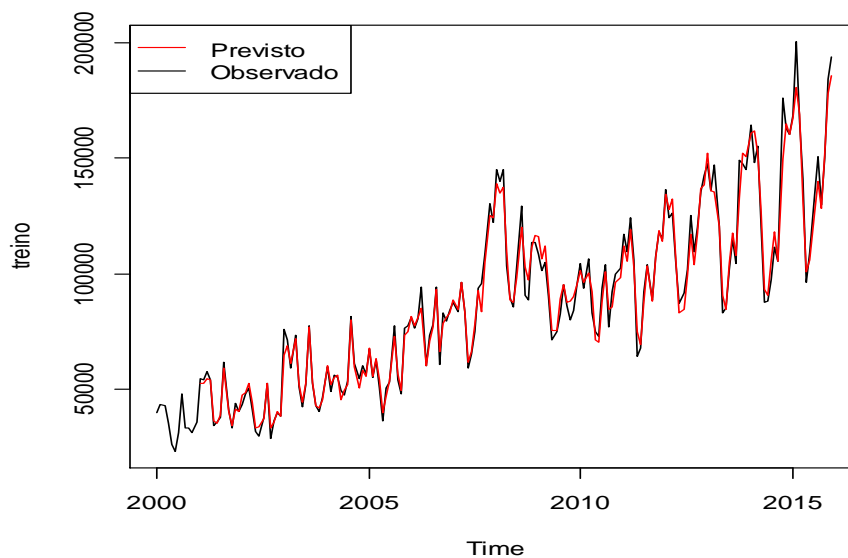


Figura:A6. 9: Ajustamento do modelo **m28** RNAR(10,1,6)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m28** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-452.6058	5156.775	3594.441	-0.1374148	3.979838	0.03985706	0.3132747

Seguidamente é feita a validação do modelo **m28** comparando-o com os dados do conjunto de teste com apoio da Figura A6.10 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-5359.541	16302.41	8474.026	-2.995443	6.517164	0.7121561	0.6307547

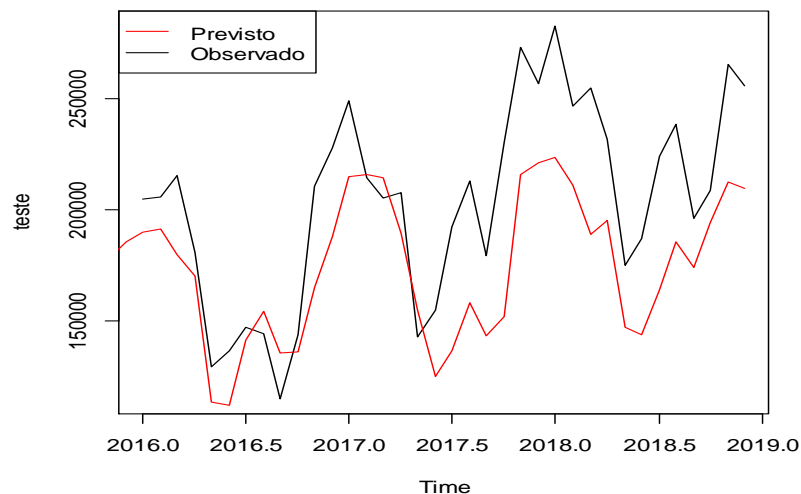


Figura:A6. 10: Ajustamento do modelo m28 RNAR(10,1,6)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m29: NNAR(12,1,7)[12]

O ajustamento do modelo:

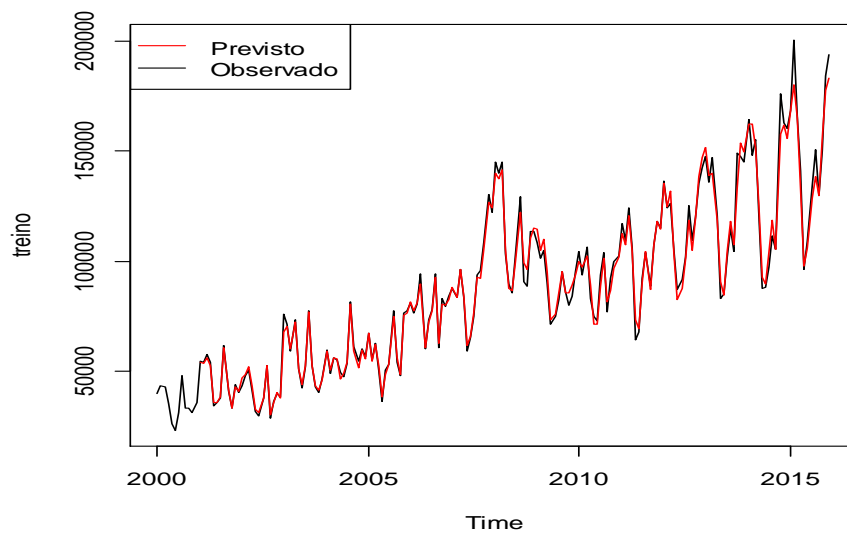


Figura:A6. 11: Ajustamento do modelo **m29** RNAR(12,1,7)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m29** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-332.9067	4216.471	2782.633	-0.06640752	2.890974	0.05712367	0.2203097

Seguidamente é feita a validação do modelo **m29** comparando-o com os dados do conjunto de teste com apoio da Figura A6.12 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-5873.639	17294.81	8265.826	-3.379434	5.965962	0.7539001	0.6088223

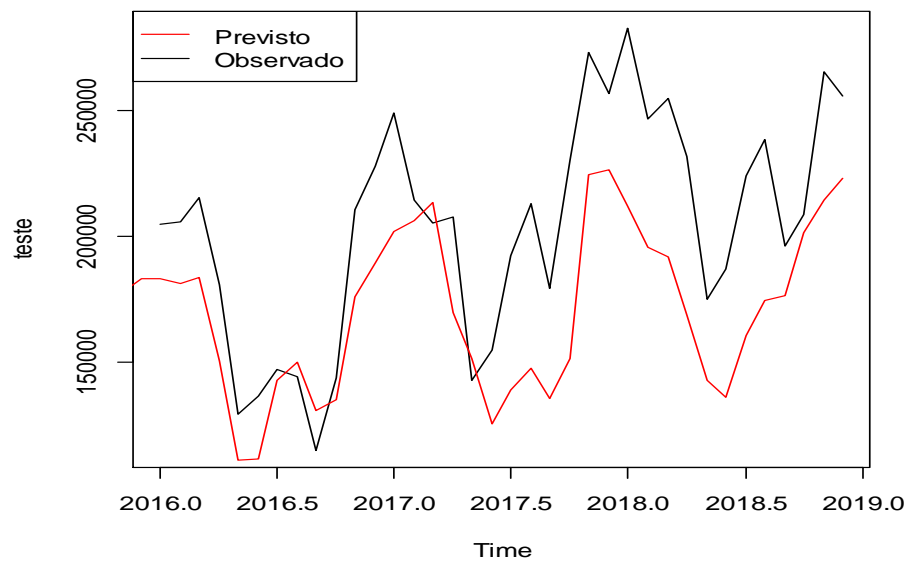


Figura:A6. 12: Ajustamento do modelo m_{29} RNAR(12,1,7)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m_{30} : NNAR(14,1,8)[12]

O ajustamento do modelo:

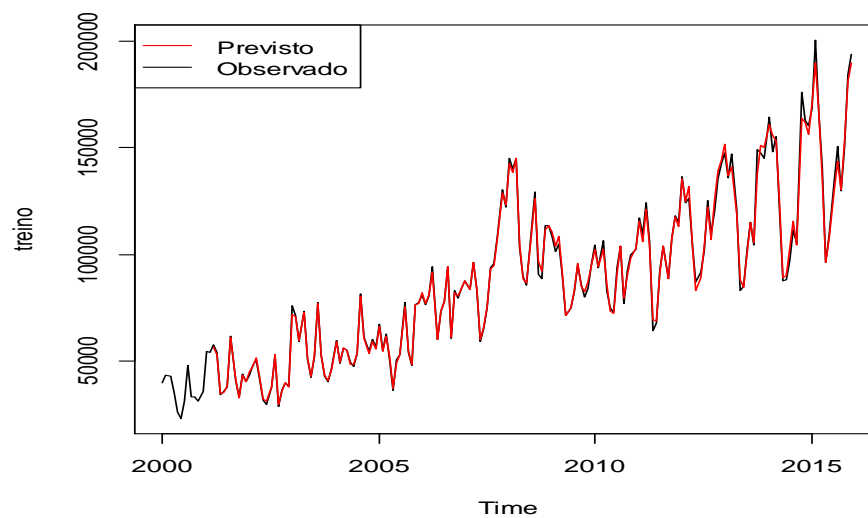


Figura:A6. 13: Ajustamento do modelo m_{30} RNAR(14,1,8)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m30** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -182.2528 2607.126 1678.734 -0.01996619 1.705275 0.01191185 0.1253805
```

Seguidamente é feita a validação do modelo **m30** comparando-o com os dados do conjunto de teste com apoio da Figura A6.14 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -4268.095 13525.07 6086.818 -2.388152 4.127908 0.7425577 0.4299042
```

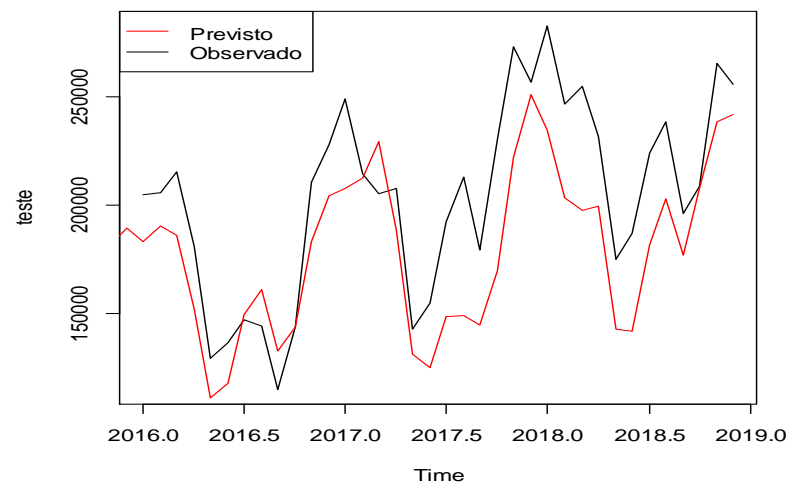


Figura:A6. 14: Ajustamento do modelo **m30** RNAR(14,1,8)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m31: NNAR(16,1,9)[12]

O ajustamento do modelo:

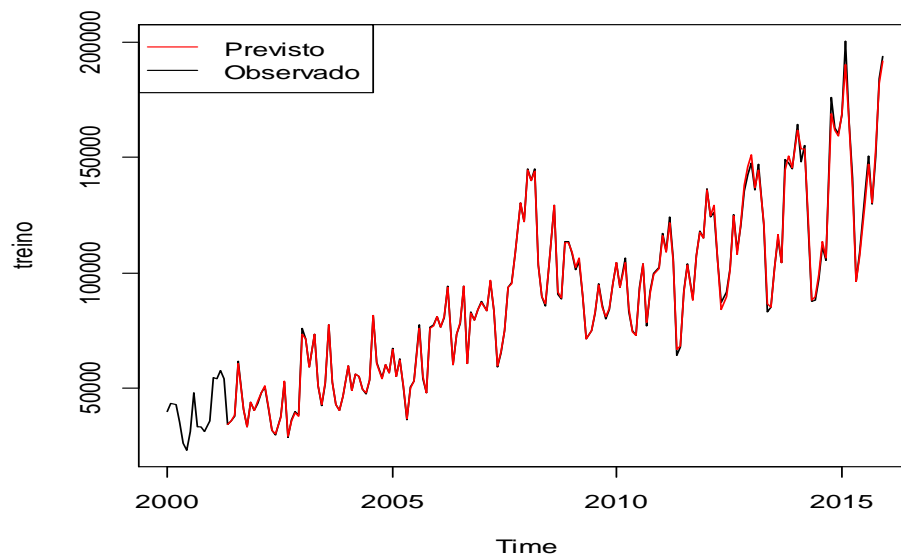


Figura:A6. 15: Ajustamento do modelo **m31** RNAR(16,1,9)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m31** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-98.40146	1596.345	888.2884	-0.003909117	0.8524025	0.06123596	0.06956897

Seguidamente é feita a validação do modelo **m31** comparando-o com os dados do conjunto de teste com apoio da Figura A6.16 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-4864.899	15404.77	6326.785	-2.801911	3.971229	0.7460516	0.4590831

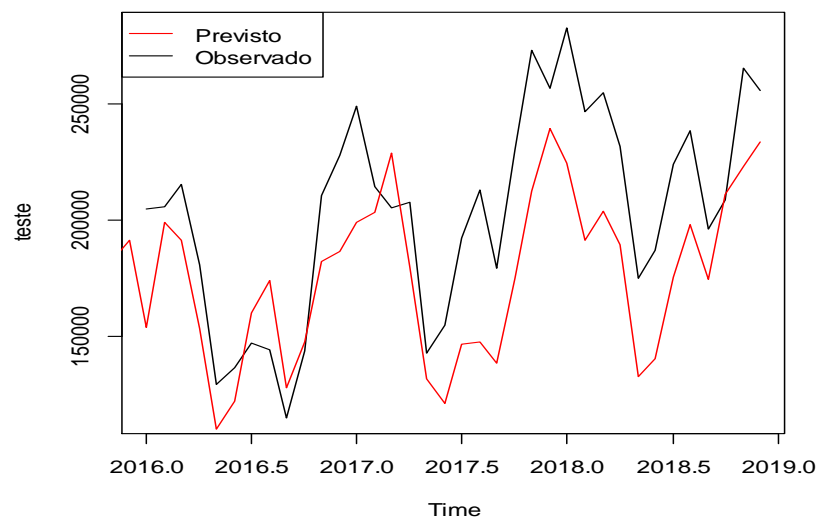


Figura:A6. 16: Ajustamento do modelo **m31** RNAR(16,1,9)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m32: NNAR(18,1,10)[12]

O ajustamento do modelo:

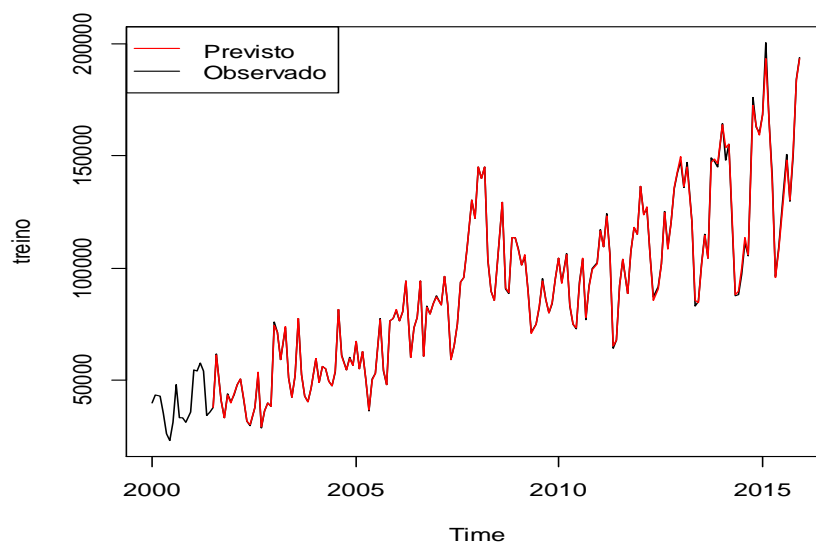


Figura:A6. 17: Ajustamento do modelo **m32** RNAR(18,1,10)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m32** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-43.48964	1003.196	480.1871	0.002142562	0.4517474	0.1290983	0.04064403

Seguidamente é feita a validação do modelo **m32** comparando-o com os dados do conjunto de teste com apoio da Figura A6.2 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-4300.887	14555.74	5445.761	-2.504617	3.292603	0.6674803	0.4121625

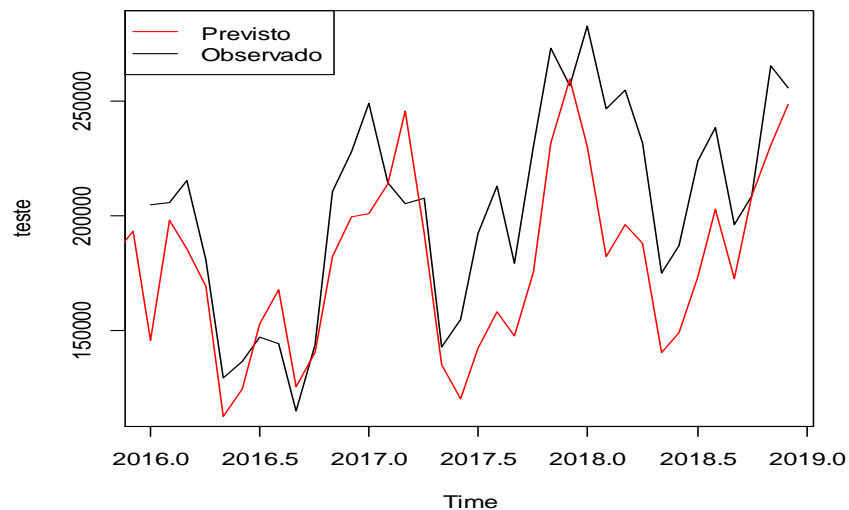


Figura:A6. 18: Ajustamento do modelo **m32** RNAR(18,1,10)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m33: NNAR(20,1,11)[12]

O ajustamento do modelo:

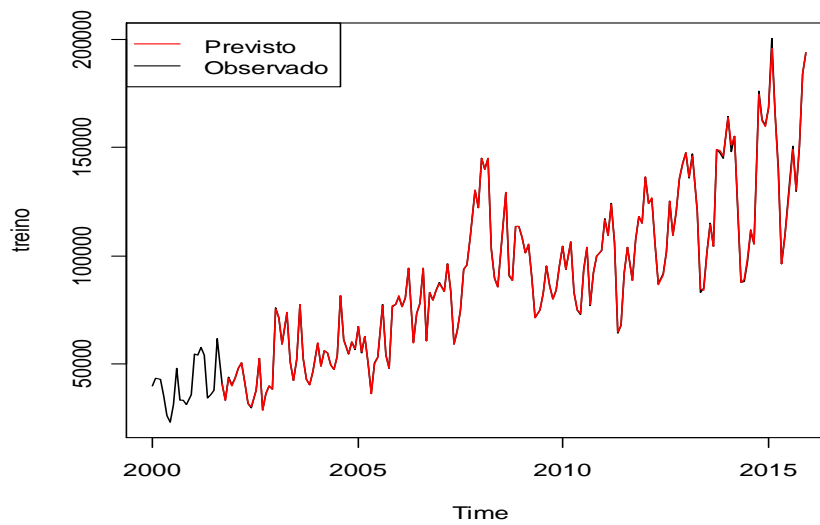


Figura:A6. 19: Ajustamento do modelo **m33** RNAR(20,1,11)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m33** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -29.1942  572.2741  256.0083 -0.004645538  0.2376959  0.1960908  0.02248498
```

Seguidamente é feita a validação do modelo **m33** comparando-o com os dados do conjunto de teste com apoio da Figura A6.20 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -5575.761  16905.62  6452.319 -3.397512  4.006112  0.8300387  0.5223628
```

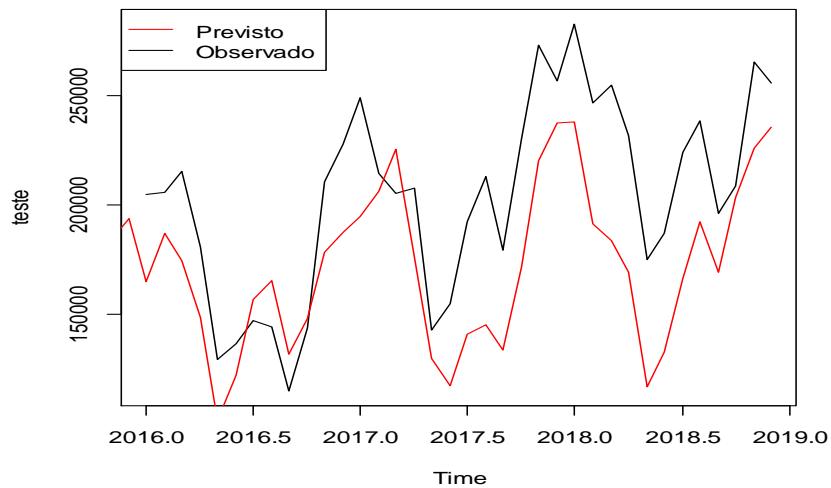


Figura:A6. 20: Ajustamento do modelo **m33** RNAR(20,1,11)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m34: NNAR(22,1,12)[12]

O ajustamento do modelo:

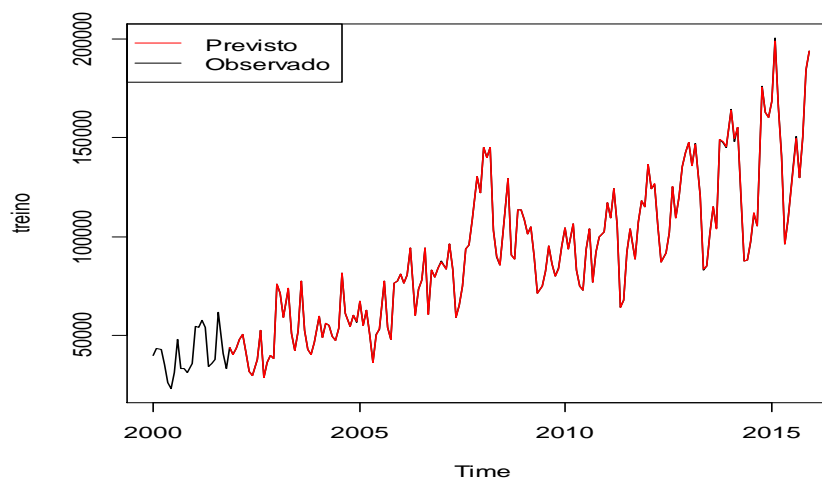


Figura:A6. 21: Ajustamento do modelo **m34** RNAR(22,1,12)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

O ajustamento do modelo **m34** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-9.802019	231.3688	103.5167	-0.001284329	0.09238996	0.1273328	0.00810196

Seguidamente é feita a validação do modelo **m34** comparando-o com os dados do conjunto de teste com apoio da Figura A6.22 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-5756.485	17250.07	6366.627	-3.499993	3.925491	0.8347531	0.5393656

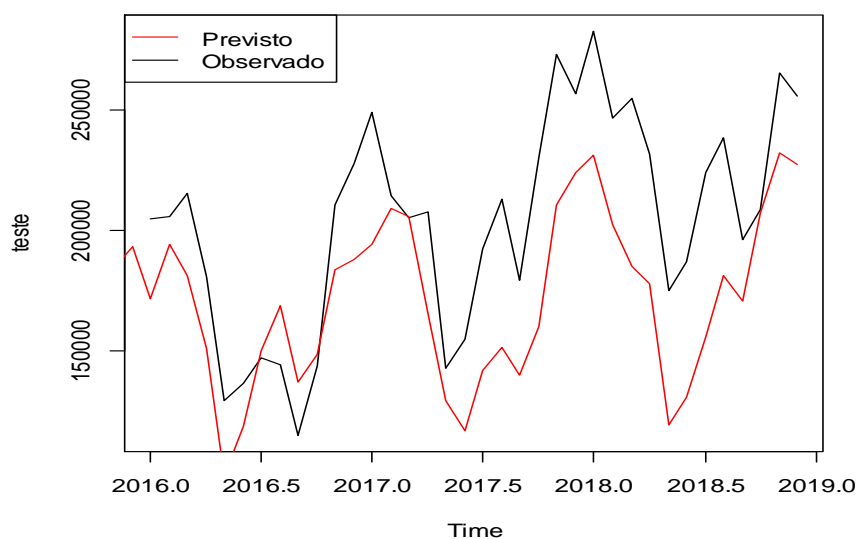


Figura:A6. 22: Ajustamento do modelo **m34** RNAR(22,1,12)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

Modelo m35: NNAR(24,1,13)[12]

O ajustamento do modelo m35 comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dados,rede_treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	2.455267	168.009	73.23709	0.00661965	0.0635572	0.1431038	0.006349224

Seguidamente é feita a validação do modelo **m35** comparando-o com os dados do conjunto de teste com apoio da Figura A6.24 e os valores da acurácia:

```
> accuracy(dados,rede_teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-6531.985	18394.49	6888.424	-3.981028	4.246093	0.877515	0.5757865

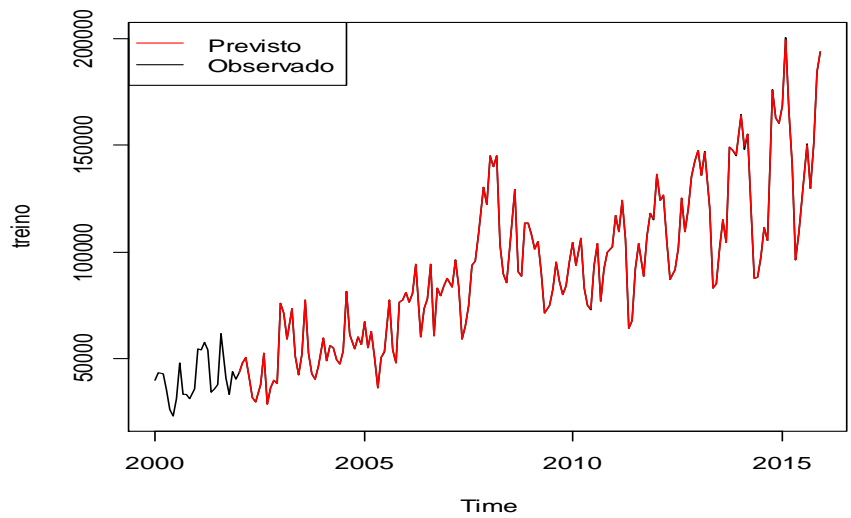


Figura:A6. 23: Ajustamento do modelo **m35** RNAR(24,1,13)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12]

Fonte: Elaboração própria

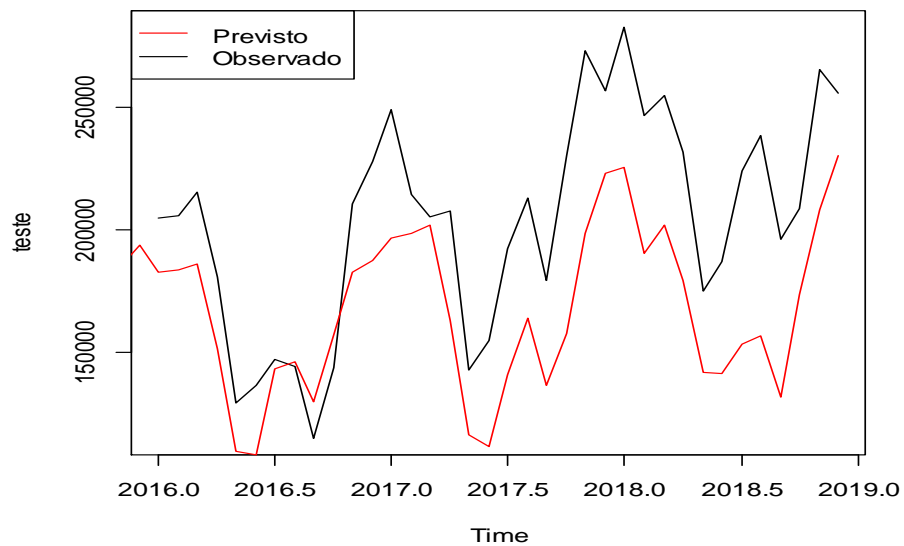


Figura:A6. 24: Ajustamento do modelo **m35** RNAR(24,1,13)[12] comparando-o com os dados do conjunto de teste, [2000:01, 2015:12]

Fonte: Elaboração própria

ANEXO VII: RESULTADOS DAS ANÁLISES DOS MODELOS 1, 2, 3 E 4

Modelo m1: SARIMA(2,1,1)(1,1,1)₁₂

```
> coeftest(m1)
t test of coefficients:

Estimate Std. Error t value Pr(>|t|)
ar1 -0.310624  0.330094 -0.9410  0.34670
ar2 -0.262703  0.108148 -2.4291  0.01514 *
ma1 -0.018795  0.345734 -0.0544  0.95665
sar1 -0.099658  0.100162 -0.9950  0.31975
sma1 -0.639554  0.079390 -8.0558 7.894e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(m1)
Series: train
ARIMA(2,1,1)(1,1,1)[12]
sigma^2 estimated as 0.01611: log likelihood=129.48
AIC=-246.96  AICc=-246.53  BIC=-227.08
```

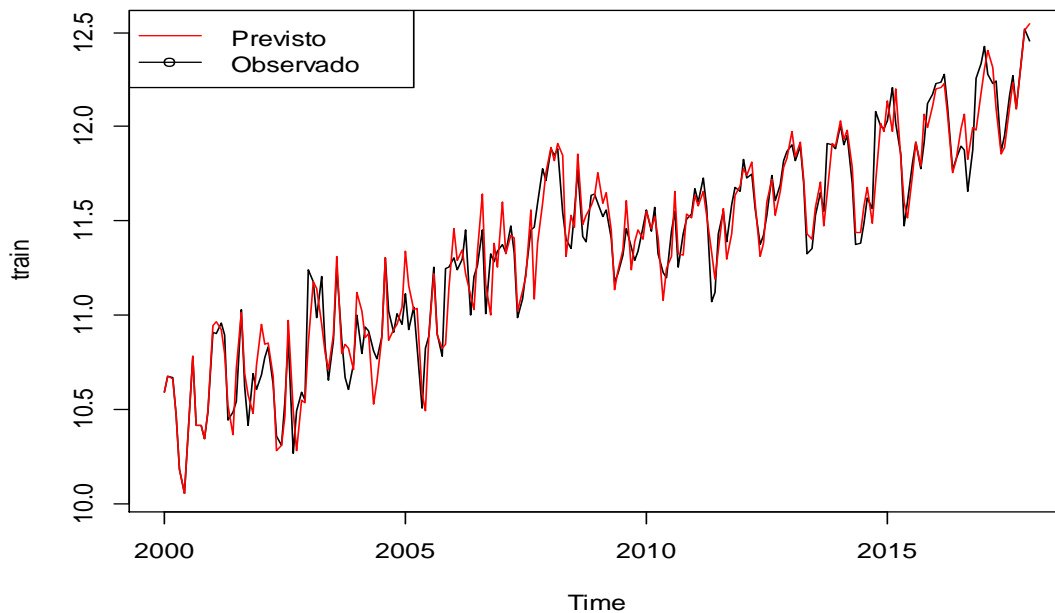


Figura: A7.1: Ajustamento do modelo **m1**, SARIMA(2,1,1) (1,1,1)₁₂ comparando-o com os dados do conjunto de treino, [2000:01, 2017:12].

Fonte: Elaboração própria

Qualidade do ajustamento

```
> accuracy(train,m1$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.001141636	0.1215264	0.08734913	0.00478809	0.7727594	-0.000778992	0.5564589

Previsão da série para o conjunto de teste

```
> plot.ts(pred$pred, col = "red", ylim = c(11, 13),
+ ylab = "",
+ main = "Previsão do modelo SARIMA(2, 1, 1)(1, 1, 1)[12]")
> lines(test, type = "o")
> legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
+ pch = c(NA, 1), col = c("red", "black"))
```

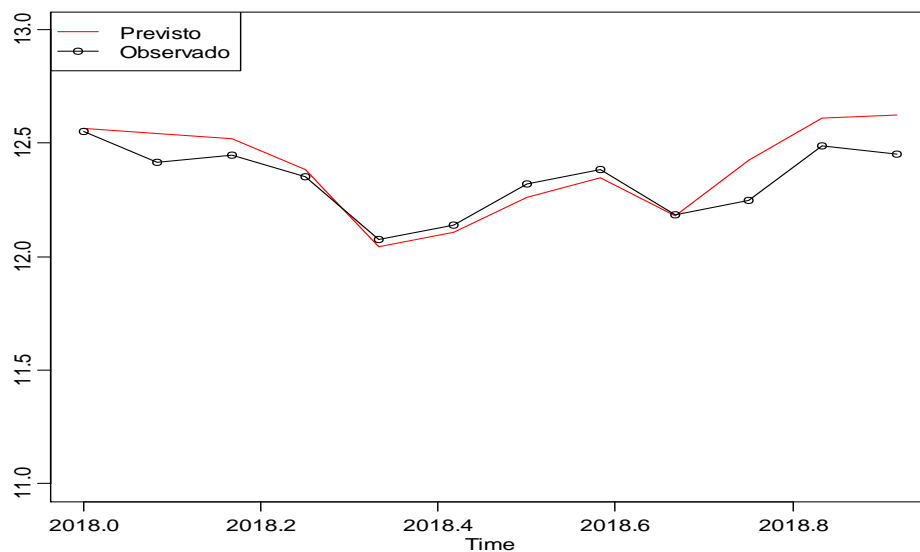


Figura: A7.2: Ajustamento do modelo **m1** SARIMA(2,1,1) (1,1,1)₍₁₂₎ comparando-o com os dados do conjunto de teste, [2018:01, 2018:12].

Fonte: Elaboração própria

Qualidade da previsão

```
> accuracy(test,pred$pred)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.04586797	0.09357349	0.07307514	0.3629342	0.5859115	0.5007219	0.5946807

Modelo m2. SARIMA(1,1,1)(0,1,1)₁₂

```
> coeftest(m2)
t test of coefficients:

Estimate Std. Error t value Pr(>|t|)
ar1  0.585892  0.107446  5.4529 4.956e-08 ***
ma1  -0.904472  0.069767 -12.9642 < 2.2e-16 ***
sma1 -0.723728  0.051110 -14.1603 < 2.2e-16 ***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(m2)
Series: train
ARIMA(1,1,1)(0,1,1)[12]
sigma^2 estimated as 0.01591: log likelihood=128.8
AIC=-249.59 AICc=-249.39 BIC=-236.34
```

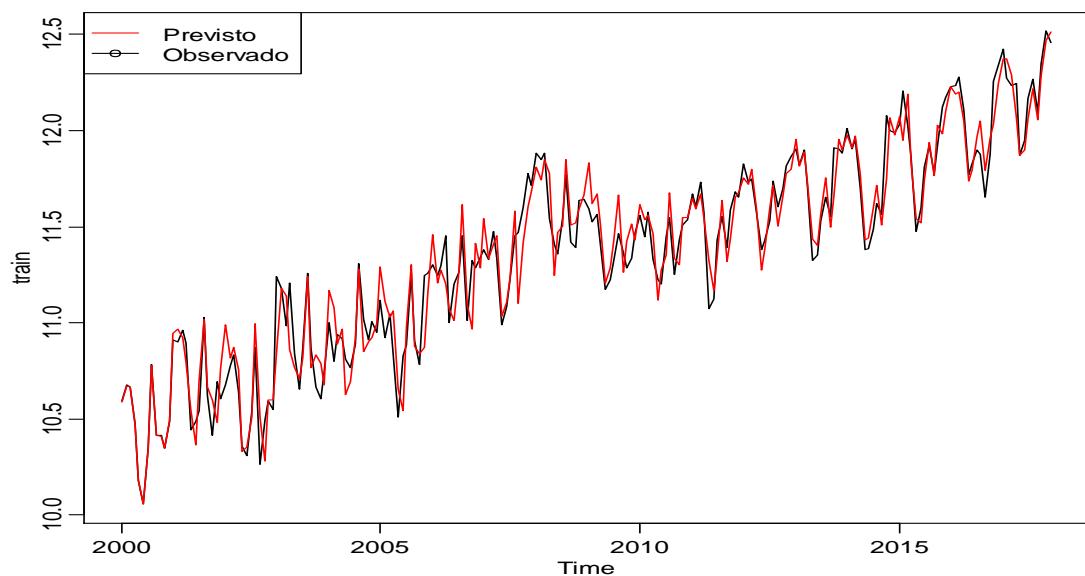


Figura: A7.3: Ajustamento do modelo **m2** SARIMA(1,1,1) (0,1,1)₁₂ comparando-o com os dados do conjunto de treino, [2000:01, 2017:12].

Fonte: Elaboração própria

Qualidade do ajustamento

```
> accuracy(train,m2$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.002538441	0.121375	0.08974515	0.02004729	0.7940888	-0.04430739	0.5504231

Previsão da série para o conjunto de teste

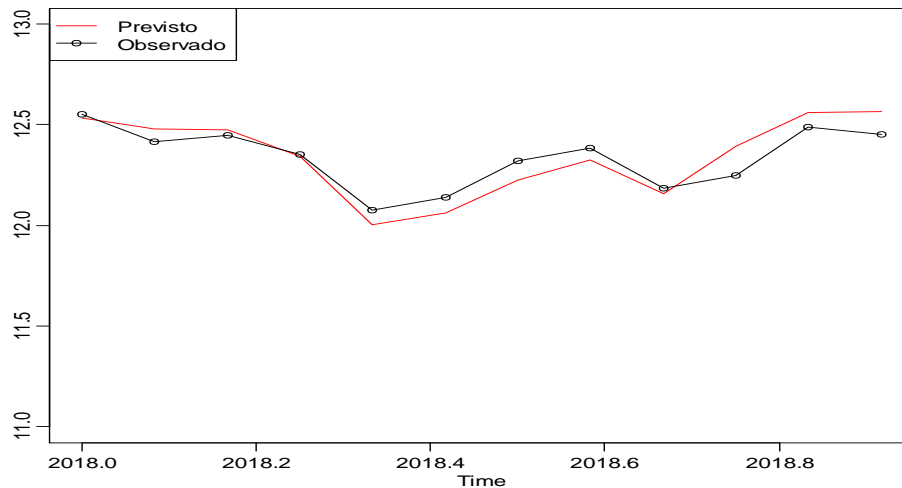


Figura: A7.4: Ajustamento do modelo **m2** SARIMA(1,1,1) (0,1,1)₍₁₂₎ comparando-o com os dados do conjunto de teste, [2018:01, 2018:12].

Fonte: Elaboração própria

Qualidade da previsão

```
> accuracy(test,pred$pred)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.005332413	0.07549765	0.06486781	0.03653867	0.5256627	0.5180242	0.4873328

Modelo m3. SARIMA(2,1,0)(0,1,1)₁₂

```
> coeftest(m3)
t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
ar1 -0.330662  0.068537  -4.8246 1.403e-06 ***
ar2 -0.265289  0.067522  -3.9289 8.532e-05 ***
sma1 -0.691886  0.052809 -13.1018 < 2.2e-16 ***
-----
Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(m3)
```

Series: train

ARIMA(2,1,0)(0,1,1)[12]

sigma^2 estimated as 0.01603: log likelihood=128.99

AIC=-249.98 AICc=-249.78 BIC=-236.73

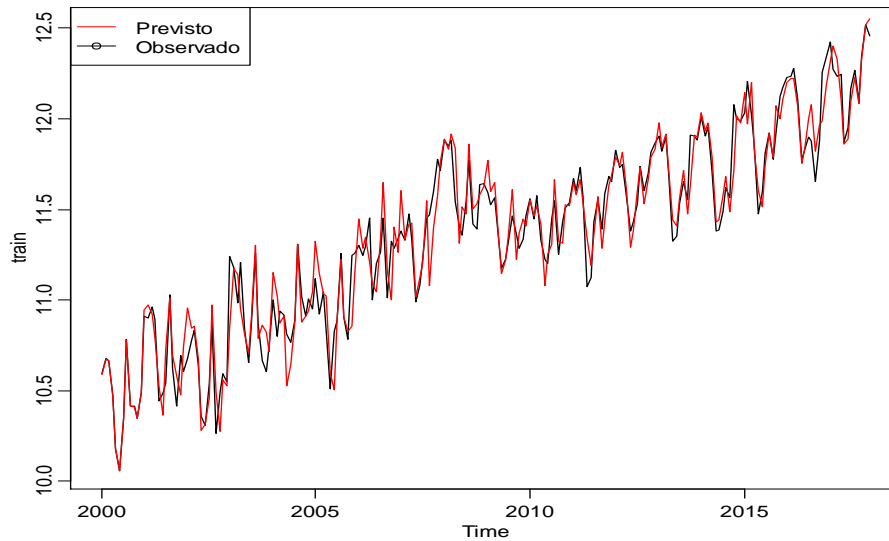


Figura: A7.5: Ajustamento do modelo **m3** SARIMA(2,1,0) (0,1,1)₍₁₂₎ comparando-o com os dados conjunto de treino, [2000:01, 2017:12].

Fonte: Elaboração própria

Qualidade do ajustamento

```
> accuracy(train,m3$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.001303442	0.1218472	0.08747936	0.006092257	0.7741481	0.0008746874	0.5559805

Previsão da série para o conjunto de teste

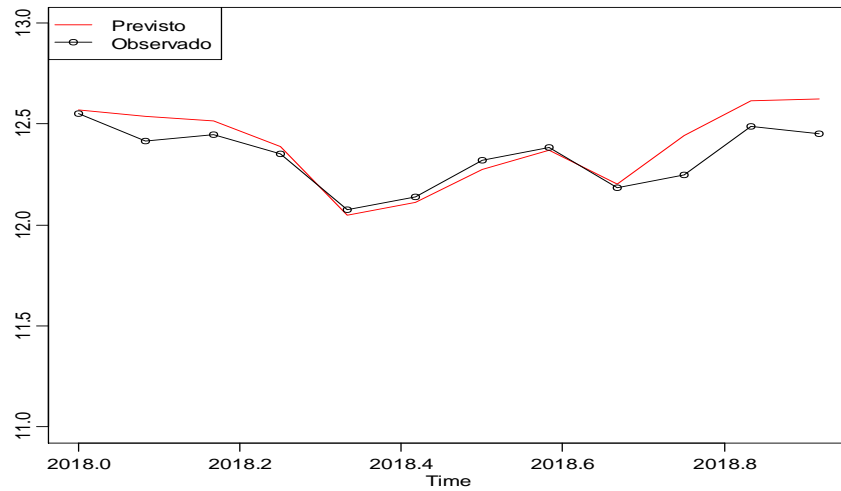


Figura: A7.6: Ajustamento do modelo **m3** SARIMA(2,1,0) (0,1,1)₍₁₂₎ comparando-o com os dados do conjunto de teste, [2018:01, 2018:12].

Fonte: Elaboração própria

Qualidade da previsão

```
> accuracy(test,pred$pred)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.05365535	0.09481637	0.07213158	0.426213	0.5777618	0.5044857	0.6073923

Modelo m4. SARIMA(2,1,0)(0,1,2)₁₂

```
> coeftest(m4)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
ar1	-0.327553	0.068545	-4.7787	1.765e-06 ***
ar2	-0.266780	0.067508	-3.9518	7.756e-05 ***
sma1	-0.747263	0.075066	-9.9547	< 2.2e-16 ***
sma2	0.083003	0.076055	1.0914	0.2751

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> summary(m4)
```

Series: train

ARIMA(2,1,0)(0,1,2)[12]

sigma^2 estimated as 0.01601: log likelihood=129.58

AIC=-249.15 AICc=-248.85 BIC=-232.59

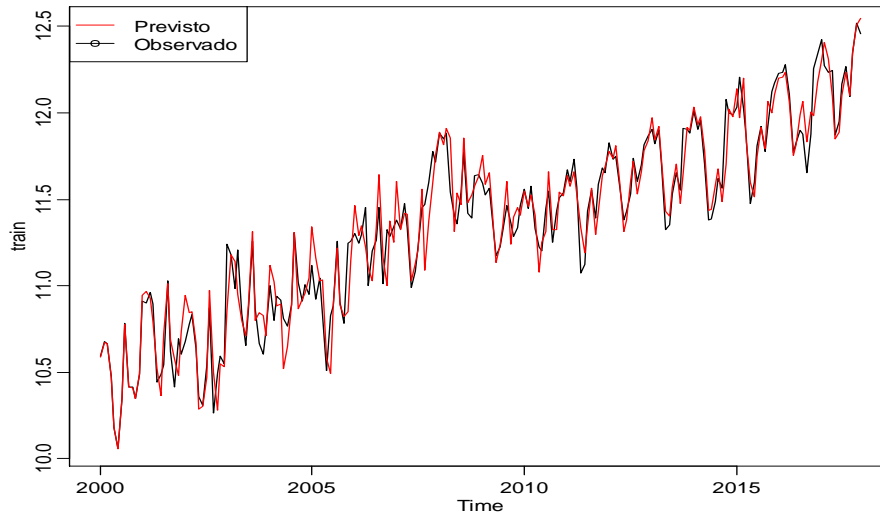


Figura: A7.7: Ajustamento do modelo **m4** SARIMA(2,1,0) (0,1,2)₍₁₂₎ comparando-o com os dados do conjunto de treino, [2000:01, 2017:12].

Fonte: Elaboração própria

Qualidade do ajustamento

```
> accuracy(train,m4$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.001112549	0.1214652	0.08737794	0.004537384	0.7729454	-0.002336056	0.5564847

Previsão da série para o conjunto de teste

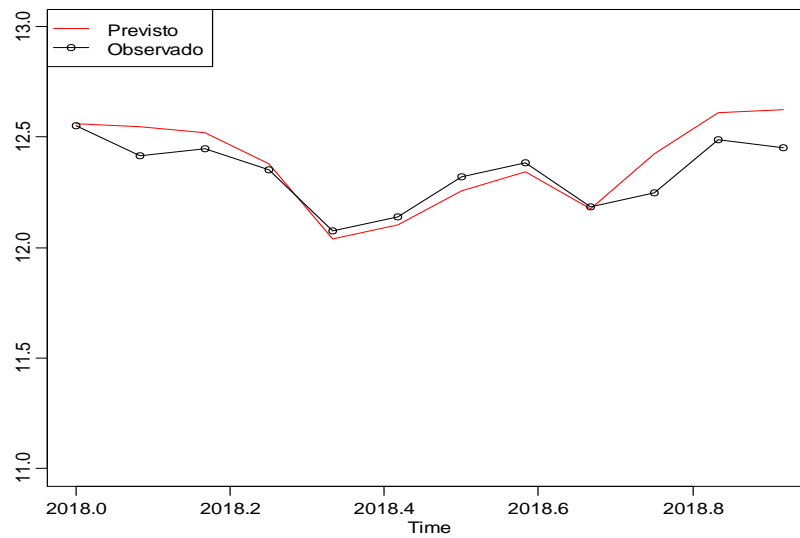


Figura: A7.8: Ajustamento do modelo **m4** SARIMA(2,1,0) (0,1,2)₍₁₂₎ comparando-o com os dados do conjunto de teste, [2018:01, 2018:12].

Fonte: Elaboração própria

Qualidade da previsão

```
> accuracy(test,pred$pred)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.04339857	0.09374902	0.07415148	0.3428064	0.5949246	0.501067	0.5941759

ANEXO VIII: RESULTADOS DA IMPLEMENTAÇÃO DA FUNÇÃO *MLP* ()

Modelo m6: RNAR(2,1,2)[12]

O comando seguinte faz a implementação da função `mlp()` para uma rede neuronal de 2 lags de entrada e dois neurónios na camada oculta e com a elaboração automática da Figura A8.1 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

```
> rede.treino<-mlp(ts(dadost[-c(1:16)]), start=2000,frequency = 12), hd = 2, lags = 2, reps = 100,
comb = "median", keep = NULL, difforder = c(1,12),outplot =TRUE, sel.lag = FALSE,allow.det.season
= TRUE, det.type = "auto",hd.auto.type = FALSE, hd.max = NULL, retrain = FALSE)
```

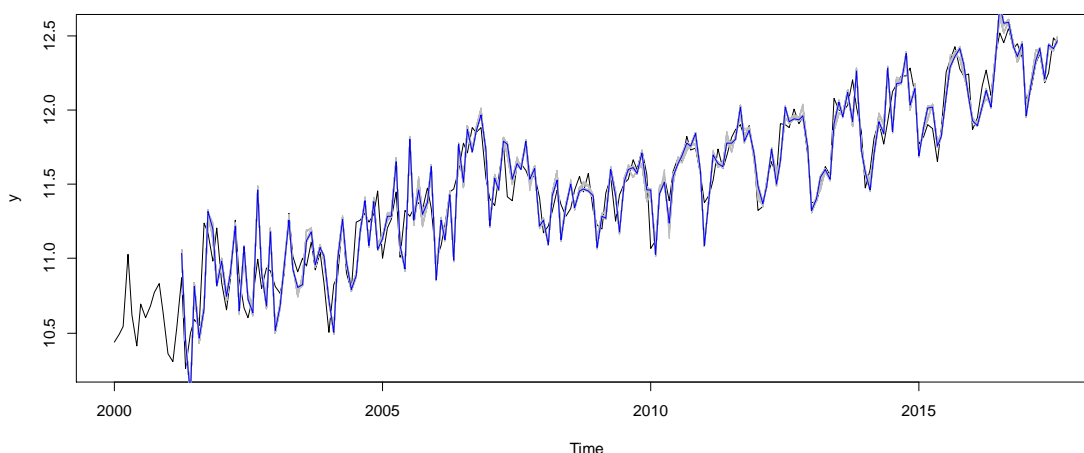


Figura:A8.1: Ajustamento do modelo **m6** RNAR(2,1,2)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m6** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.1398067 0.380849 0.3062223 1.160269 2.643151 0.5003195 1.320898
```

Seguidamente é feita a validação do modelo **m6** comparando-o com os dados do conjunto de teste com apoio da Figura A8.2 e os valores da acurácia:

```
> rede.teste<-mlp(dadost,teste,model = rede.treino)
>
> plot(teste,
```

```
+ main="")
> lines(rede.teste$fitted,col="red")
> legend("topleft", legend = c("Previsto", "Observado"), lty = 1,
+       pch = c(NA, NA), col = c("red", "black"))
```

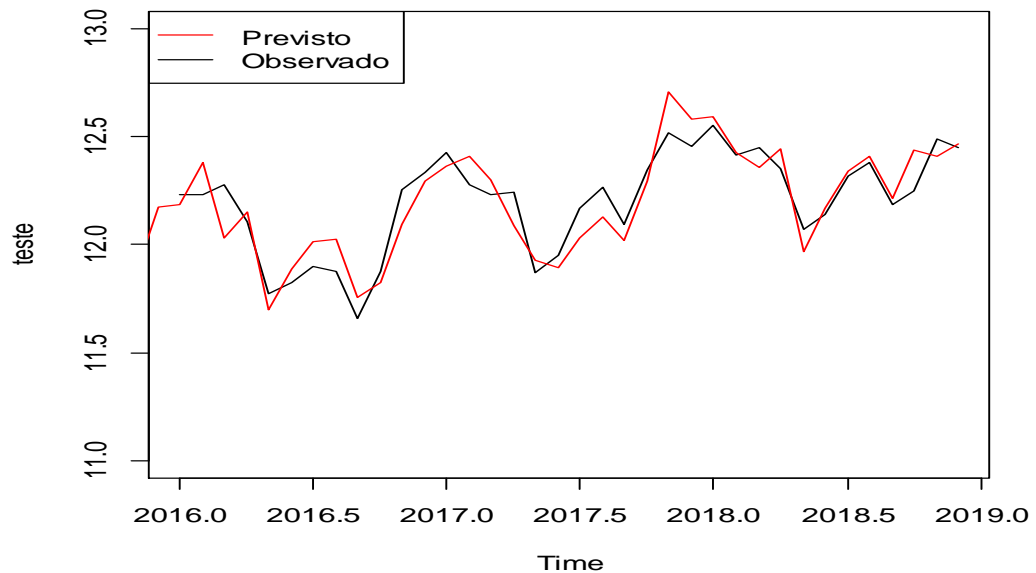


Figura:A8.2: Validação do modelo **m6** RNAR(2,1,2)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.001585986	0.1652311	0.1256562	-0.001890005	1.112105	-0.3409493	0.5755729

Modelo m7: RNAR(4,1,3)[12]

Este modelo de rede neuronal de 4 lags de entrada e 3 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.3 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

O ajustamento do modelo **m7** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1426374	0.3819335	0.3088357	1.183036	2.663976	0.5080798	1.346314

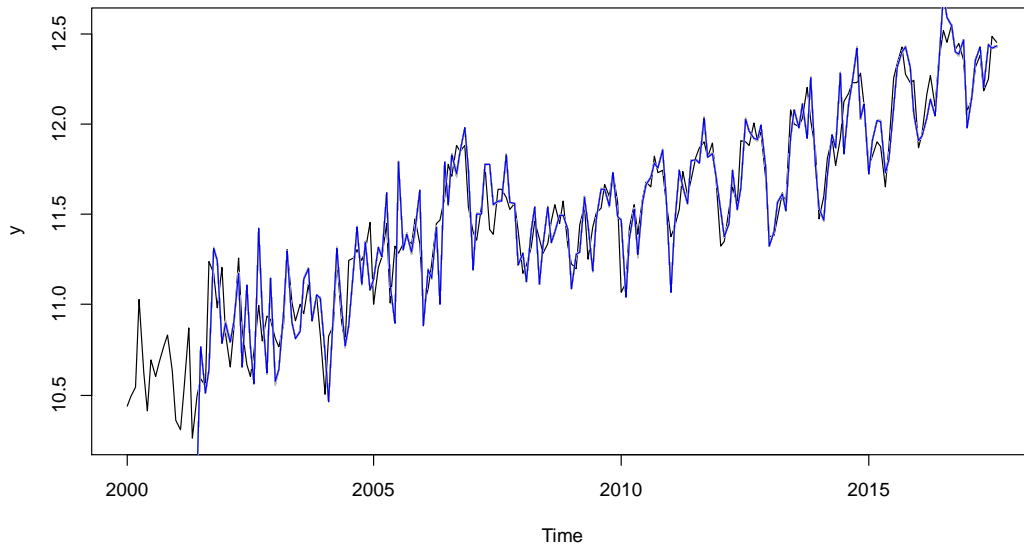


Figura:A8.3: Ajustamento do modelo **m7** RNAR(4,1,3)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

Seguidamente é feita a validação do modelo **m7** comparando-o com os dados do conjunto de teste com apoio da Figura A8.4 e os valores da acurácia.

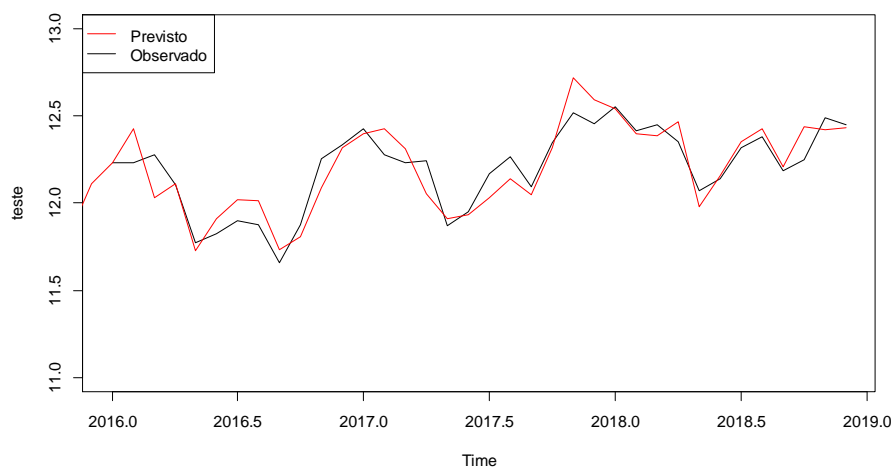


Figura:A8.4: Validação do modelo **m7** RNAR(4,1,3)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.00317544	0.1677717	0.1253795	0.01013874	1.108621	-0.2928713	0.58864

Modelo m8: RNAR(6,1,4)[12]

Este modelo de rede neuronal de 6 lags de entrada e 4 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.5 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

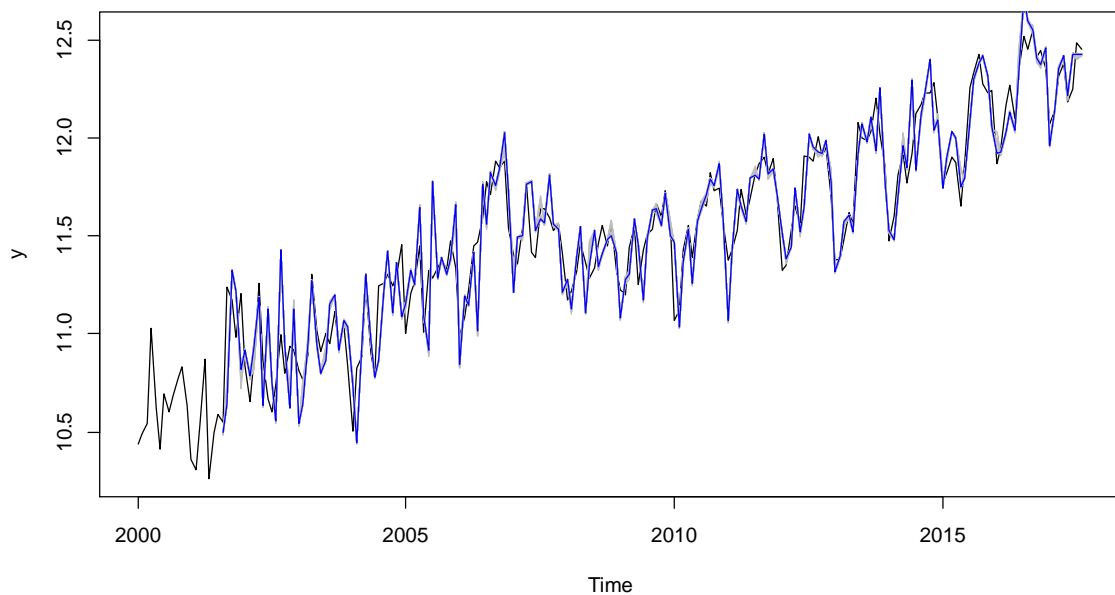


Figura:A8.5: Ajustamento do modelo **m8** RNAR(6,1,4)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m8** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1438388	0.383879	0.3091362	1.194341	2.663862	0.5157339	1.370395

Seguidamente é feita a validação do modelo **m8** comparando-o com os dados do conjunto de teste com apoio da Figura A8.6 e os valores da acurácia.

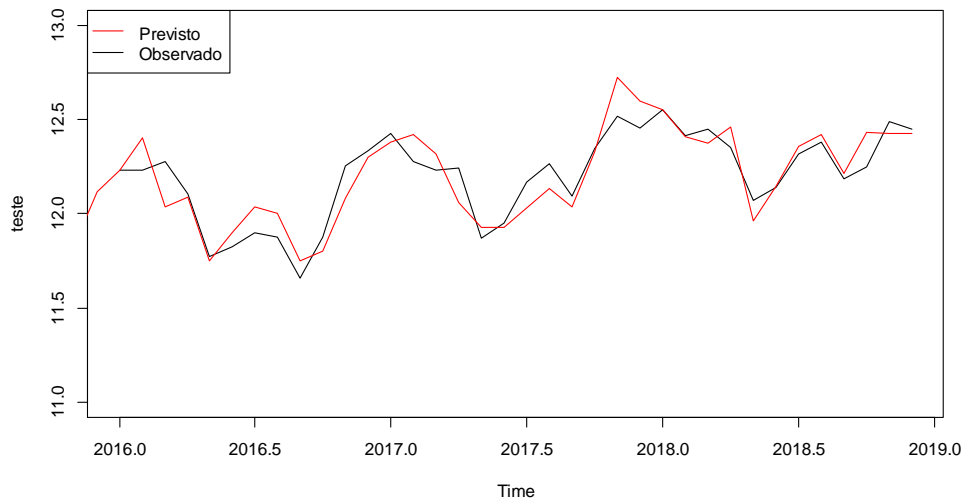


Figura:A8.6: Validação do modelo **m8** RNAR(6,1,4)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.001325938	0.1677528	0.124909	-0.006495551	1.104041	-0.2876351	0.5877792

Modelo m9: RNAR(8,1,5)[12]

Este modelo de rede neuronal de 8 lags de entrada e 5 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.7 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

O ajustamento do modelo **m9** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1446499	0.3811533	0.3072238	1.205265	2.645449	0.5327807	1.386682

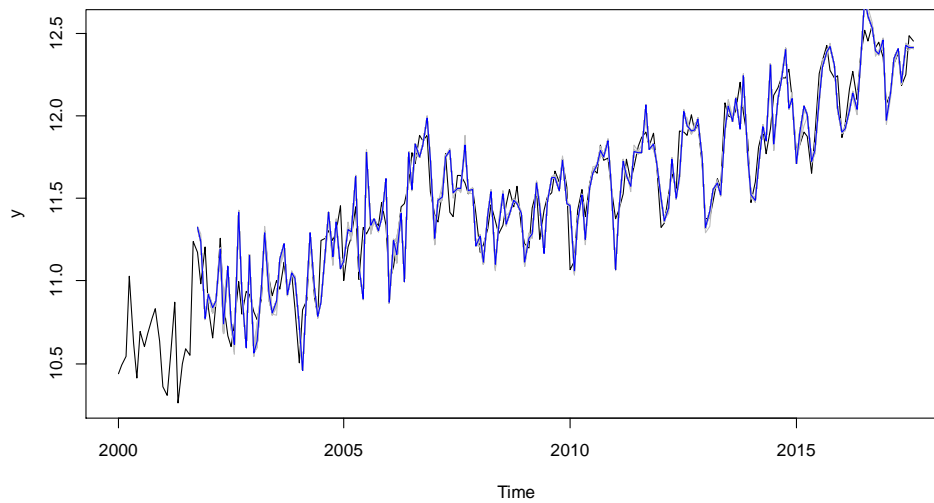


Figura:A8.7: Ajustamento do modelo **m9** RNAR(8,1,5)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

Seguidamente é feita a validação do modelo **m9** comparando-o com os dados do conjunto de teste com apoio da Figura A8.8 e os valores da acurácia.

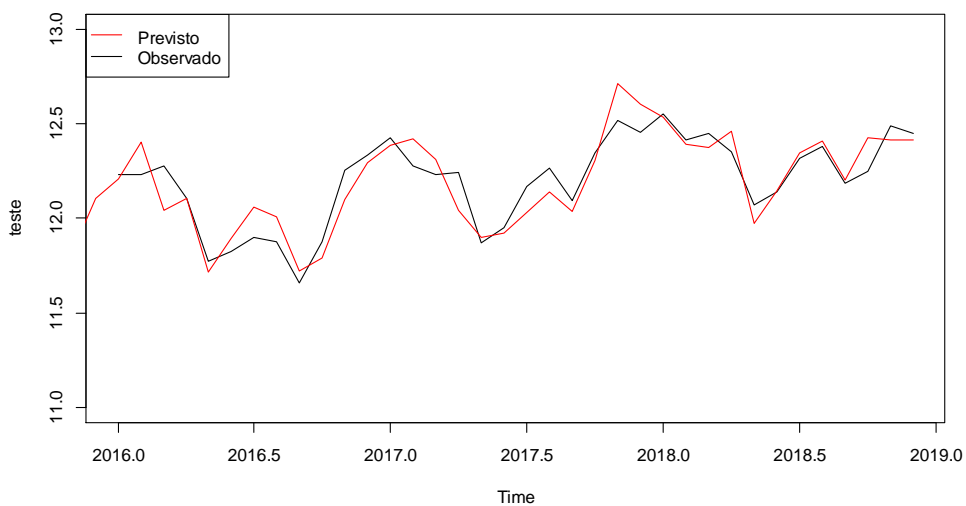


Figura:A8.8: Validação do modelo **m9** RNAR(8,1,5)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -0.001900986  0.1674  0.1245285 -0.03332909  1.099675 -0.2744008  0.5941054
```

Modelo m10: RNAR(10,1,6)[12]

Este modelo de rede neuronal de 10 lags de entrada e 6 neurónios na camada oculta, resultante da aplicação do comando da função $mlp()$ dada anteriormente, permite a elaboração automática da Figura A8.9 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

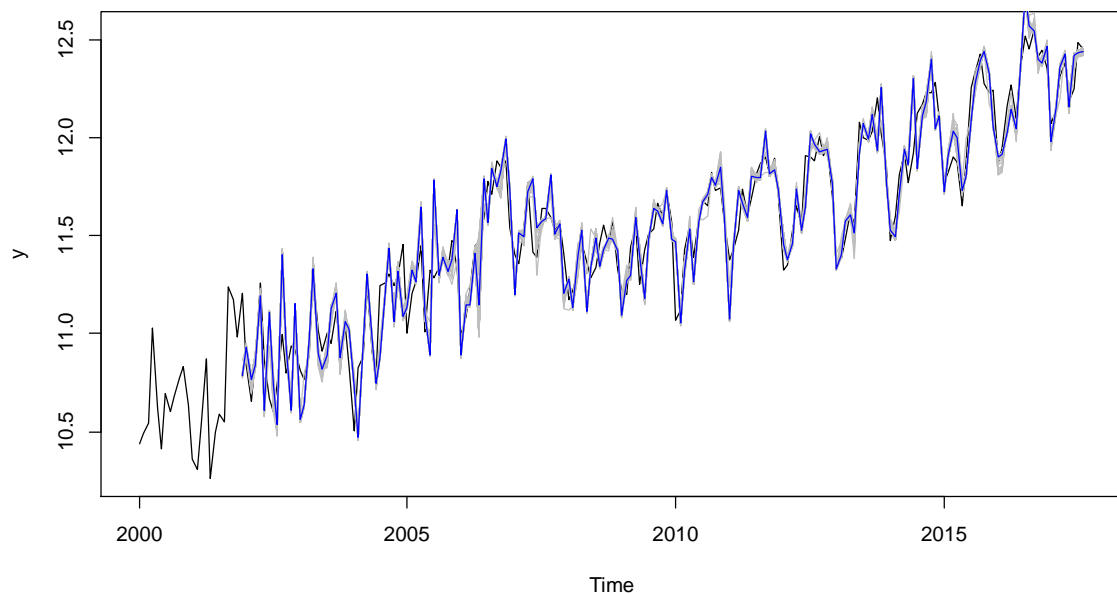


Figura:A8.9: Ajustamento do modelo **m10** RNAR(10,1,6)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m10** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set  0.1401707  0.3767322  0.3053143  1.162377  2.6261  0.5263674  1.375904
```

Seguidamente é feita a validação do modelo **m10** comparando-o com os dados do conjunto de teste com apoio da Figura A8.10 e os valores da acurácia.

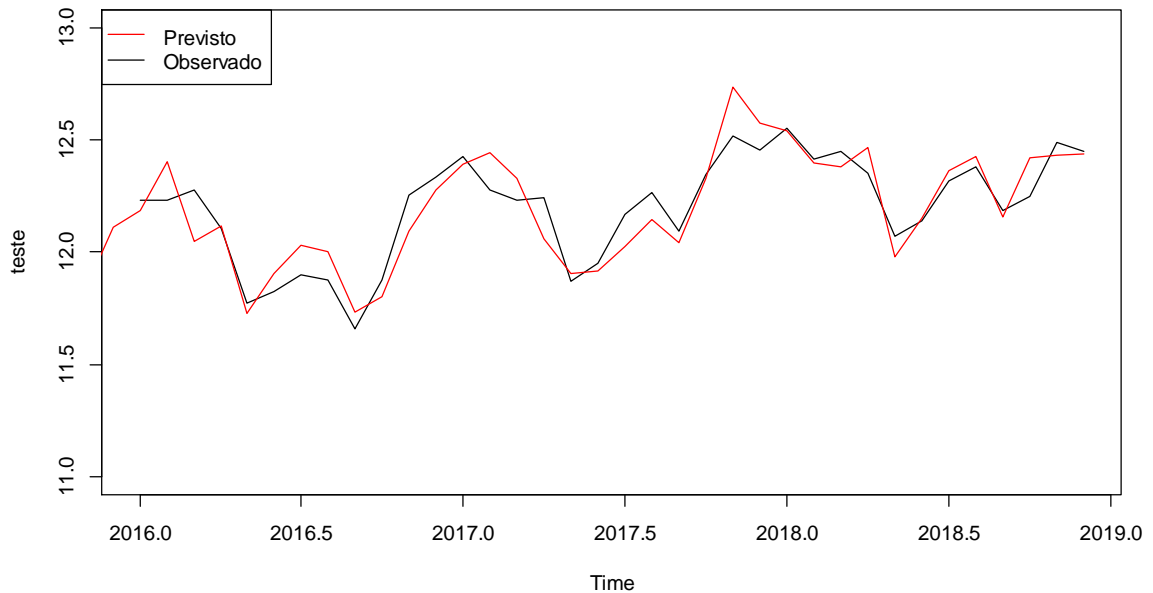


Figura:A8.10: Validação do modelo **m10** RNAR(10,1,6)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

> accuracy(dadost,rede.teste\$fitted)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.0001207956	0.1647609	0.122942	-0.0190589	1.084625	-0.2887221	0.5789354

Modelo m11: RNAR(12,1,7)[12]

Este modelo de rede neuronal de 12 *lags* de entrada e 7 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.11 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

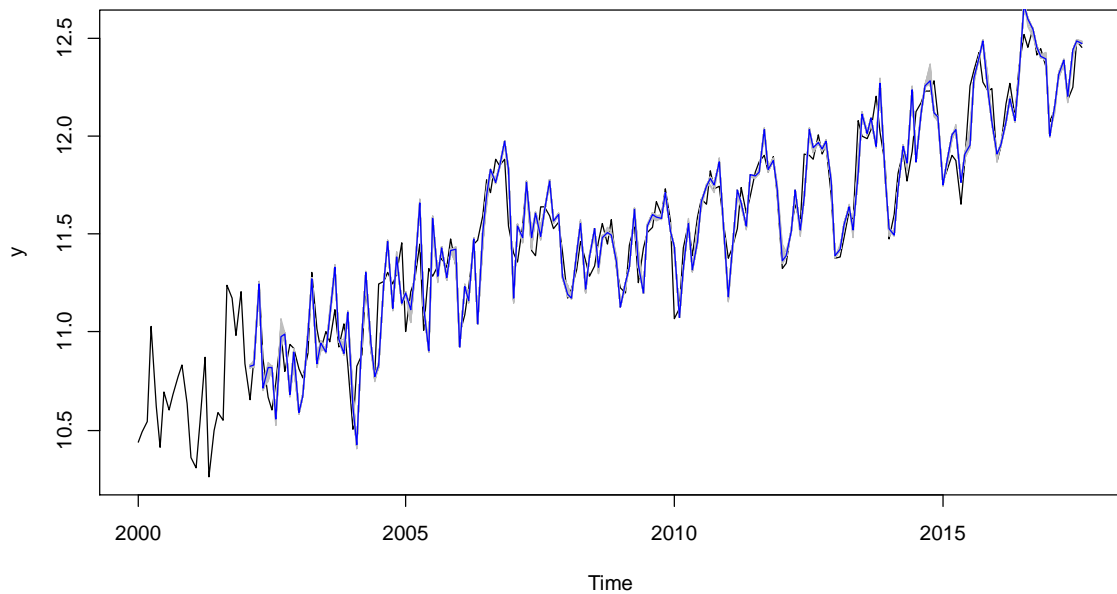


Figura:A8.11: Ajustamento do modelo **m11** RNAR(12,1,7)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m11** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

> accuracy(dadost,rede.treino\$fit)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1380919	0.3617152	0.2979164	1.147435	2.563723	0.5884511	1.50372

Seguidamente é feita a validação do modelo **m11** comparando-o com os dados do conjunto de teste com apoio da Figura A8.12 e os valores da acurácia:.



Figura:A8.12: Validação do modelo **m11** RNAR(12,1,7)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

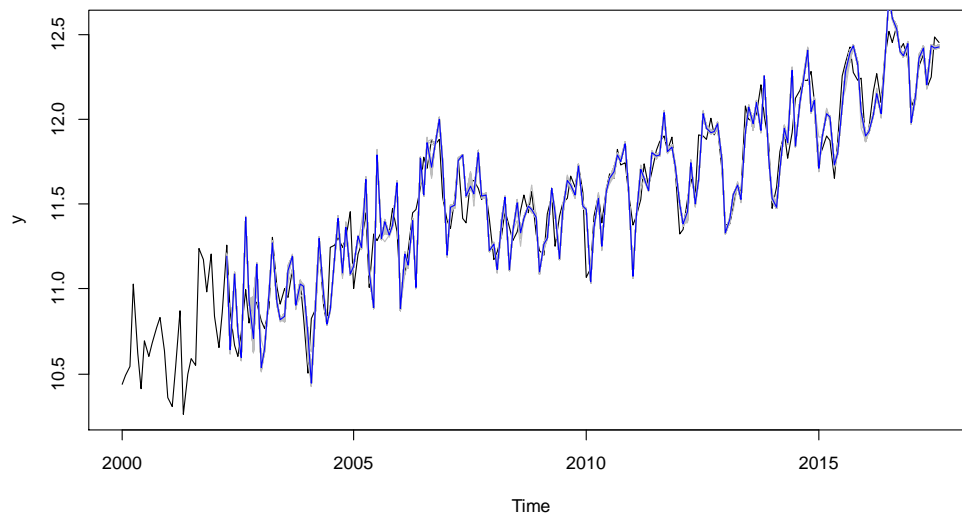
Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.002874792	0.1428481	0.1042434	-0.04038602	0.9184138	-0.2544051	0.5818922

Modelo m12: RNAR(14,1,8)[12]

Este modelo de rede neuronal de 14 *lags* de entrada e 8 neurónios na camada oculta, resultante da aplicação do comando da função *m1p()* dada anteriormente, permite a elaboração automática da Figura A8.13 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.



*Figura:A8.13: Ajustamento do modelo **m12** RNAR(14,1,8)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].*

Fonte: Elaboração própria

O ajustamento do modelo **m12** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1404468	0.3803216	0.3086291	1.162832	2.653421	0.535477	1.381759

Seguidamente é feita a validação do modelo **m12** comparando-o com os dados do conjunto de teste com apoio da Figura A8.14 e os valores da acurácia.

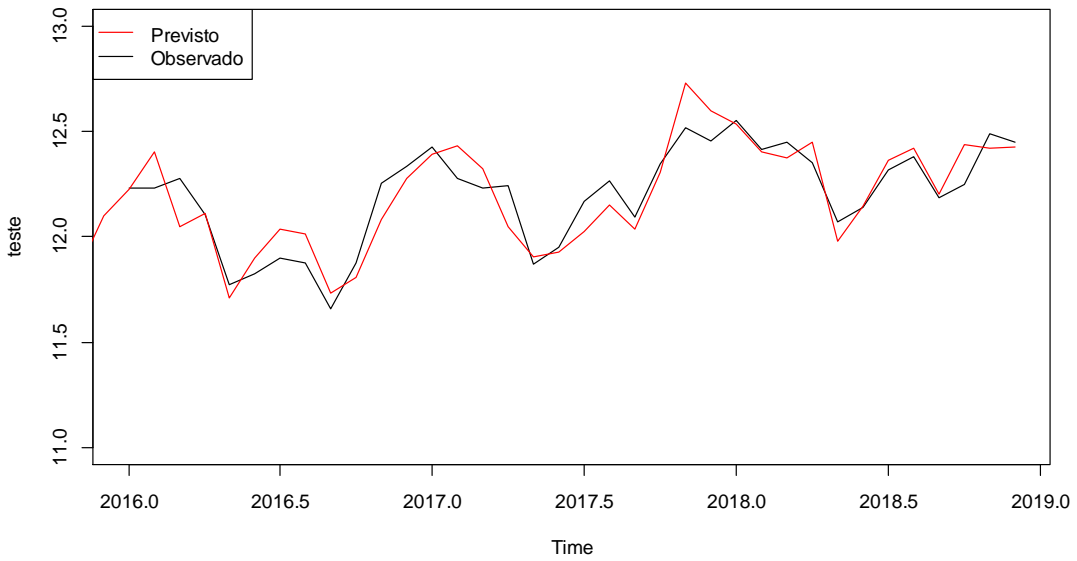


Figura:A8.14: Validação do modelo **m12** RNAR(14,1,8)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

> accuracy(dadost,rede.teste\$fitted)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.002218885	0.1642093	0.1225275	-0.037588	1.078954	-0.291928	0.5784803

Modelo m13: RNAR(16,1,9)[12]

Este modelo de rede neuronal de 16 *lags* de entrada e 9 neurónios na camada oculta, resultante da aplicação do comando da função *mip()* dada anteriormente, permite a elaboração automática da Figura A8.15 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

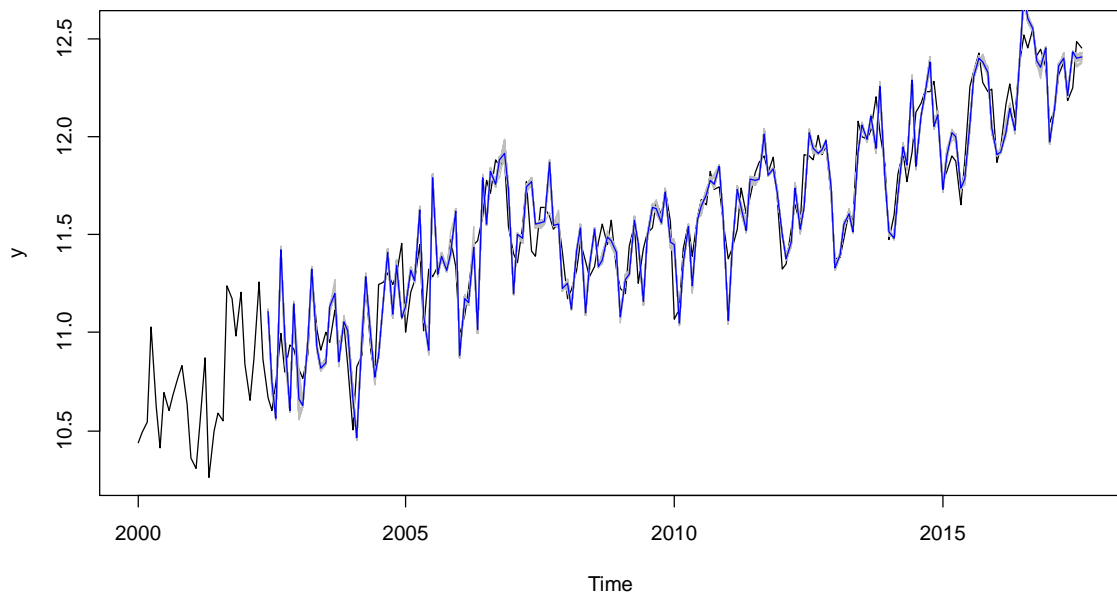


Figura:A8.15: Ajustamento do modelo **m13** RNAR(16,1,9)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m13** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia:

```
> accuracy(dadost,rede.treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.1349493 0.3774405 0.3055836 1.113579 2.62539 0.5315088 1.372168
```

Seguidamente é feita a validação do modelo **m13** comparando-o com os dados do conjunto de teste com apoio da Figura A8.16 e os valores da acurácia.



Figura:A8.16: Validação do modelo **m13** RNAR(16,1,9)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

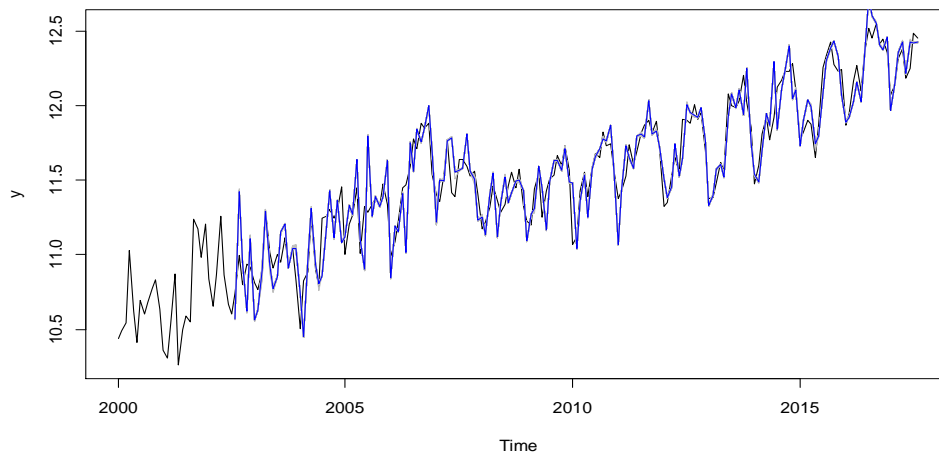
Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.004464129	0.1651288	0.1221664	-0.05668503	1.07524	-0.2945316	0.582961

Modelo m14: RNAR(18,1,10)[12]

Este modelo de rede neuronal de 18 *lags* de entrada e 10 neurónios na camada oculta, resultante da aplicação do comando da função `mlp()` dada anteriormente, permite a elaboração automática da Figura A8.17 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.



*Figura:A8.17: Ajustamento do modelo **m14** RNAR(18,1,10)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].*

Fonte: Elaboração própria

O ajustamento do modelo **m14** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1344614	0.3776711	0.3058449	1.104555	2.624004	0.5434664	1.38867

Seguidamente é feita a validação do modelo **m14** comparando-o com os dados do conjunto de teste com apoio da Figura A8.18 e os valores da acurácia.

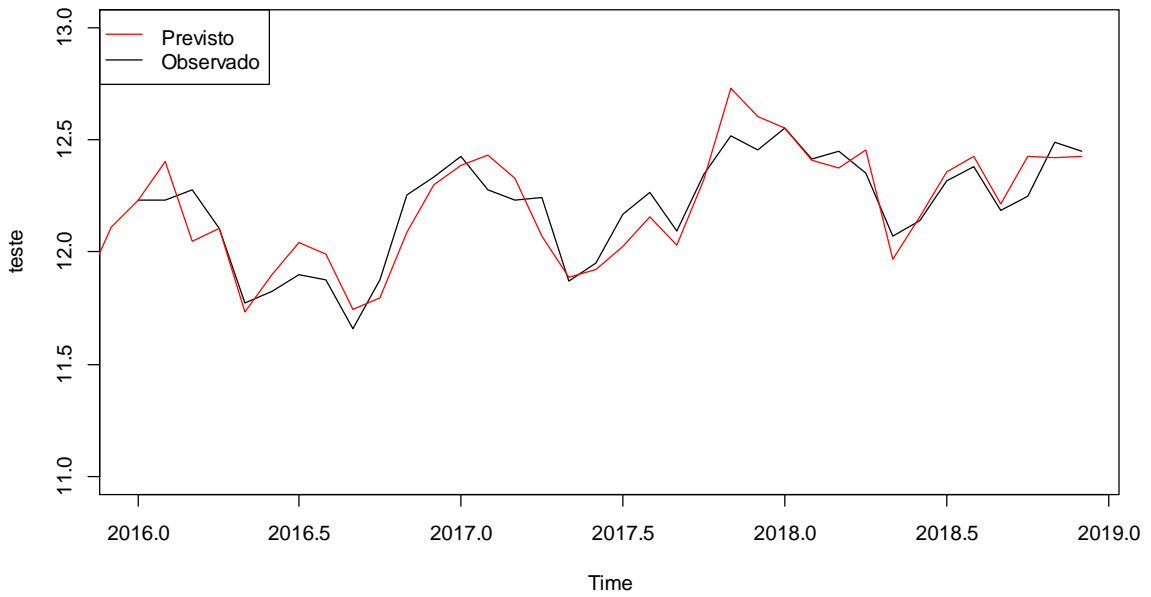


Figura:A8.18: Validação do modelo **m14** RNAR(18,1,10)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

>accuracy(dadost,rede.teste\$fitted)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	6.388907e-06	0.1657383	0.1223155	-0.01823573	1.075348	-0.2687375	0.5916798

Modelo m15: RNAR(20,1,11)[12]

Este modelo de rede neuronal de 20 *lags* de entrada e 11 neurónios na camada oculta, resultante da aplicação do comando da função *m1p()* dada anteriormente, permite a elaboração automática da Figura A8.19 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

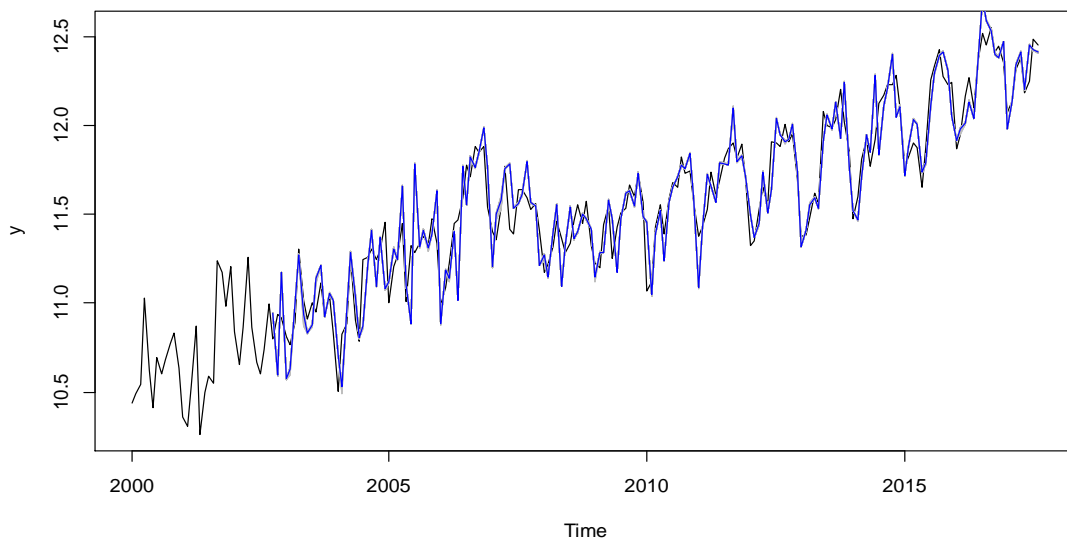


Figura:A8.19: Ajustamento do modelo **m15** RNAR(20,1,11)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m15** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
>accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1325233	0.3678325	0.3002805	1.089714	2.573162	0.5731273	1.398865

Seguidamente é feita a validação do modelo **m15** comparando-o com os dados do conjunto de teste com apoio da Figura A8.20 e os valores da acurácia.

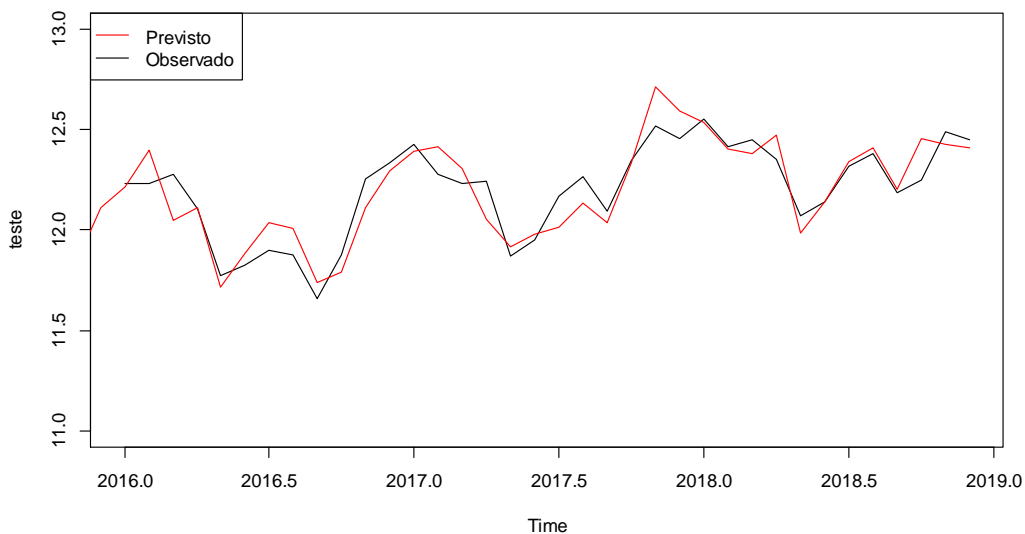


Figura:A8.20: Validação do modelo **m15** RNAR(20,1,11)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.0002453553	0.1645655	0.1216046	-0.02004417	1.06725	-0.2773661	0.5873937

Modelo m16: RNAR(22,1,12)[12]

Este modelo de rede neuronal de 22 lags de entrada e 12 neurónios na camada oculta, resultante da aplicação do comando da função `mlp()` dada anteriormente, permite a elaboração automática da Figura A8.21 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

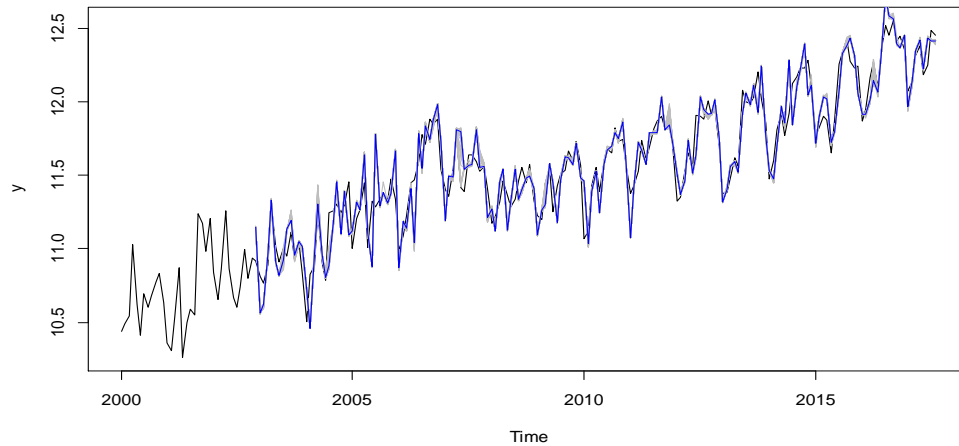


Figura:A8.21: Ajustamento do modelo **m16** RNAR(22,1,12)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m16** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.1304265 0.3696347 0.3021989 1.068619 2.588978 0.5809922 1.399948
```

Seguidamente é feita a validação do modelo **m16** comparando-o com os dados do conjunto de teste com apoio da Figura A8.22 e os valores da acurácia.



Figura:A8.22: Validação do modelo **m16** RNAR(22,1,12)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

> accuracy(dadost,rede.teste\$fitted)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.0002391167	0.1621449	0.1201	-0.01866632	1.051251	-0.2710674	0.594031

Modelo m17: RNAR(24,1,13)[12]

Este modelo de rede neuronal de 24 *lags* de entrada e 13 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.23 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

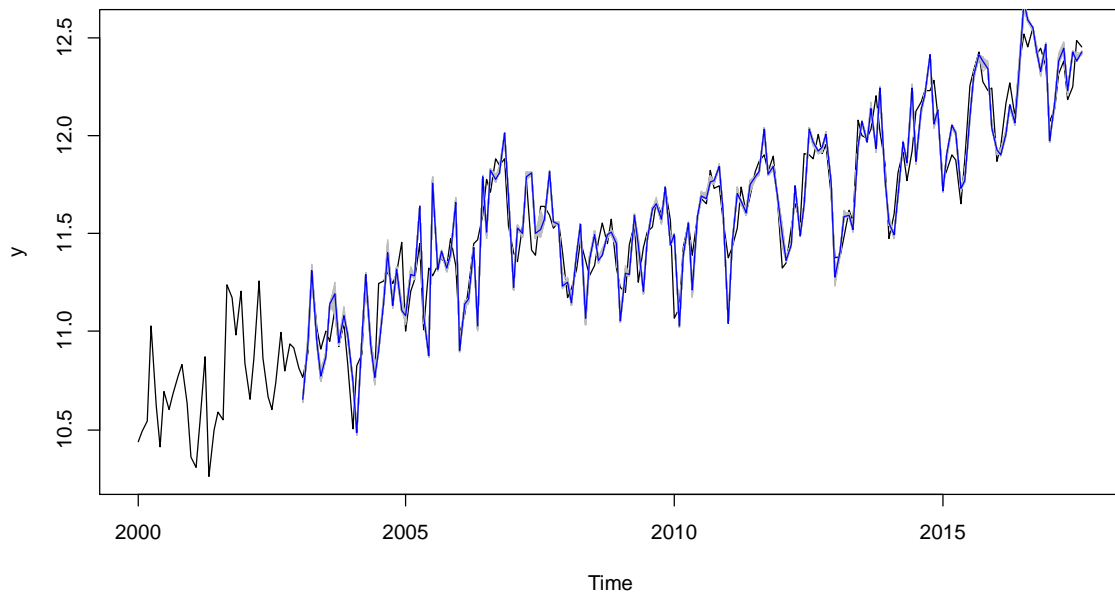


Figura:A8.23: Ajustamento do modelo **m17** RNAR(24,1,13)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m17** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

> accuracy(dadost,rede.treino\$fitted)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1321906	0.3640064	0.2975351	1.084973	2.543438	0.5951504	1.39575

Seguidamente é feita a validação do modelo **m17** comparando-o com os dados do conjunto de teste com apoio da Figura A8.24 e os valores da acurácia.



Figura:A8.24: Validação do modelo **m17** RNAR(24,1,13)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.003036492	0.157199	0.1177041	0.01118598	1.027754	-0.2862533	0.5740612

Modelo m18: RNAR(26,1,14)[12]

Este modelo de rede neuronal de 26 *lags* de entrada e 14 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.25 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

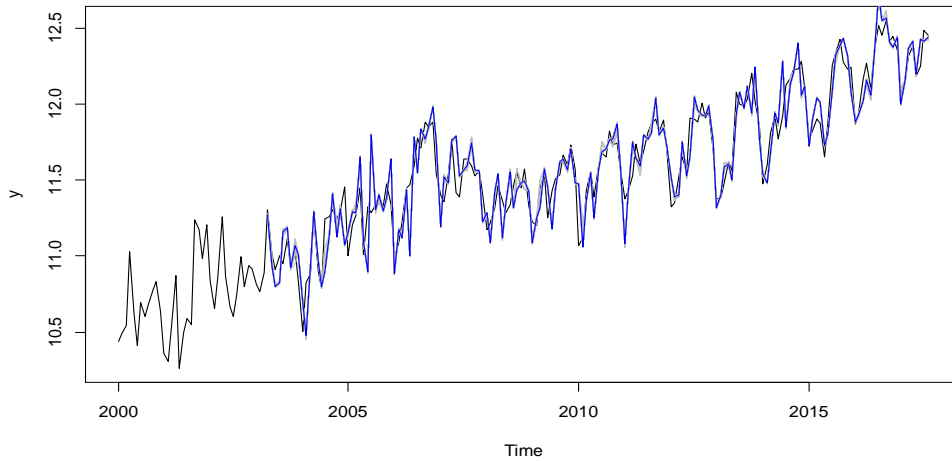


Figura:A8.25: Ajustamento do modelo **m18** RNAR(26,1,14)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m18** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1366733	0.3646572	0.2976005	1.125131	2.541785	0.6096035	1.424143

Seguidamente é feita a validação do modelo **m18** comparando-o com os dados do conjunto de teste com apoio da Figura A8.26 e os valores da acurácia.



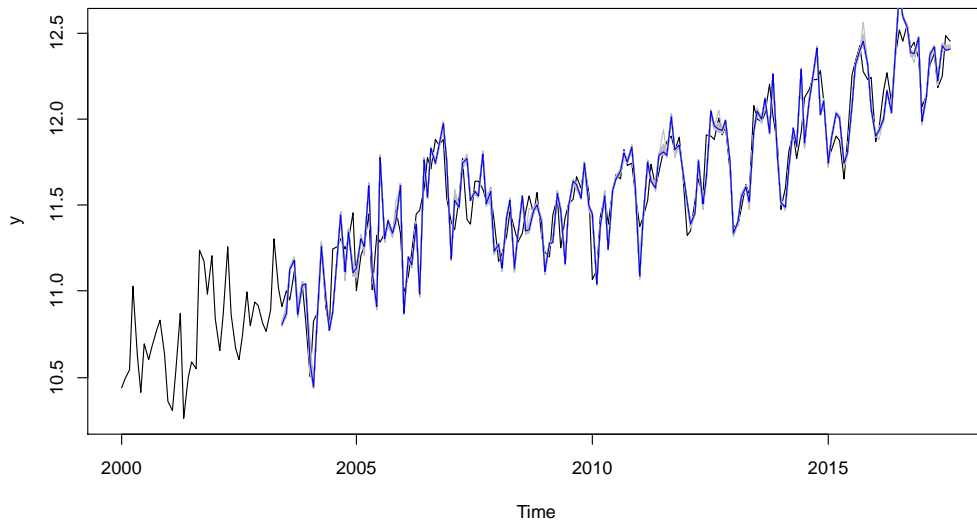
Figura:A8.26: Validação do modelo **m18** RNAR(26,1,14)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

> accuracy(dadost,rede.teste\$fitted)							
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.0004458404	0.1583547	0.117308	-0.01239367	1.025633	-0.2775414	0.5703731

Modelo m19: RNAR(28,1,15)[12]

Este modelo de rede neuronal de 28 lags de entrada e 15 neurónios na camada oculta, resultante da aplicação do comando da função `mlp()` dada anteriormente, permite a elaboração automática da Figura A8.27 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.



*Figura:A8.27: Ajustamento do modelo **m19** RNAR(28,1,15)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].*

Fonte: Elaboração própria

O ajustamento do modelo **m19** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1366276	0.3664639	0.3012434	1.12293	2.572418	0.6065073	1.427902

Seguidamente é feita a validação do modelo **m19** comparando-o com os dados do conjunto de teste com apoio da Figura A8.28 e os valores da acurácia.

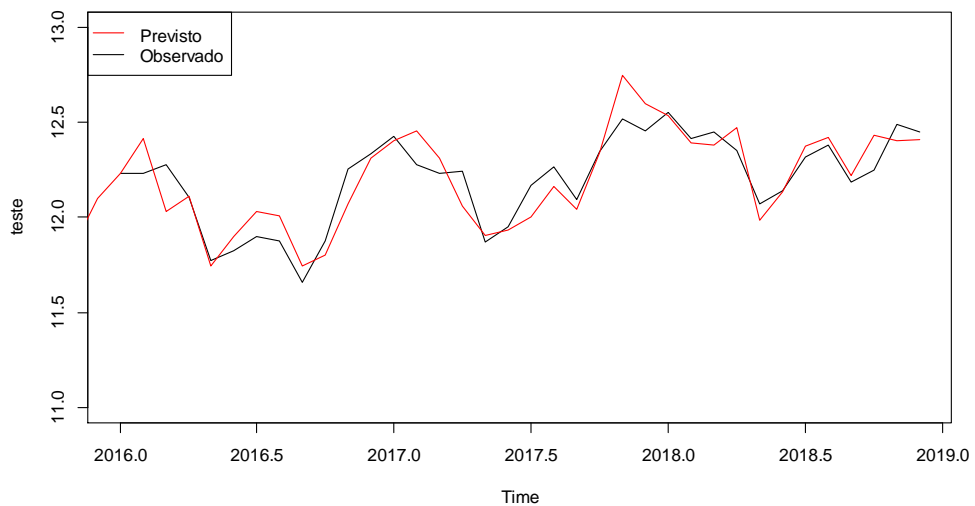


Figura:A8.28: Validação do modelo **m19** RNAR(28,1,15)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

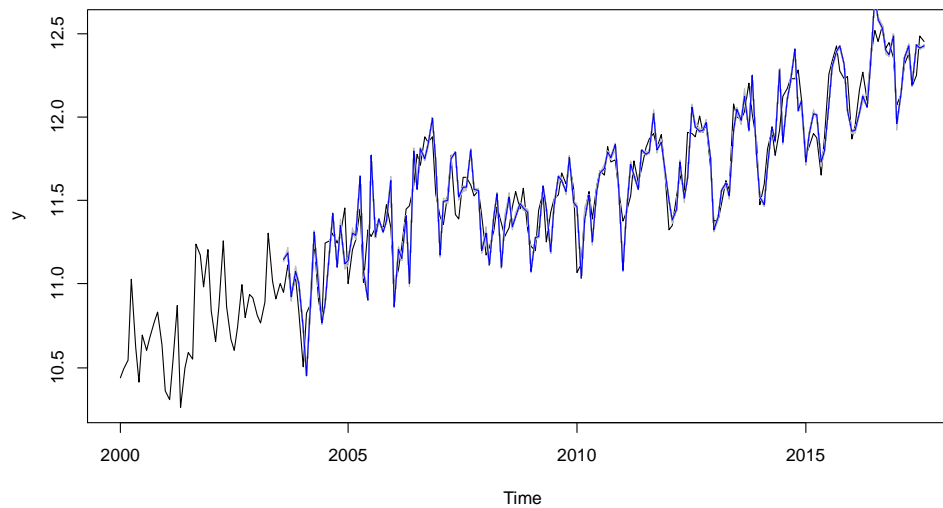
Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.001669493	0.1541444	0.1134527	-0.001090784	0.9889894	-0.2757298	0.5671134

Modelo m20: RNAR(30,1,16)[12]

Este modelo de rede neuronal de 30 *lags* de entrada e 16 neurónios na camada oculta, resultante da aplicação do comando da função `mlp()` dada anteriormente, permite a elaboração automática da Figura A8.29 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.



*Figura:A8.29: Ajustamento do modelo **m20** RNAR(30,1,16)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].*

Fonte: Elaboração própria

O ajustamento do modelo **m20** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.1358762	0.3674519	0.3011872	1.117034	2.571721	0.6060532	1.424077

Seguidamente é feita a validação do modelo **m20** comparando-o com os dados do conjunto de teste com apoio da Figura A8.30 e os valores da acurácia.



Figura:A8.30: Validação do modelo **m20** RNAR(30,1,16)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-0.0004569471	0.1561364	0.1173089	-0.0189286	1.023307	-0.2899199	0.570518

Modelo m21: RNAR(32,1,17)[12]

Este modelo de rede neuronal de 32 *lags* de entrada e 17 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.31 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

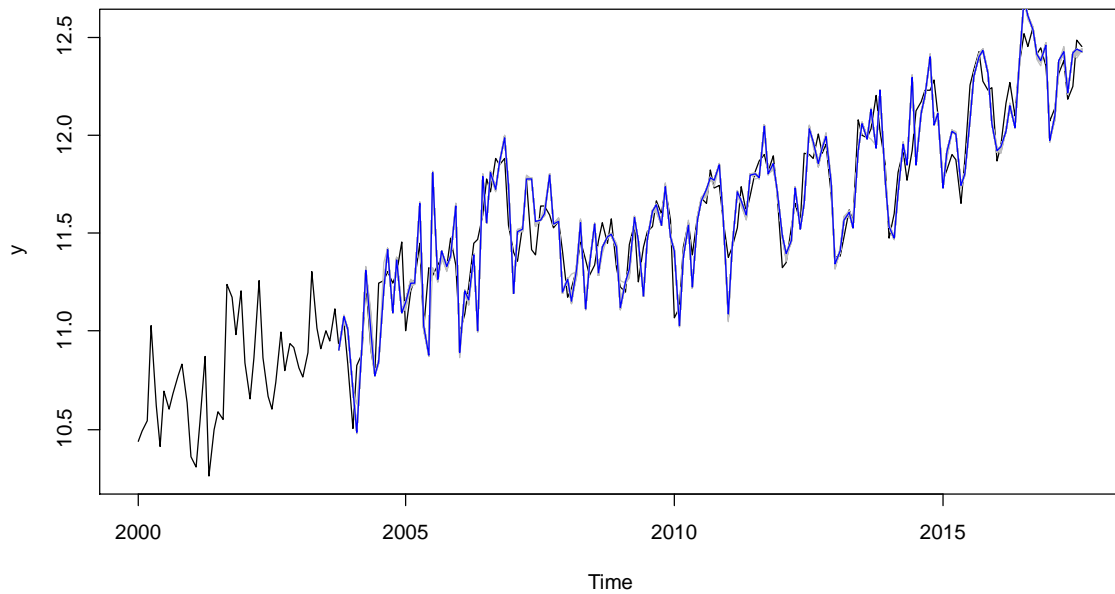


Figura:A8.31: Ajustamento do modelo **m21** RNAR(32,1,17)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m21** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.137871	0.3698161	0.3037696	1.132196	2.591462	0.6111848	1.427808

Seguidamente é feita a validação do modelo **m21** comparando-o com os dados do conjunto de teste com apoio da Figura A8.32 e os valores da acurácia.



Figura:A8.32: Validação do modelo **m21** RNAR(32,1,17)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

```
> accuracy(dadost,rede.teste$fitted)
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	0.002250831	0.1552447	0.1151495	0.004593456	1.003598	-0.260206	0.564828

Modelo m22: RNAR(34,1,18)[12]

Este modelo de rede neuronal de 34 *lags* de entrada e 18 neurónios na camada oculta, resultante da aplicação do comando da função `m1p()` dada anteriormente, permite a elaboração automática da Figura A8.33 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

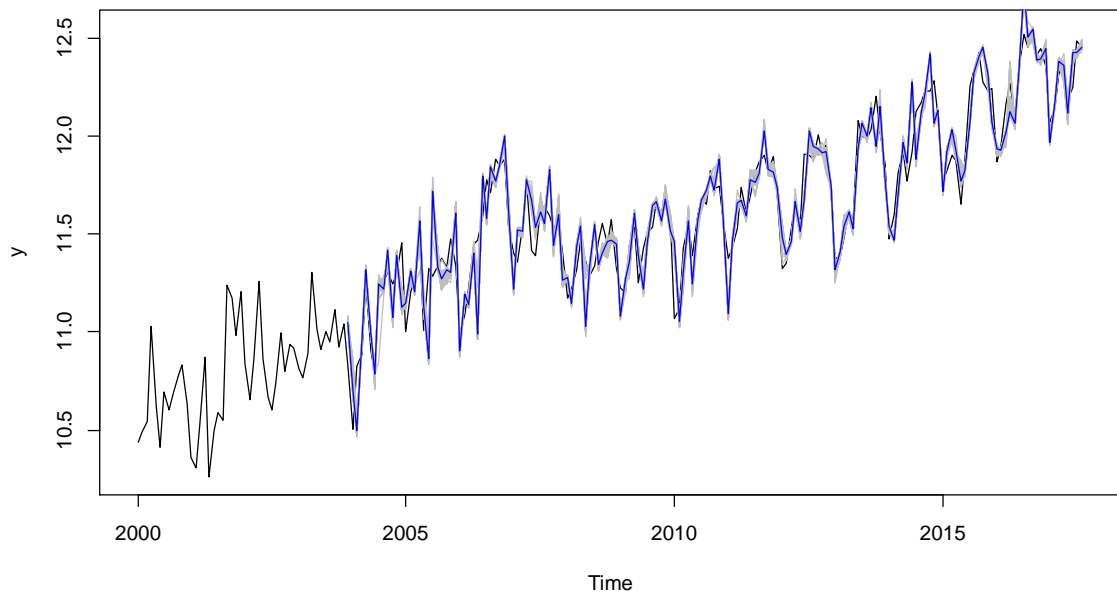


Figura:A8.34: Ajustamento do modelo **m22** RNAR(34,1,18)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

O ajustamento do modelo **m22** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.135873 0.3658066 0.3013604 1.115332 2.570319 0.6040649 1.426573
```

Seguidamente é feita a validação do modelo **m22** comparando-o com os dados do conjunto de teste com apoio da Figura A8.35 e os valores da acurácia.

```
> accuracy(dadost,rede.teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set -0.0007688965 0.1451405 0.1094947 -0.02063319 0.9537433 -0.3087611 0.5462495
```



Figura:A8.35: Validação do modelo **m22** RNAR(34,1,18)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

Modelo m23: RNAR(36,1,19)[12]

Este modelo de rede neuronal de 36 *lags* de entrada e 19 neurónios na camada oculta, resultante da aplicação do comando da função *m1p()* dada anteriormente, permite a elaboração automática da Figura A8.36 correspondente ao ajustamento do modelo comparando-o com os dados do conjunto de treino.

O ajustamento do modelo **m23** comparando-o com os dados do conjunto de treino produz os seguintes valores de acurácia.

```
> accuracy(dadost,rede.treino$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.1396601 0.3708131 0.3037732 1.14541 2.587191 0.6120512 1.441516
```

Seguidamente é feita a validação do modelo **m23** comparando-o com os dados do conjunto de teste com apoio da Figura A8.37 e os valores da acurácia.

```
> accuracy(dadost,rede.teste$fitted)
      ME      RMSE      MAE      MPE      MAPE      ACF1      Theil's U
Test set 0.0008220893 0.149091 0.1112064 -0.007926348 0.9665606 -0.2995882 0.5782896
```

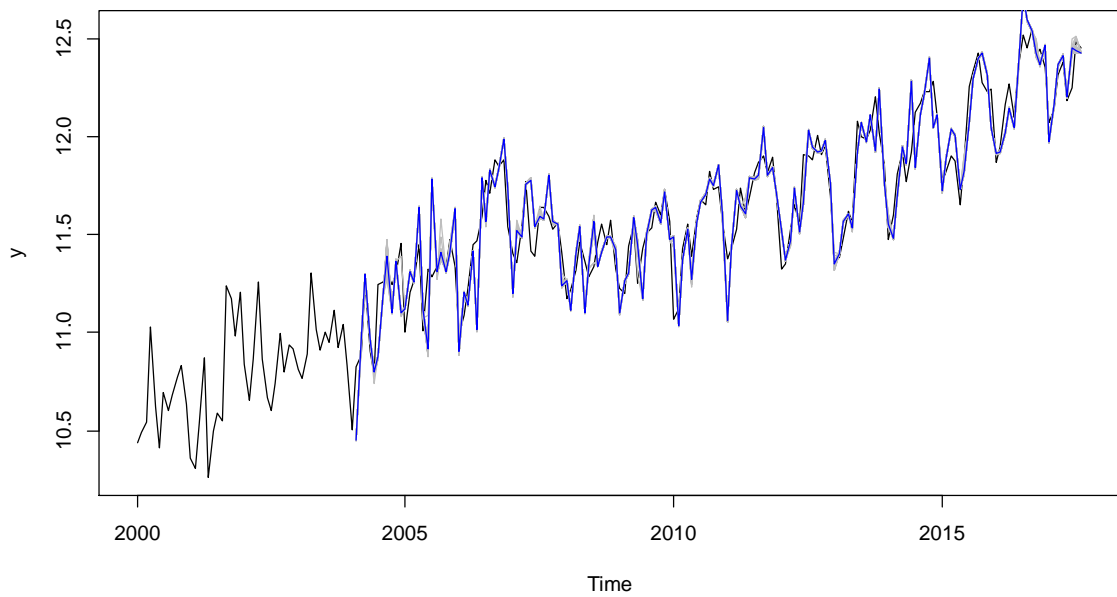


Figura:A8.36: Ajustamento do modelo **m23** RNAR(36,1,19)[12] comparando-o com os dados do conjunto de treino, [2000:01, 2015:12].

Fonte: Elaboração própria

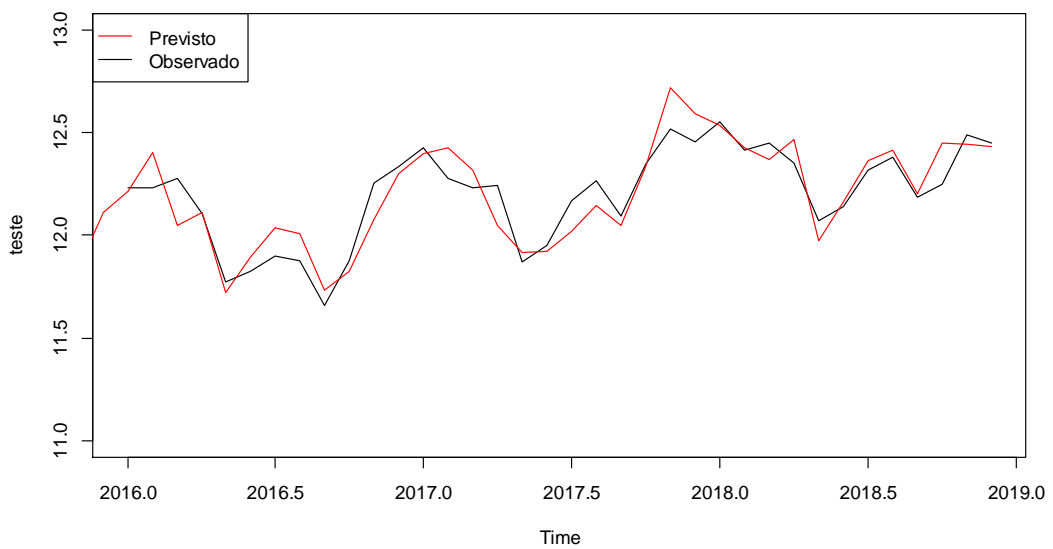


Figura:A8.37: Validação do modelo **m23** RNAR(36,1,19)[12] comparando-o com os dados do conjunto de teste, [2016:01, 2018:12].

Fonte: Elaboração própria

ANEXO IX: PUBLICAÇÕES

Application of Benford's Law to the Tourism Demand: the Case of the Island of Sal, Cape Verde

Gilberto A. Neves¹[0000-0003-1817-292X], Catarina S. Nunes²[0000-0002-8357-0994], Paula Ode-
te Fernandes³[0000-0001-8714-49010]

¹Universidade Aberta, Rua do Amial 762, 4200-055 Porto, Portugal

²Universidade Aberta, Departamento de Ciências e Tecnologia, Rua do Amial 762, 4200-055
Porto, Portugal & INEGI/LAETA, Portugal

³Applied Management Research Unit (UNIAG); Instituto Politécnico de Bragança; Campus de
Santa Apolónia, 5300-253, Bragança, Portugal

1402186@estudante.uab.pt; CatarinaS.Nunes@uab.pt; pof@ipb.pt

Abstract. This article presents Benford's Law applied for the first time to the tourism context, focusing on tourism demand. This law states that in sets of random numbers of natural events, the probability of the first digit of these numbers being 1 is approximately 30%, of being 2 is 18%, and so on, until reaching 9 with 4.6% probability. In this context, the objective is to verify if Benford's Law applies to the monthly numbers of overnight stays registered in the accommodation establishments of the Island of Sal, in the period between 2000 and 2018, to test the data reliability. This research focus on data provided by the National Statistics Institute of Cape Verde. The Chi-Square test (χ^2) was used to assess the discrepancy between the observed and expected relative frequencies. The results show that the observed χ^2 value is higher than the χ^2 critical value (5% significance level), meaning that the number of monthly overnight stays recorded in accommodation establishments on the Island of Sal does not follow Benford's Law. However, certain possible data disturbances must be considered, such as the occurrence of specific events during that time period. Other factors that could influence the results are the size of the data set and a sub notification in the data collection process. These circumstances may be the cause of the non-adaptation of the number of overnight stays to Benford's Law. The implication of this fact on the estimation of tourism demand is crucial for the development and optimization of prediction models.

Keywords: Benford's Law, Data irregularities, Overnight stays, Island of Sal.

1 Introduction

In the 19th century, Astronomer and Mathematician Simon Newcomb found a pattern that occurred with the handling of logarithmic table books. He observed, in 1881, that the first pages of the books were more worn, that is, more used, and corresponded to the pages whose numbers started with 1, 2 or 3 when compared to the last pages

whose first digits were the numbers 7, 8 or 9 [1]. He deduced, therefore, that in many quantitative phenomena, that is, in a set of numbers that appeared in nature, more numbers started with digit 1 than with digit 2, and more with 2 than with 3, and more with 3 than with 4, and more with 4 than with 5, and more with 5 than with 6, and so forth. He designed the probability distribution of the occurrence of the first digit of numbers in a set of data, thus visualizing a logarithmic distribution [1].

The topic was taken up and studied, and the theory was formally demonstrated and disseminated in 1939 by the American Physicist Frank Benford, becoming known as the Benford Law, or the First Digit Law, or the Newcomb-Benford Law, or the Law of Numbers Anomalous. Of course, for the reader who is neither a physicist nor a mathematician, this Law seems somewhat intriguing, as we can see: in a game that involves rolling a six-sided dice the probability of getting anyone side is 1/6, and the probability of getting heads or tails by tossing a coin is 1/2. In this logic, our common census would say that in a data set any number that we randomly choose would have 1 as the first digit with probability 1/9 and 2 as the first digit with the same probability, and so on. That is, all digits (1, 2 ..., 9) would have the same frequency, meaning they would occur in practically the same quantity, however an uniform distribution does not occur in many phenomena studied by Benford [2]. He analysed several sets of data from various sources, such as river lengths, number of city populations, etc., thus proving what Newcomb had discovered fifty-eight years ago.

In fact, lately Benford's Law has been used to detect evidence of irregularities and data manipulation, electoral and tax fraud, economic and psychological data. The work in [3] showed that it is possible to detect false or fraudulent reporting of accounting data using Benford's Law on data drawn from companies listed on the Johannesburg Stock Exchange. This information was then used by auditors and revenue services, but also by financial analysts as input to prediction models. Regarding electoral fraud, there are publications pointing into different directions, the work in [4] showed how in this context Benford's Law can be a problematic forensic tool. Benford's Law was successfully applied as a lie detector, presenting an advance in cognitive psychology [5]. The work in [6], showed that the relative frequencies of the first and second digits in economic research are broadly consistent with Benford's Law. Tourism demand data follows patterns of economic growth, the review in [6] supports the question that Benford's Law could be applied to this area.

It should be noted that Benford's law has empirical evidence, research and practical application; however, like any law/model it does not have a universal explanation. Therefore, Benford's law can be used as a test of whether a dataset contains real numbers [7].

In the present study, Benford's law is a preliminary approach to check the reliability of the monthly numbers of overnight stays data registered in the accommodation establishments in the Island of Sal, Cape Verde. To our knowledge there is only a recent work which explores the application of this law in the context of tourism traffic in Croatia [8]. Nevertheless, it could be very important to analyse the applicability of the law to these numbers considering the hypothesis that the number of overnight stays occurs at random. The implication of this fact on the estimation of tourism demand is crucial for the development and optimization of prediction models.

Tourism today represents a highly competitive sector and has been assumed as an alternative or even as the main driver of development for countries and localities,

whose dynamism and complexity permeate other crosscutting segments, corroborating its potential for profitability and sustainability. However, when these countries have integrated tourism products and a defined geographical area, establishing themselves as a tourism destination, they end up facing competitors for the same source markets [9].

In this paper, the concept of tourism demand is presented first. Then, an overview of tourism on the Island of Sal is presented, followed by the Benford's Law methodology and data description. The next section presents the implementation and analysis of results. Finally, the conclusions summarise the essential elements of the study.

2 Tourism Demand

Tourism demand can be split between those who travel (Effective Demand) and those who do not (Latent Demand) to a specific destination and focuses on three fundamental aspects: the existence of a destination, the availability of goods and services and the price to be paid throughout the tourism process. It can be understood or expressed as follows: number of visitors travelling to a specific destination, far from their place of residence or work, by the number of goods and services they consume at a specific price that they are willing to pay during their stay and in a given period of time [10-14].

Its measurement can be done in different ways, depending on whether the approach is, for example, physical or monetary. Physical tourism demand can be measured through volume variables, namely through 'arrivals at the borders of each country'/'number of visitors' and/or by the 'number of overnight stays'/'number of guests' in the means of accommodation. The monetary approach can be obtained through 'visitor consumption' and/or 'visitor expenses'. However, the factors that can influence tourism demand have been classified by Uysal into [15]: psychosociological, economic and exogenous. Cunha, in turn, categorized them as [13]: structural, conjunctural and psychosociological. Alternatively, sometimes there are those who present the factors divided into three types [16]: push factors, pull factors and resistance factors. Divisekera creates two categories [17]: factors that affect travel (ex-ante) and factors that influence the choice of destinations (ex-post). Many of these classified and categorised factors have variables that are reflected in the tourism destination Island of Sal.

For this study, the number of overnight stays will be used. The series overnight stays in accommodation establishments per month is considered a significant indicator of tourism activity, since it provides information about the number of visitors that have taken advantage of tourism facilities [18]. The registration of the number of monthly overnights stays from visitors in accommodation establishments on the Island of Sal is recorded in two ways: (i) for accommodation establishments that have computer support, the National Statistics Institute of Cape Verde (INE) sends a digital form to fill in the data on the movement of guests; (ii) for accommodation establishments without computer support, a paper sheet is sent to them. Once the forms are filled in, they are sent to INE, which is responsible for processing and disseminating the data.

3 Overview of Tourism in the Island of Sal

The island has been growing demographically at an average annual rate of 4.5%, and currently has a resident population of at least 39,693 inhabitants [19], with approximately 7.2% of the population residing in Cape Verde. Its economy is markedly based on tourism activity with a Gross Domestic Product (GDP) of 197,197.66 euros¹ with a predominance of the tertiary sector that generates 2/3 of the island's wealth, with a GDP/per capita of 4,496.69 euros. Transport and storage are the main economic activity with 23.6% of the island's wealth, determined mainly by the airport.

In 2019, the island had 30 new accommodation establishments, 10.6% of the total existing in the country and an accommodation capacity of 13,219 guests and 372,785 guests (45.5%) and 2,655,932 overnight stays. Of visitors arriving on the Island of Sal as a tourism destination, those from the United Kingdom are the most frequent (89,798) followed by Portugal (39,533), Belgium and the Netherlands (39,515) and Germany (32,759). There are 4,787 people employed by accommodation establishments, of which 54% are Cape Verdeans and 56% women. In terms of the number of rooms and beds, the Island of Sal is one of the islands with the highest expression (46.2% and 45.4%) at the national level [20].

Considering the tourism potential of the Island of Sal, different types of tourism products can be developed and offered. From 'sun and beach' tourism (diving, windsurfing, surfing, kitesurfing, kite beaching, shark bay, fishing and sports); nature tourism (hiking, cycling, quads, tour of the island), cultural tourism (Museum of Sal, handicrafts, *Palmeira* town, *Pedra de Lume* village, gastronomy) to health tourism (massage/spa, mud baths, baths salt, specialized health services). *Santa Maria* beach with 3 km of white sands was named by TripAdvisor as one of the 25 most beautiful beaches in the world and one of the seven wonders of Cape Verde.

In the period between 2000 and 2018, investments in sports infrastructures, combined with hotel infrastructures and the excellent natural conditions of the Island of Sal, contributed to the realization of several important events at an international level. Some of the events that mobilised foreign and national visitors were: Kite-Surf World Tour (2018: February); finals of the national championships in 2018 (Handball - May/June, Basketball - July, Karate - September, Athletics - June/July, Boxing - August, Chess - March). In addition, the opening of new internationally renowned hotels also increased the number of visitors, such as the Hilton Hotel (2017: October; 2018: January), Riu Funaná (2006: October) renewed for Club-Hotel Riu Funaná (2016), Riu Garopa (2011) renewed for Riu Palace Cabo Verde (2016: March), Meliá Lana (2016: December), Meliá Tortuga (2011: May), Meliá Dunas (2014: November), Tui Sensimar (2016: December). At the same time, some cultural events were held, such as the World of Literature (2017 and 2018) and the Santa Maria Festival (2000 to 2018), which also boosted the movement and increased the number of visitors on the island [21, 22].

¹Monetary value converted into euros using the *Banco de Portugal* currency at www.bportugal.pt on 03/05/2021.

4 Methodology

4.1 Benford's Law

Benford's law has increasingly been used as a diagnostic tool to isolate large data sets with irregularities that merit closer inspection. Performing a systematic sweep is extremely arduous; therefore, these data are a natural candidate for initial screening compared to Benford distributions [23]. Benford's Law can be defined as a collection of empirical evidence related to the frequency distribution of the leading digits in numerical data sets. The best-known version of the law states that in those data sets representing a collection of 'natural' data, the probability of seeing a particular digit in the first position is inversely related to its rank [24].

The proposal conceived by Newcomb and demonstrated by Benford shows that the first digit of numbers, in a data set, follows a logarithmic distribution, where the probability of occurrence of the first digit (d) in an universe of random numbers is given by [1, 2]:

$$P(\text{leading digit} = d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, \dots, 9. \quad (1)$$

Employing equation (1) it is then possible to determine the probability of occurrence of the numbers whose first digit is 1, 2, ..., or 9 as shown in Table 1. For some reason, it seemed that numbers beginning with digits 1 and 2 are much more frequent than numbers beginning with digits 8 and 9. For it seemed that numbers beginning with 1 and 2 occurred more frequently in nature than numbers beginning with 8 and 9 [25].

Table 1. Benford's law for the first digit of numbers in a data set.

First Digit (d)	1	2	3	4	5	6	7	8	9	Total
Probability $P(d)$	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1
(%)	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6	100

Source: Own elaboration, from the theoretical model of the Benford's Law.

Through equation (1) and Table 1, it is easy to understand that, according to Benford's Law, the distribution of leading digits in numbers is far away from uniform. A randomly selected number should begin with the digit 1 about 30% of the time: more precisely, the proportion should be 0.301, or $\log_{10}(2)$. The frequency of numbers with 2 as the first digit should be about 17.6% (obtained from $\log_{10}(3/2)$), those with 3 as the first digit should be about 12.5% (from $\log_{10}(4/3)$), and then successively until the frequency of digit 8 should be about 5.1% and for the digit 9 about 4.6% [25].

4.2 Goodness of fit

To compare the observed frequencies of the first digit of the number of monthly overnight stays in accommodation establishments on the Island of Sal with the expected frequencies according to Benford's Law, the Chi-Square frequency comparison test (χ^2) was used. This test verifies the existence (or not) of discrepancies between the observed and expected frequencies, and is given by [26]:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \quad (2)$$

where,

- o_i - observed frequencies of the i digit, where $i = 1, 2, \dots, 9$;
- e_i - expected frequencies of the i digit, where $i = 1, 2, \dots, 9$.

The null hypothesis (H_0) and the alternative hypothesis (H_1) are:

- H_0 : The number of monthly overnight stays registered in accommodation establishments on the island of Sal follows Benford's Law.
- H_1 : The number of monthly overnight stays recorded in accommodation establishments on the island of Sal does not follow Benford's Law.

4.3 Data description

To verify Benford's Law, monthly data on the number of overnight stays in accommodation establishments on Island of Sal, between 2000:01 and 2018:12, provided by the National Statistics Institute of Cape Verde (INE), were used.

Figure 1 displays the values of the data under analysis, it is visible the existence of irregular oscillations that suggest a non-stabilisation of the average and the presence a seasonality (maximum values in the winter months: between October and December, and minimum values in the summer months: between May and September), indicating that the values of overnights stays in accommodation establishments depend on the time of the year. It also shows an increasing trend after the year 2010. Although, at the end of 2007 and beginning of 2008, it showed high values when compared with the previous year. Since 2014, there has been a significant increase because of the promotion of the island and even of the Cape Verde archipelago among potential visitors in Europe.

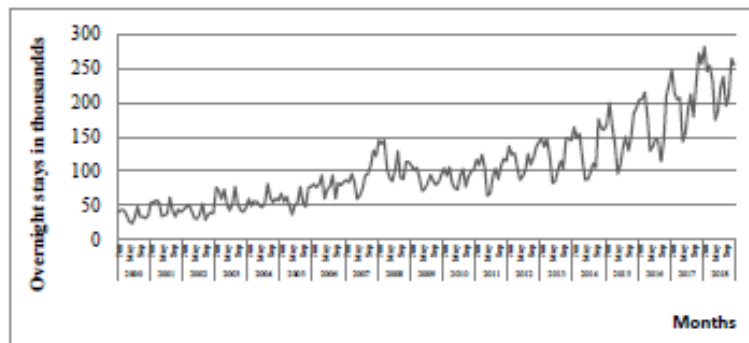


Fig. 1. Number of monthly overnight stays in accommodation establishments on Island of Sal [2000:01-2018:12]. Source: Own elaboration with data provided by INE.

5 Implementation and Analysis of Results

In this section, the Benford's law is applied to the monthly data of the number of overnight stays in accommodation establishments on the Island of Sal, between 2000:01 and 2018:12, to test the data reliability.

Based on the data presented in Table 2 and Figure 1, it can be said that:

- Digit 1 as the first digit appears for the first time when the number of overnight stays reaches six digits, which occurs in the last quarter of the year 2007. From then until 2015, the entire number of overnight stays with six digits starts with the digit 1, except for one or another exception, and this is one year after the opening of the international hotel Riu Funaná;
- Digit 2 as the first digit appears in 2000 a few times when the number of overnight stays is five digits, and it only reappears more frequently from 2015, when the opening of large international hotels occurs. In 2018, 2 is the first digit in practically all overnight stays. This may be due to the six national events and one international event that took place in seven months of that year;
- The number of overnight stays has been increasing as a result of the opening of internationally renowned hotels and occasional international and national sporting and cultural events; and, the observed percentage of digit 1 numbers exceeded the expected, since the lack of significant events lead to a slow increase of the number of overnight from 100,000 to 200,000;
- The range of values starting with the digit 1 or with digit 2 is wider than the range of values starting with the other digits. Since there is a visible increasing trend in the number of overnight stays; there are different rates of growth in the number of overnight stays, showing a possible change in tourism demand after 2014; and, there is a possible underlying oscillatory/random behaviour of the data, which could be partially explained by the Benford's law.

Table 2. Number of monthly overnight stays recorded in accommodation establishments on Island of Sal [2000:01-2018:12].

Year	Months												Total
	01	02	03	04	05	06	07	08	09	10	11	12	
2000	39,922	43,318	42,915	35,405	26,436	23,331	31,187	48,101	33,393	33,274	31,144	35,850	424,276
2001	54,739	54,243	57,621	53,946	34,270	35,893	37,967	61,574	41,054	33,305	44,007	40,282	548,901
2002	43,475	47,852	50,716	41,622	31,606	29,952	37,752	52,602	28,679	36,154	39,775	38,200	478,385
2003	76,066	71,431	59,082	73,671	50,992	42,512	52,307	77,439	52,478	42,879	40,342	46,312	685,511
2004	59,823	48,949	56,204	55,241	49,687	47,462	53,659	81,435	61,035	54,653	60,187	56,863	685,198
2005	67,218	55,353	62,544	50,586	36,578	50,323	53,355	77,364	54,350	48,221	76,610	77,480	709,982
2006	81,181	76,419	80,718	94,386	59,979	73,422	78,073	94,167	60,510	82,888	79,553	84,125	945,421
2007	87,408	83,494	96,368	83,681	59,295	65,312	75,100	93,974	95,706	108,642	130,322	122,340	1 101,642
2008	145,022	140,073	144,959	103,156	89,902	85,462	102,912	129,286	90,889	88,494	113,392	113,529	1 347,076
2009	108,226	101,263	105,014	90,069	71,362	74,777	82,424	95,145	86,053	79,853	83,930	95,184	1 073,300
2010	104,440	93,575	106,354	83,054	75,022	73,158	93,555	103,653	77,147	92,095	99,716	102,235	1 104,004
2011	116,879	109,266	124,106	105,980	64,291	67,778	92,090	103,965	88,646	107,536	118,280	115,249	1 214,066
2012	136,565	124,112	126,403	105,639	87,333	91,836	101,751	125,410	109,590	120,029	135,459	142,418	1 406,544
2013	147,746	135,967	146,822	120,591	83,032	85,033	101,882	114,924	104,153	149,004	147,795	145,033	1 481,980
2014	164,238	148,293	155,164	120,828	87,399	87,914	97,299	111,508	105,286	176,098	162,936	160,443	1 577,406
2015	168,420	200,288	165,904	140,721	96,388	109,098	134,288	150,369	129,895	150,165	184,360	193,707	1 823,603
2016	204,661	205,643	215,444	180,732	129,660	136,703	147,332	144,155	115,291	143,846	210,489	228,078	2 062,034
2017	249,162	214,312	205,295	207,714	143,093	154,921	192,532	212,943	179,214	230,459	273,066	256,776	2 519,487
2018	282,466	246,397	254,670	231,471	175,247	187,246	223,923	238,628	196,074	208,881	265,270	255,454	2 765,727

Source: Own elaboration with data provided by INE.

For this study, the calculation of the Chi-Square χ^2 was based on the sum of the nine possible digits to be the first position in the data set (Table 2). For a significance level of 5% and with 8 degrees of freedom ($gl = 9 - 1 = 8$), we have, when consulting the Chi-Square table [26], that the χ^2 critical value is 15,507.

Table 3 shows the steps performed in Excel to calculate the χ^2 value for the data set. In this case, the calculated χ^2 (= 28,940) is greater than the critical χ^2 (= 15,507), so the null hypothesis (H_0) is rejected, at a significance level of 5%. The test decision is that the number of monthly overnight stays in accommodation establishments on the Island of Sal in the period between 2000 and 2018 does not follow the Benford's Law. These results were confirmed by IBM SPSS Software, with a p-value < 0.001.

Table 3. Calculation of the χ^2 value for the number of overnight stays data set.

Digit (d)	α_i	Observed %	e_i	Expected %	$(\alpha_i - e_i)$	$(\alpha_i - e_i)^2$	$\frac{(\alpha_i - e_i)^2}{e_i}$
1	79	34.6%	68,635	30.1%	10,365	107,437	1,565
2	27	11.8%	40,149	17.6%	-13,149	172,891	4,306
3	17	7.5%	28,486	12.5%	-11,486	131,929	4,631
4	17	7.5%	22,095	9.7%	-5,095	25,964	1,175
5	23	10.1%	18,053	7.9%	4,947	24,470	1,355
6	9	3.0%	15,264	6.7%	-6,264	39,236	2,571
7	18	7.9%	13,222	5.8%	4,778	22,828	1,726
8	21	9.2%	11,663	5.1%	9,337	87,184	7,475
9	17	7.5%	10,433	4.6%	6,567	43,129	4,134
Total	228					χ^2calculated = 28,940	

Source: Own elaboration.

Figure 2 shows the discrepancy between the observed and expected relative frequencies. Although there is an approximation to the shape of the curve, the observed frequencies are above the expected frequencies (according to the Benford's Law) for the number of overnight stays beginning with digits 1, 7, 8 and 9. For the number of overnight stays beginning with digits 2 and 3 the observed frequencies are below the expected, but for the numbers beginning with digits 4, 5 and 6 the observed frequencies are close to the expected values. This shows how sensitive the Chi-Square χ^2 test is, but also how there could be sporadic disturbances on the data and even the necessity of increasing the data collection period.

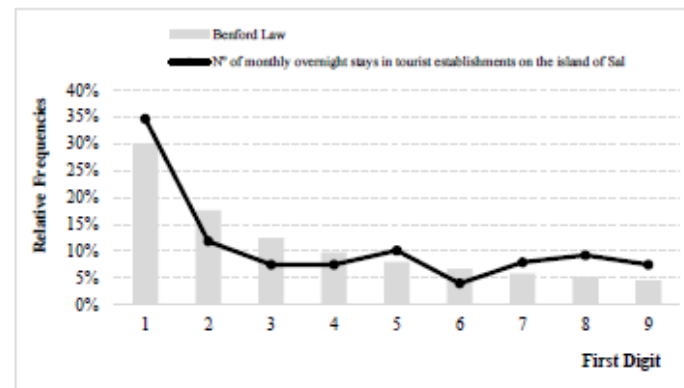


Fig. 2. Comparison of the observed and expected relative frequencies of the first digit of the number of monthly overnight stays in accommodation establishments on the Island of Sal [2000:01-2018:12]. Source: Own elaboration.

6 Conclusion

In this study, the hypothesis that the number of monthly overnight stays in accommodation establishments followed the Benford's Law was rejected, according to the data of Island of Sal. The discrepancies between the observed and expected relative frequencies were statistically significant, although there is a similarity in the shape of the curve.

Benford's law was a preliminary approach to check the reliability of the monthly numbers of overnight stays data registered in the accommodation establishments in the Island of Sal, and to our knowledge, it is its first application in the context of tourism demand in Cape Verde.

The delay in the diversification of offers in terms of infrastructure, and cultural and sporting events can lead to the number of overnight stays being outside the Benford's Law, falling short and beyond expectations. In addition, it is also necessary to consider the possibility of visitors who spend their nights in complementary accommodation, which is not accounted for in these data (such as private accommodation) and could make a difference in the results special for national visitors.

A recent study [8], Benford's Law was applied to the number of domestic and foreign overnight stays, in a monthly series over a period of three years in Croatia. That study, only the domestic overnight stays did not follow Benford's Law. However, the authors did not identify the causes of adaptation and non-adaptation of the number of overnight stays to Benford's Law, they assumed that this occurred mainly due to the lack of accuracy and absence of records in tourist accommodations.

In the Island of Sal there are still other legal tourist accommodations that, not being known to INE, do not provide data for official statistics. This could have an impact on

the adaptation of the Benford's Law and is in line with the conclusions in [8]. In this sense, it is necessary that the local government, through the Tourism Office, continues the effort to update the information on these accommodation providers and that the data in reported not just for statistical purposes but to enhance the tourism experience of Island.

In contrast with the work in [8], the data for the Island of Sal is not separated into domestic and foreign overnight stays and the size of the database is approximately six times larger, which enables a detailed data comparison. One aspect that could impact on the adaptation of Benford's Law to the data from this work, is the fact that the data does not span over a large order of magnitudes.

There are aspects that could be considered in future research mainly: incorporating more data and data from other islands in the archipelago; quantify the magnitude of the impact of specific events; look at the rate of change of overnight stays; discriminate between national and foreign tourists' overnight stays.

The main conclusion of this study points out that the results obtained not only support the reliability of secondary data (from INE), but also suggest that the baseline number of monthly overnight stays in accommodation establishments can be simulated using Benford's law and events can be considered as possible disturbances to the law.

Tourism today represents a highly competitive sector and has been assumed as an alternative or even as the main driver of development for countries and localities, whose dynamism and complexity permeate other crosscutting segments, corroborating its potential for profitability and sustainability.

Nevertheless, it was important to analyse the applicability of the law to data considering the hypothesis that the number of overnight stays occurs at random. The implication of this fact on the estimation of tourism demand is crucial for the future development and optimization of prediction models.

7 Acknowledgement

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support by national funds FCT/MCTES to UNIAG, under Project no. UIDB/04752/2020 and to INEGI under LAETA project UIDB/5022/2020.

G. Neves would also like to acknowledge the Sal City Council (*Câmara Municipal do Sal*) for their support of the PhD Scholarship.

8 References

1. Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. *American Journal of Math*, 4(1), 39-40 (1881).
2. Benford, F.: The law of anomalous numbers. In *Proceedings of the American Philosophical Society* 78(4), 551-572 (1938).
3. Saville, A.: Using Benford's Law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange. *South African Journal of Economic and Management Sciences*, 9(3), 341-354 (2014).
4. Deckert, J., Myagkov, M., & Ordeshook, P.: Benford's Law and the Detection of Election Fraud. *Political Analysis*, 19(3), 245-268 (2011).
5. Gauvrit, N., Houillon, J.C., Delahaye, J.P.: Generalized Benford's law as a lie detector. *Advances in Cognitive Psychology*, 13(2), 121-127 (2017).
6. Günnel, S., Tödter, K.: Does Benford's Law Hold in Economic Research and Forecasting?. *Bundesbank Series 1 Discussion Paper No. 2007*, 32 (2007).
7. Gabaix, X.: Power Laws in Economics: An Introduction. *Journal of Economic Perspectives*, 30(1), 185-206 (2016).
8. Matakovic, H.: Tourism Data in Croatia Assessed by Benford's Law. *Journal of Public Administration, Finance and Law*, 19, 166-184 (2021).
9. Neves G.A., Nunes C.S., Fernandes P.O.: Challenges and Opportunities for Island Tourist Destinations: The Case of the Island of Sal, Cape Verde. In: Abreu A., Liberato D., González E.A., Garcia Ojeda J.C. (eds) *Advances in Tourism, Technology and Systems. ICOTTS 2020. Smart Innovation, Systems and Technologies*, vol 209, p. 498-508. Springer, Singapore (2021).
10. Pearce, D.G.: *Tourism Development*. 1st ed. John Wiley, New York (1989).
11. Song, H., Witt, S.F.: *Tourism demand modelling and forecasting: Modern econometric approaches*. 1st ed. Taylor & Francis, New York (2000).
12. Cooper, C., Fletcher, J., Wanhill, S., Gilbert, D., Sheperd, R.: *Turismo, Principios e Prática*. 3.^a Ed.; Bookman, Porto Alegre (2007).
13. Cunha, L., Abrantes A.: *Introdução ao Turismo*, 6th ed. Lidel, Lisboa (2019).
14. Matias, A.: *Economia do turismo: teoria e prática*. 1st ed. Instituto Piaget, Lisboa (2007).
15. Uysal, M.: The determinants of tourism demand: A theoretical perspective. In D. Ioannides & K. G. Debbage (Eds.), *The economic geography of the tourist industry: A supply-side analysis* (pp. 79-95). New York: Routledge. (1998).
16. ETC/UNWTO: *Handbook on Tourism Forecasting Methodologies*. UNWTO, Madrid (2008).
17. Divisekera, S.: Tourism Demand Models: Concepts and Theories. In: C. A., Tisdell (Eds.), *Handbook of Tourism Economics: Analysis, New applications and Case studies*. (pp. 33-66). Hackensack, NJ: World Scientific Publishing Co.Pte.Ltd. (2013).
18. Fernandes, P.O., Teixeira, J., Ferreira, J., Azevedo, S.: Modelling Tourism Demand: A Comparative Study between Artificial Neural Networks and the Box Jenkins Methodology. *Romanian Journal of Economic Forecasting*, 5(3), 30-50 (2008).
19. INE: *Estatísticas das Famílias e Condições de Vida – Inquérito Multi-Objetivo Contínuo 2019*. <https://ine.cv>, last accessed 2021/03/01.
20. INE: *Estatísticas do Turismo – Ano de 2019*. <https://ine.cv>, last accessed 2021/03/01.
21. CMS: *Câmara Municipal do Sal – Relatório de Atividade* (2017).
22. CMS: *Câmara Municipal do Sal – Relatório de Atividade* (2018).
23. Cho, W., Gaines, B.: Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance. *The American Statistician*, 61(3), 218-223 (2007).
24. Kopczewski, T., Okhrimenko, I.: Playing with Benford's Law. *e-mentor*, 80(3), 34-44 (2019).
25. Fewster, R.M.: A Simple Explanation of Benford's Law. *The American Statistician* 63(1), 26-32 (2009).
26. Ross, S.: *Probability and Statistics for Engineers and Scientists*. 4th ed. Elsevier Academic Press, London, UK (2009).

Seasonal Autoregressive Integrated Moving Average Time Series Model for Tourism Demand: the Case of Sal Island, Cape Verde

Gilberto A. Neves¹{0000-0003-1817-292X}, Catarina S. Nunes²{0000-0002-8357-0994}, and Paula Odete Fernandes³{0000-0001-8714-49010}

¹Universidade Aberta, Rua do Amial 762, 4200-055 Porto, Portugal

²Universidade Aberta, Departamento de Ciências e Tecnologia, Rua do Amial 762, 4200-055 Porto, Portugal & DNEGL/LAETA, Portugal

³UNIAG, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253, Bragança, Portugal

1402186@estudante.uab.pt; CatarinaS.Nunes@uab.pt; pof@ipb.pt

Abstract. This article appears as an essential contribution for decision-makers in the Cape Verdean tourism sector given the impact that the number of overnight stays has on the economy of the country and the Sal Island, which until 2018 had been increasing every year. Since seasonality is a strong feature of the island's tourism, decision-makers are interested in knowing the seasonal variation in tourism demand. Thus, this study focused on the application of the Box-Jenkins method to the time series of the monthly number of nights stays in tourist establishments on the Sal Island, Cape Verde, over the period from January 2000 to December 2018, to find a model that better describes the series and with good forecast results for the year 2019. Several SARIMA models were studied using the Box-Jenkins method, with the SARIMA(1,1,1)(0,1,1)₁₂ and the SARIMA(2,1,0)(0,1,1)₁₂ demonstrating the best predictive performance in the test phase. However, in forecasting the series for the year 2019 the SARIMA(2,1,0)(0,1,1)₁₂ achieved the best results with a MAPE=8.78%. This model can be used to simulate and analyse the number of overnight stays that be expected on the Island, if the tourism sector were not affected by the pandemic caused by COVID-19.

Keywords: Time Series, SARIMA models, Number of overnight stays, Sal Island.

1 Introduction

The term tourist describes any person who travels to a place different from their usual environment for less than twelve consecutive months and whose main purpose of the trip is any activity within the destination, as long as it is not paid, thus generating the

so-called tourism demand [1]. Different factors influence tourism demand, namely economic, social, technological and political factors that take place in the “Tourism Destination”. This demand has been the focus of many studies considering its analysis and forecast accompanying the growth of this sector. To get an idea, 420 studies were published from 1960 to 2002, many of them on the modelling and forecasting of tourism demand [2]. From simple to more complex approaches, time series methods are one of those that have marked this field of research since the 1980s. Models based on dynamic regressions have also started to become popular. Several reviews show the advances in tourism demand modelling from 1960 to 2018 [2-6].

Most of the studies reviewed used time series models, a total of 202 in which 34% used the Box Jenkins methods [5]. Of the 211 studies published the majority (55.9%) used time series models, of which 60% were ARIMA models [3]. However, seasonal versions of the ARIMA models and specifically the SARIMA, were used in several studies, according to [5], revealing a good performance and increasing its application. The SARIMA's popularity is because many decision-makers are interested in knowing the seasonal variation of tourism demand [2].

In this research, it will be applied the Box-Jenkins method to analyse the time series of the number of monthly overnights stays in tourist establishments on Sal Island, from January 2000 to December 2018, to identify the model that adjusts to the data and presents a better prediction performance. A brief review of the tourism demand concept will be presented; a brief description of the state of tourism on the Island; the concept of seasonality as well as the SARIMA model; and finally, the conclusions summarizing the study.

2 Tourism Demand

Tourism demand can be split between those who travel (Effective Demand) and those who do not (Latent Demand) to a specific destination and focuses on three fundamental aspects: the existence of a destination, the availability of goods and services and the price to be paid throughout the tourism process. It can be understood or expressed as follows: the number of visitors travelling to a specific destination, far from their place of residence or work, by the number of goods and services they consume at a specific price that they are willing to pay during their stay and in a given period of time [7-11]. More details are in [12].

For this study, the number of overnight stays will be used. The monthly series of overnight stays in accommodation establishments is considered a significant indicator of tourism activity since it provides information about the number of visitors that have taken advantage of tourism facilities [13]. The registration of the number of monthly overnights stays from visitors in accommodation establishments on the Sal Island is recorded in two ways: (i) for accommodation establishments that have computer support, the National Statistics Institute of Cape Verde (INE) sends a digital form to fill in the data on the movement of guests; (ii) for accommodation establishments without computer support, a paper sheet is sent to them. Once the forms are filled in, they are sent to INE, which is responsible for processing and disseminating the data.

3 Overview of Tourism in the Sal Island

The overview of tourism in Sal Island can be consulted in [12]. In the current research, it was added the facts that in the period between 2000 and 2018, investments in sports infrastructures, combined with hotel infrastructures and the excellent natural conditions of the Sal Island, contributed to the realization of several important events at an international level. Some of the events that mobilised foreign and national visitors were: Kite-Surf World Tour (2018: February); finals of the national championships in 2018 (Handball - May/June, Basketball - July, Karate - September, Athletics - June/July, Boxing - August, Chess - March). In addition, the opening of new internationally renowned hotels also increased the number of visitors, such as the Hilton Hotel (2017: October; 2018: January), Riu Funaná (2006: October) renewed for Club-Hotel Riu Funaná (2016), Riu Garopa (2011) renewed for Riu Palace Cape Verde (2016: March), Meliá Lana (2016: December), Meliá Tortuga (2011: May), Meliá Dumas (2014: November), Tui Sensimar (2016: December). At the same time, some cultural events were held, such as the World of Literature (2017 and 2018) and the Santa Maria Festival (2000 to 2018), which also boosted the movement and increased the number of visitors on the island [14, 15].

4 Methodology

4.1 Seasonality of the Time Series

The seasonality of a time series can be deterministic (the seasonal pattern is regular and stable) or stochastic (the seasonal pattern varies with time) [16]. The deterministic approach is the most considered in traditional forecasting methods in the tourist context, but more recent methods already suggest stochastic seasonality in the modelling of tourism demand forecasting, especially the Box-Jenkins method and its SARIMA model [17].

In the tourism context, seasonality can be considered as any systematic ups and downs over time in specific tourism activities, including tourist overnight stays [18, 19]. A permanent feature of seasonality is the sudden change in demand in a given period of the year [20]. Also points out two important causes of seasonality in tourism, namely: natural causes, as a result of regular variations in climatic conditions, and institutional causes, as a result of human decisions in a combination of different factors [19]. In addition, there are different areas of impact such as: economic, socio-cultural and ecological [21].

4.2 The SARIMA Model

The SARIMA model is an extension of the ARIMA model, known as a Box-Jenkins methodology, which integrates autoregressive (AR) and moving average (MA) models [22]. SARIMA, known as seasonal ARIMA, incorporates non-seasonal and seasonal factors into a multiplicative model in the form of $(p, d, q) \times (P, D, Q)_s$, where p is the non-seasonal autoregressive order; d is the non-seasonal differentiation; q is the order

of the nonseasonal moving average; P is the seasonal autoregressive order; D is the seasonal differentiation; Q is the order of the seasonal moving average; and, s is the seasonal pattern repeat time interval. According to [23] for a monthly time series, $s=12$. This model is popular because of its ability to handle both stationary and non-stationary series. In the case of non-stationary time series, the variable values are taken as the first differences of the original data to estimate the model. Formally, a time series $\{X_t\}$ is said to follow the SARIMA model if defined as [24]:

$$\Phi(B^s)\Phi(B)(1-B^s)^D(1-B)^dX_t = \Theta(B^s)\Theta(B)\theta_t, \quad t \in T \quad (1)$$

where,

$\Phi(B) = 1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p$ is the non-seasonal autoregressive operator; $\Phi(B^s) = 1 - \phi_1B^s - \phi_2B^{2s} - \dots - \phi_pB^{ps}$ is the seasonal autoregressive operator; $\Theta(B) = 1 + \theta_1B + \theta_2B^2 + \dots + \theta_qB^q$ is the nonseasonal moving average operator; $\Theta(B^s) = 1 + \theta_1B^s + \theta_2B^{2s} + \dots + \theta_qB^{qs}$ is the seasonal moving average operator; $(1-B^s)^D(1-B)^d$ is the seasonal differentiation of order d and non-seasonal differentiation of order D ; X_t is the value of the series observed at a given time t .

4.3 Autocorrelation

The sample autocorrelation function (ACF) of X_t is generally used to determine whether or not the time series is stationary. A stationary time series will cause the autocorrelation to drop to zero quickly, but for a non-stationary time series, it will drop slowly. The ACF has a greater effect on the stationary series and is given by expression (2), considering the pairs of values $(x_1, x_2), \dots, (x_{n-1}, x_n)$, thus obtaining the correlation between x_t and x_{t+1} [25]:

$$r_1 = \frac{\sum_{t=1}^{n-1}(x_t - \bar{x}_1)(x_{t+1} - \bar{x}_2)}{\sqrt{\sum_{t=1}^{n-1}(x_t - \bar{x}_1)^2} \sqrt{\sum_{t=1}^{n-1}(x_{t+1} - \bar{x}_2)^2}} \quad (2)$$

where the sample means are given by

$$\bar{x}_1 = \sum_{t=1}^{n-1} x_t / (n-1) \text{ and } \bar{x}_2 = \sum_{t=2}^n x_t / (n-1) \quad (3)$$

The partial autocorrelation function (PACF) is the correlation of the time series with a delay of itself, removing the linear dependence between lags. It is defined by the correlation between the variables X_t and X_{t+k} without any influence of the intermediate variables $X_{t+k'}, \dots, X_{t+k-1}$ and is given by (4).

$$\rho(k, k) = \frac{\text{Cov}(X_t, X_{t+k} | X_{t+1}, \dots, X_{t+k-1})}{\sqrt{\text{Var}(X_t | X_{t+1}, \dots, X_{t+k-1})} \sqrt{\text{Var}(X_{t+k} | X_{t+1}, \dots, X_{t+k-1})}} \quad (4)$$

The correlogram is a graph that represents the ACF and PACF functions and presents the k first autocorrelation coefficients as a function of k [23]. It allows identifying some characteristics of the time series, such as randomness, stationarity and seasonality. In

case of randomness, it is expected that $r_k = 0$. Dashed horizontal lines, usually indicate the lower and upper limits calculated for a given confidence level; for example, 95% confidence limits are defined by $\pm 2/\sqrt{n}$ [23]. Thus, all ACF or PACF values in each k , that are outside the limits are considered non-zero, and those that are inside are considered null. The ACF interpretation is also useful in identifying the parameters of the non-seasonal orders p, d, q and the seasonal orders P, D, Q .

4.4 Stationarity Analysis

The stationarity of a time series pertains to the following assumption “[...] it develops in time randomly around a constant mean, reflecting some form of stable equilibrium” [22, p.4]. Chatfield states that most time series are stochastic [26], but, Moretin and Toloï [27], state that series are stationary if, at different phases, maintain the same mean, variance and autocovariance. The temporal behaviour can be analysed graphically, taking into account the data variation over time around a horizontal line. If non-stationarity is verified, the integration order (the number of necessary differences) is used to transform the series. This consists of taking, generally, one (ΔX_t), two ($\Delta^2 X_t$) or $\Delta^d X_t$ successive differences from the original series until obtaining a stationary series. The 1st difference is made if the series is growing at a constant rate and the 2nd difference is used if the series is growing at a decreasing rate [16]. Quoting [28], in the case of the seasonal model, SARIMA, the first difference is given by (5).

$$\Delta_s X_t = X_t - X_{t-s} \tag{5}$$

However, this technique has its disadvantage, because data could be lost, that is, when we apply, for example, differentiation in a monthly series of size n whose seasonal component is of period s , we will have a series of seasonal differences of size $n - s$ [23].

5 Results

In this section, the time series monthly number of overnight stays in accommodation establishments on Sal Island, between 2000:01 and 2018:12, is analysed (Figure 1). The R Software version 4.0.5 packages, functions and commands, is used in this research. Initially, a graph of the series under study is made; trend and seasonality removed; identified the parameters of non-seasonal order p, d, q and the seasonal order P, D, Q ; temporary seasonal models identified, and their parameters estimated; the diagnosis of the models was carried out and the best model was chosen based on the Akaike Information Criteria (AIC) and training and testing error (MAPE-Mean absolute error).

When analysing the behaviour of the series, a growing trend is observed (confirmed by the Cox-Stuart test with a $p\text{-value}=2.2e-16<0.05$) as well as the presence of seasonality (confirmed by the Weibel-Ollech test with a $p\text{-value} = 0.00<0.05$). It is visible how the variance grows over time. Such a trend leads to the specification of a non-stationary model. This non-stationarity was observed in the simple correlogram of the series and

confirmed with the application of the Dickey-Fuller unit root test with $p\text{-value}=0.05773>0.05$.

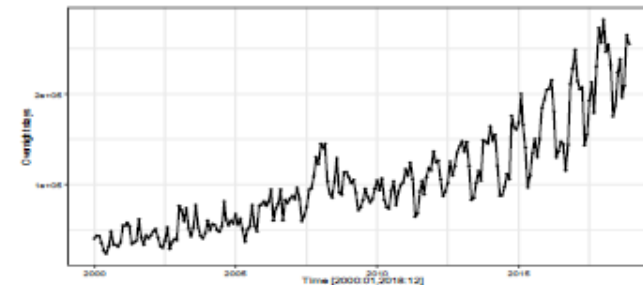


Fig. 1. Number of monthly overnight stays in accommodation establishments on Sal Island [2000:01-2018:12]. Source: Own elaboration with data provided by INE Cape Verde.

To stabilize the variation, the series was transformed using the logarithm $\Delta \log(X_t)$. As part of the modelling, 90% of the series data was taken for training and 10% for testing (2000:01-2017:12 and 2018:01-2018:12). The trend was removed by applying the first simple difference from the \log series. The Dickey-Fuller test was also used, obtaining a $p\text{-value} = 0.01<0.05$, confirming that the series became stationary with this first non-seasonal difference. Then, with the $nsdiff()$ function, it was determined how many seasonal differences would be needed (in this case one), resulting in the correlogram in Figure 2 which confirms that very little correlation remains in the series after taking two differences (the simple and the seasonal).

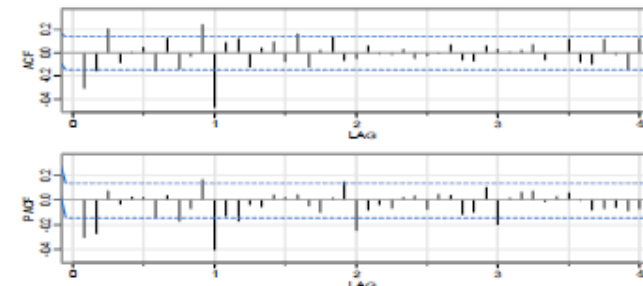


Fig. 2. ACF and PACF after the first seasonal difference of the \log series.

After analysing the correlogram (Figure 2) three models were considered (Table 1). The $auto.arima()$ function was used to estimate the model parameters and calculate the

Akaike information criterion (AIC) and the second-order Akaike Information Criterion (AICc).

Table 1. Initial time-series models for the monthly overnight stays in tourist establishments on the Sal Island.

#	Models	AIC	AICc	MAPE (%) (Training data)	MAPE (%) (Testing data)
1.	SARIMA(1,1,1)(0,1,1) ₁₂	-249.59	-249.39	0.794	0.526
2.	SARIMA(2,1,0)(0,1,1) ₁₂	-249.98	-249.78	0.774	0.578
3.	SARIMA(2,1,0)(0,1,2) ₁₂	-249.15	-248.85	0.773	0.595

Based on the error indicators of the models in the two modelling moments of the series (training and testing), and by the value of the AIC/AICc, the first two models were selected, as they presented the best MAPE, in addition to the fact that all their coefficients are all significant. Residual analysis was carried out to assess the adequacy of the models. When analysing the residues, the data and the respective ACF and PACF, taking into account the characteristics that the residues must have, such as white noise, normal distribution and constant variance, both models are acceptable. However, concerning the assumption of normality, it was found that the standardized residuals $(\hat{\epsilon}_t/S_t^2)$, do not deviate much from normality when observed through the histogram and QQ plot.

Figure 3 shows the adjustment of the model SARIMA(2,1,0)(0,1,1)₁₂ in the training (years 2000 to 2017) and testing (the year 2018) data. Using the accuracy() function the MAPE=0.794%. The model improved its performance with the testing data set (the year 2018), with a MAPE=0.526%.

Finally, the data from the year 2019 was used for prediction only (one-step forward). It was found that the SARIMA(2,1,0)(0,1,1)₁₂ model presented the lowest prediction error, MAPE=6.77%, against the model SARIMA(1,1,1)(0,1,1)₁₂ with MAPE=11.96%, which had a lower AIC value. It can be noted that the obtained forecast quality showed a MAPE <10% indicating a high accuracy [29].

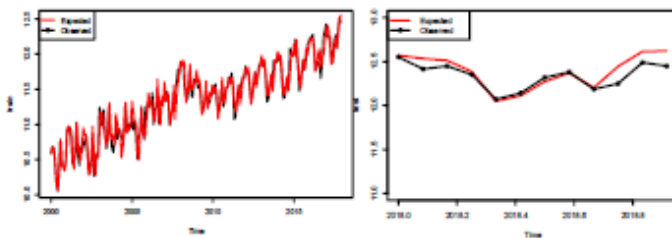


Fig. 3. SARIMA(2,1,0)(0,1,1)₁₂ model results: training data (left graph) and testing data (right graph).

Through expression (1), the SARIMA(2,1,0)(0,1,1)₁₂ can be presented with the following form:

$$(1 + 0.2672B + 0.1976B^2)(1 - B^{12})(1 - B)X_t = (1 - 0.5948B^{12}) \quad (6)$$

Table 2 and Figure 4, shows the predicted values for the next 12 months of the year 2019 with model SARIMA(2,1,0)(0,1,1)₁₂, which can be visualized in Figure 3.

Table 2. Number of predicted and observed monthly overnight stays and the respective confidence intervals at 80% and 95%, for the prediction data of the year 2019.

2019	Observed data	Expected data	Lo 80	Hi 80	Lo 95	Hi 95
Jan	267,940	273,666	257,276	290,056	248,600	298,732
Feb	237,203	253,985	233,666	274,304	222,909	285,061
Mar	233,508	254,351	231,730	276,973	219,755	288,948
Apr	207,956	234,672	209,358	259,986	195,958	273,386
May	152,691	181,532	153,773	209,290	139,079	223,984
Jun	192,645	191,583	161,686	221,479	145,860	237,305
Jul	217,145	220,787	188,871	252,702	171,976	269,598
Aug	235,091	234,487	200,660	268,315	182,752	286,222
Sep	193,380	201,811	166,181	237,440	147,320	256,301
Oct	224,602	230,599	193,255	267,942	173,486	287,711
Nov	257,670	276,033	237,049	315,019	216,412	335,656
Dec	236,604	271,384	230,826	311,943	209,355	333,414

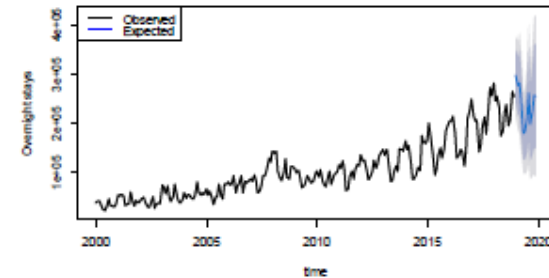


Fig. 4. Predicted values for the 12 months of the year 2019 with model SARIMA(2,1,0)(0,1,1)₁₂.

6 Conclusion

Empirically, considering the nature of tourism and the graphic visualization of the series, it was already perceived that one of the models of the Box-Jenkins method that could be applied would be the SARIMA, specially due to the presence of seasonality in tourism on the Sal Island. The results confirmed the presence of seasonality in the data. After performing all the analyses in search of the best SARIMA model that could describe the series of monthly overnight stays, the choice fell on the

SARIMA(2,1,0)(0,1,1)₁₂ model. This model presented the lowest average percentage error rate, in addition to complying with all the assumptions required for identification and diagnosis. The results of the model in one step ahead prediction (data from 2019) showed little difference between the predicted values and the observed number of overnight stays in each month of 2019, with an average error of 8.78%, lower than any other analysed model in this work. Since the data series is non-stationary, the SARIMA model, based on the results obtained demonstrated that it can deal well with this type of series, performing well in the training and testing phases, as well as in forecasting. As long as seasonality continues to be a strong feature in tourism on Sal Island, the SARIMA model may be the best choice in the forecast of the number of future monthly overnight stays in hotel establishments. However, we need to mention that 2020 and 2021 are atypical years due to the COVID-19 pandemic with a great impact on tourism. As soon as the tourist market re-establishes itself, the model can prove to be very useful to obtain good forecasts for the number of overnight stays in Cape Verde, with valuable information.

Note that the SARIMA(2,1,0)(0,1,1)₁₂ model was also the model chosen by the automatic choice procedure, the *auto.arima()* function, which corroborates the analysis of the ACF and PACF.

Although the tourism sector has drastically reduced its activities due to the pandemic caused by COVID 19, as a consequence it certainly did not record many overnight stays in 2020 as in previous years, even so, the proposed SARIMA model can contribute to simulate and analyse the percentage of overnight stays that did not take place due to the pandemic.

However, monitoring and updating the series is an ongoing challenge, so other model structures should be studied and applied to overnight stays series, such as neural networks or econometric models.

7 Acknowledgement

The authors are grateful to the UNIAG, R&D unit funded by the FCT – Portuguese Foundation for the Development of Science and Technology, Ministry of Science, Technology and Higher Education under Project no. UIDB/04752/2020 and to INEGI under LAETA project no. UIDB/5022/2020. G. Neves would also like to acknowledge the Sal City Council (*Câmara Municipal do Sal*) for their support of the PhD Scholarship.

8 References

- Poon, S.: A practical guide to forecasting financial market volatility, Chichester, John Wiley and Sons (2005).
- Song, H., Li, G.: Tourism demand modelling and forecasting – A review of recent research. *Tourism Management* 29 (2008) 203–220 (2008).
- Song, H., Qiu, R.T.R., Park, J.: A review of research on tourism demand forecasting. *Annals of Tourism Research* 75 (2019) 338–362 (2019).
- Wu, D.C., Song, H., Shen, S.: New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, Vol.29 No.1, pp.507–529 (2017).
- Goh, C., Law, R.: The Methodological Progress of Tourism Demand Forecasting: A Review of Related Literature. *Journal of Travel & Tourism Marketing*, 28:3, 296–317 (2011).
- Witt, S. F., Witt, C. A.: Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11, 447–475 (1995).
- Pearce, D. G.: *Tourism Development*. 1st ed. John Wiley, New York (1989).
- Song, H., Witt, S. F.: *Tourism demand modelling and forecasting: Modern econometric approaches*. 1st ed. Taylor & Francis, New York (2000).
- Cooper, C., Fletcher, J., Wanhill, S., Gilbert, D., Sheperd, R.: *Turismo, Principios e Prática*. 3.ª Ed.; Bookman, Porto Alegre (2007).
- Cunha, L., Abrantes, A.: *Introdução ao Turismo*, 6th ed. Lidel, Lisboa (2019).
- Matias, A.: *Economia do turismo: teoria e prática*. 1st ed. Instituto Piaget, Lisboa (2007).
- Neves, G.A., Nunes, C.S., Fernandes, P.O.: Challenges and Opportunities for Island Tourist Destinations: The Case of the Island of Sal, Cape Verde. In: Abreu A., Liberato D., González E.A., Garcia Ojeda J.C. (eds) *Advances in Tourism, Technology and Systems*. ICOTTS 2020. *Smart Innovation, Systems and Technologies*, vol 209, p. 498–508. Springer, Singapore (2021).
- Fernandes, P.O., Teiveira, J., Ferreira, J., Azevedo S.: Modelling Tourism Demand: A Comparative Study between Artificial Neural Networks and the Box Jenkins Methodology. *Romanian Journal of Economic Forecasting*, 5(3), 30–50 (2008).
- CMS: *Câmara Municipal do Sal – Relatório de Atividade* (2017).
- CMS: *Câmara Municipal do Sal – Relatório de Atividade* (2018).
- Xavier, N.M.J.: *Análise e Previsão de Séries Temporais com Modelos ARIMA e Análise Espectral Singular*. Dissertação de Mestrado, Universidade Aberta (2016).
- Chen, L., Li, G., Wu, C., Shen, S.: Forecasting Seasonal Tourism Demand Using a Multi-Series Structural Time Series Method. *Journal of Travel Research* (2017).
- Butler, R. W.: Seasonality in Tourism: Issues and Problems. In A. V. Seaton (eds.), *Tourism: the State of the Art*, Chichester: Wiley & Sons (1994).
- Butler, R.W.: Seasonality in Tourism: Issues and Implications. In Baum, T., and Lundtorp, S., (eds.). *Seasonality in Tourism, Advances in Tourism Research Series*, Pergamon, UK: Elsevier Science (2001).
- Biederman, P.S.: *Travel and tourism: an industry example*, Pearson Education, Upper Saddle River (2008).
- Canas, R.: An Overview of Tourism Seasonality: Key Concepts and Policies. *Alma Tourism – Journal of Tourism, Culture and Territorial Development*, 3(5), 40–58 (2012).
- Moretin, P.A., Toloi, C.M.C.: *Análise de Séries Temporais*. Editora Blucher (2006).
- Cryer, J. D., Chan, K.S.: *Time Series Analysis with Application in R*. 2nd Edn., Springer, New York. ISBN-10: 0387759581, pp: 491 (2008).
- Mishra, P.K., Rout, H.B., Pradhan, B.B.: Seasonality in Tourism and Forecasting Foreign Tourist Arrivals in India. *Iranian Journal of Management Studies (IJMS)*, Vol.11, No.4, Autumn, pp. 629–658 (2018).
- Ehlers, R.S.: *Análise de Séries Temporais*. Departamento de Estatística, UFPR (2005). <http://www.each.usp.br/rvicente/AnaliseDeSeriesTemporais.pdf>, last accessed 2018/04/16.
- Chatfield, C.: *Time Series Forecasting*, Chapman & Hall/CRC (2000).
- Moretin, P.A., Toloi, C. M.: *Modelos para Previsão de Séries Temporais*, Edgard Blucher (1981).
- Maddala, G.S., Lahiri, K.: *Introduction to Econometrics*. 4ª Edição, John Wiley & Sons Ltd, UK (2009).
- Lewis, C.D.: *Industrial and Business Forecasting Method*. Butterworth Scientific, London (1982).

