UNIVERSIDADE DE CABO VERDE
FACULDADE DE CIÊNCIAS E TECNOLOGIA
MESTRADO em MATEMÁTICA E APLICAÇÕES

Análise e Previsão de Séries Temporais: uma aplicação a dados da economia Cabo-Verdiana

Kelton Santos
Orientadora: M. Cristina Miranda

Com o apoio da Fundação Calouste Gulbenkian

Praia, Dezembro de 2021

UNIVERSIDADE DE CABO VERDE
FACULDADE DE CIÊNCIAS E TECNOLOGIA
MESTRADO em MATEMÁTICA E APLICAÇÕES

Análise e Previsão de Séries Temporais: uma aplicação a dados da economia Cabo-Verdiana.

Kelton Santos
Orientadora: M. Cristina Miranda

Praia, Dezembro de 2021

Resumo

Uma grande parte dos dados económico-financeiros podem ser representados em forma de séries temporais. Essas séries, que são caracterizadas por valores observados com igual espaçamento temporal podem ser analisadas e utilizadas para fazer previsões. Existem vários métodos usados na análise e previsão de séries temporais, sendo que a maioria usa valores históricos da série para criar um modelo usado para a representação e previsão da mesma. O objetivo deste trabalho foi de apresentar alguns desses métodos e aplicar ao Produto Interno Bruto (PIB) de Cabo Verde. O modelo ARIMA(1, 2, 1) foi o que produziu melhor resultado, somente não conseguindo prever com eficiência a queda que aconteceu em 2020 em consequência da pandemia da COVID-19.

Conteúdo

Lista de Tabelas

Lista de Figuras

Agradecimentos

Agradeço a Deus e à minha família pela força que me deram durante esta jornada. Também deixo um voto de agradecimento à minha orientadora, a todos os professores e colegas que contribuíram de forma direta ou indireta para a materialização deste trabalho.

Capítulo 1

Introdução

1.1 Enquadramento

No setor financeiro, o acesso aos dados para suporte na tomada de decisões é fundamental para o sucesso de qualquer empresa ou organização. Os dados devem ser fidedignos e devem prover informações úteis aos decisores. Atualmente, as organizações têm dado mais importância aos dados do que davam no passado, devido aos diversos métodos desenvolvidos para análise de dados, suportados por *softwares* capazes de processar grande volume de informações em pouco tempo, fornecendo assim informações de grande utilidade às empresas.

Um dos problemas que as organizações enfrentam é o acesso aos dados no momento certo para a tomada de decisões. Se os dados não estiverem disponíveis no momento exato ela poderá deixar de ser relevante para a tomada de determinadas decisões. Frequentemente, usam-se dados históricos para analisar o desempenho de determinada organização e tirar ilações sobre como melhorar no futuro. As organizações,normalmente, sentem necessidade de recorrer a previsões para tomarem decisões com um nível significativo de confiança. Existem vários tipos de dados com características diferentes e, por isso, a metodologia utilizada nas previsões varia consoante as características dos dados.

Neste trabalho serão analisados dados em forma de *séries temporais*, que são muito frequentes no setor financeiro. *Séries temporais* são sequências de observações sobre uma

determinada variável, ordenada no tempo, e registadas em períodos regulares [4]. Por exemplo, temperaturas diárias num determinado local ao longo do ano, vendas mensais de uma empresa, cotação diária de uma determinada ação na bolsa de valores, produto interno bruto (PIB) anual ou trimestral de um determinado país, etc.

Para a análise e previsão de séries temporais existem vários métodos ou modelos, mas neste trabalho o foco incidirá na metodologia ARIMA (Modelos Autoregressivos Integrados Médias Móveis, do inglês, *Auto Regressive Integreted Moving Average*).

Parte da motivação para este trabalho reside na utilização de dados reais da economia Cabo-Verdiana como objeto de análise. Assim, na parte prática do trabalho, serão utilizados os valores do PIB de Cabo Verde em volume de 1980 a 2018 para a análise e previsão utilizando a metodologia ARIMA. O *software R* [19] será utilizado para a análise dos dados, por ser muito potente e de fácil utilização.

1.2 Objetivos

A necessidade de ter informações atualizadas na tomada de decisões é imperativo para o sucesso das empresas e organizações que operam no setor financeiro. Em Cabo Verde, as previsões económicas são feitas normalmente pelo Banco de Cabo Verde (BCV), Ministério das Finanças e o Instituto Nacional de Estatísticas (INE), que são todas entidades do Estado ou relacionadas com o Estado. Essas previsões normalmente são de um número limitado de indicadores económicos (normalmente o PIB, a taxa de desemprego, taxa de inflação, etc.), daí a necessidade de ter pesquisas independentes que forneçam outras previssões que poderão servir para confrontação ou comparação com aqueles publicados pelas entidades acima, fornecendo ao público outras fontes de pesquisa. Pretende-se com esse trabalho alcançar os seguintes objetivos:

- Apresentar alguns métodos usados para análise e previsão de séries temporais;
- Aplicar a metodologia ARIMA para a análise e previsão de séries temporais, recorrendo a dados da economia Cabo-Verdiana.

1.3 Hipóteses de estudo

Para a realização desse estudo parte-se do pressuposto que:

- Vários indicadores económico-financeiros da economia Cabo-Verdiana podem ser representados através de séries temporais;
- A metodologia ARIMA produzirá previsões com alta precisão, do Produto Interno Bruto de Cabo Verde;

1.4 Motivação

Os dados do setor financeiro muitas vezes podem ser representados em forma de séries temporais devido a sua característica referente ao tempo de observação ou ocorrência (anuais, trimestrais, mensais, diárias, etc.). Devido aos vários métodos de análise e previsão desenvolvidos ao longo do tempo e do avanço no domínio computacional, é possível tentar obter previsões das séries temporais com melhor precisão do que se costumava obter no passado.

Cabo Verde é um país que tem se esforçado para disponibilizar a tempo útil as informações económico-financeiros, para não perder oportunidades de investimento a nível nacional e internacional, assim que a apresentação de ferramentas adicionais de análise e previsão será uma mais valia para atingir esse objetivo.

Uma das razões para a realização deste trabalho, é a de poder contribuir com exemplos de aplicações das séries temporais na análise e previsão dos indicadores económico-financeiros de Cabo Verde e despertar assim na comunidade académica e entidades independentes, o interesse de realizar pesquisas adicionais e produzir informações periódicas sobre a economia Cabo-Verdiana, que possam ser úteis ao mercado financeiro nacional e internacional.

1.5 Metodologia

Para alcançar os objetivos estabelecidos para este trabalho, será utilizada uma abordagem analítica e descritiva na recolha e análise dos dados. Após uma apresentação dos principais métodos para análise e previsão de séries temporais, com base numa revisão bibliográfica, será feita uma aplicação com dados da economia de Cabo Verde. O foco de interesse será a análise dos dados publicados pelo Banco de Cabo Verde, Instituto Nacional de Estatística, Ministério das Finanças e outras entidades com publicações relevantes. Os dados serão analisados utilizando o *software R* com suporte do *Rstudio* [21].

Capítulo 2

Revisão da literatura

Os métodos de previsão geralmente usam valores do passado para fazer previsões de valores futuros. No início do século XIX, as previsões eram feitas recorrendo à extrapolação simples de um valor global, ajustado em função do tempo [25]. Devido à incapacidade desses métodos solucionarem problemas mais complexos, foi necessário desenvolver métodos mais eficientes e capazes de produzir melhores resultados.

As séries temporais estão incluídas no domínio das probabilidades e estatística, mais concretamente, no campo dos processos estocásticos. O início dos anos 70, foi considerado como a época de ouro no desenvolvimento do estudo das séries temporais devido aos vários trabalhos feitos e às metodologias de análise e previsão desenvolvidas [9]. O trabalho de Box e Jenkins [4] é considerado como uma obra de referência no estudo das séries temporais, pois ela integra os conhecimentos existentes sobre séries temporais até a data. Os autores apresentaram modelos estocásticos no domínio de séries temporais discretas com número mínimo de parâmetros, satisfazendo o princípio de parcimónia.

São muitos e diversos os trabalhos de investigação com aplicações de séries temporais na economia e finanças. Numa pesquisa realizada nos Estados Unidos da América por Andrei et al. [1], foram utilizadas séries temporais para se fazer a previsão do PIB dos EUA. Como a série encontrada não era estacionária, os autores usaram vários métodos estatísticos para a transformar numa série estacionária. Depois de aplicar os testes a série foi transformada numa estacionária e integrada de ordem 1. Os autores aplicaram a metodologia de

Box-Jenkins para a determinação do modelo ARMA (Modelos Auto Regressivos Médias Móveis). Usando o método dos mínimos quadrados na determinação dos parâmetros, optou-se, nesse caso, pelo modelo $ARIMA(1,1,1)$ devido à melhor performance.

Num outro trabalho similar desenvolvido para a previsão do PIB do Brasil, a preços correntes [7], os autores usaram os modelos Holt-Winters multiplicativo, SARIMA e o modelo linear dinâmico para a previsão. Os dados compreendiam observações entre 1996 e 2019 (observações trimestrais). A observação da medida de U-Theil mostrou que os modelos precisavam se ajustar a períodos em que choques económicos significativos afetam o crescimento da economia Brasileira. A série analisada e a previsão dos modelos apresentavam a necessidade de um crescimento sustentado numa economia de mercado. Também se verificou que na série analisada, o modelo linear dinâmico apresentava melhor ajuste aos dados e um desempenho preditivo mais eficiente.

Baltac [2], num pequeno artigo, estudou a possibilidade da aplicação de séries temporais para fazer uma análise económica e financeira, usando como indicador principal o *turnover* (indicador do grau de liquidez dos ativos de uma empresa, ou seja, a capacidade de uma empresa gerar receitas usando os seus ativos). No trabalho, foi realizado um estudo prático, onde foi analisada a evolução do *turnover* extraído do balanço de uma empresa da Roménia, no período de 2004 a 2014. Usando séries temporais, a autora determinou a previsão de crescimento da empresa.

Numa pesquisa sobre a previsão do crescimento económico na região de Shenzhen na China usando séries temporais [24], o autor aplicou o modelo $ARIMA(3,3,5)$, pois segundo ele, é o que melhor reflete a lei do desenvolvimento económico da região, e pode ser usado para previsão a médio e a longo prazo. Chegou-se à conclusão de que, num espaço de 5 anos, a região em estudo experimentaria uma tendência de crescimento lento.

Através de uma pesquisa empírica realizada por Zhenwei et al. [15], concluiu-se que o modelo ARIMA tradicional tem uma variância muito grande na previsão de séries temporais financeiras com alta frequência. Com a melhoria da capacidade de computação, os autores chegaram a conclusão que juntando o modelo ARIMA tradicional com a tecnologia *deep learning* pode-se melhorar as previsões de tais séries. No seu estudo, o autor usou como série

temporal o índice CSI300 da Bolsa de Valores de Shanghai.

Stock [23] numa pesquisa mais teórica, apresentou várias especificidades de séries temporais e a sua aplicação na economia. No estudo foram apresentadas as séries temporais univariáveis e multivariáveis, os detalhes a considerar na escolha de modelos, dificuldades habitualmente encontrados nas previsões, instabilidades dos modelos e possíveis falhas nos modelos.

Um aspeto a ter em conta na modelação estacionária de séries temporais, segundo Lopes [16], tem a ver com a sazonalidade dos dados. Por exemplo, segundo a autora, para certas séries mensais os dados relativos a um mesmo mês em diferentes anos têm tendência a situar-se de maneira semelhante em relação à média anual. Uma maneira de solucionar esse problema segundo a autora é aplicar a metodologia $SARIMA$ (a letra S vem do temo *sazonal*), desenvolvida por Box e Jenkins, capaz de ter em conta a sazonalidade dos dados. O princípio subjacente a esse modelo, é a aplicação do modelo $ARIMA$, após a eliminação da sazonalidade.

Capítulo 3

Análise das séries temporais

3.1 Introdução

As previsões são necessárias em várias situações para servir como suporte na tomada de decisões e muitas vezes com vários anos de antecedência. Por exemplo, para uma empresa decidir se vai investir na construção de uma nova unidade de produção, nos próximos cinco anos, requer previsões sobre a procura dos produtos e sobre o retorno do investimento no futuro. Para se decidir sobre a implementação de determinada política económica-financeira, as autoridades de um determinado país precisam de previsões para análise da viabilidade da implementação dessa política.

Alguns eventos são mais fáceis de prever do que outros. Por exemplo, pode-se prever com alto grau de precisão, a que horas será o pôr do sol daqui a uma semana, mas a previsão de qual a equipa que irá vencer o próximo campeonato do mundo de futebol não pode ser feita com 100% de certeza. Segundo Hyndman [13], a previsibilidade de um evento depende de vários fatores, tais como:

- Nível de conhecimento dos factores que afectam o evento;
- Quantidade e qualidade dos dados disponíveis;
- Efeito que o conhecimento da previsão poderá ter no próprio evento.

Por exemplo, a previsão do consumo de eletricidade pode ser feita com alto grau de precisão, pois as três condições acima são normalmente satisfeitas. Têm-se ideia dos fatores que podem afetar o consumo de eletricidade (temperatura, época do ano, condições económicas da população, etc.); normalmente têm-se dados do passado relativamente ao consumo de eletricidade; a previsão determinada normalmente não influencia o consumo de eletricidade no futuro [13]. No caso de previsões de taxas de câmbio somente a segunda condição é satisfeita, visto que se tem conhecimento limitado dos fatores externos que podem afetar as taxas de câmbio e a previsão e publicação dessas taxas de câmbio podem afetar a taxa de câmbio no futuro.

Geralmente, os métodos usados nas previsões analisam dados do passado para tentar fazer previsões sobre valores futuros [25]. O método apropriado a usar nas previsões depende muito dos dados disponíveis. Quando não existem dados disponíveis, ou se os dados não são muito relevantes para a previsão, usam-se **métodos de previsões qualitativas**, que são métodos bem desenvolvidos estruturados para obter previsões sem usar dados históricos [13].

Os **métodos de previsão quantitativos** são aplicados quando as seguintes condições são satisfeitas:

1. Dados numéricos sobre o passado estão disponíveis;
2. É razoável assumir que alguns padrões verificados no passado continuarão no futuro.

Existem vários métodos de previsões quantitativas dos mais simples aos mais complexos, normalmente desenvolvidos para áreas e propósitos específicos. Uma grande parte dos métodos quantitativos usam dados em forma de *séries temporais* (recolhidos em intervalos regulares ao longo do tempo) ou dados de *corte transversal*, também chamados *cross-sectional* (relativos a um dado instante ou período de tempo).

Em economia há dois procedimentos predominantes no processo de previsão: *modelos econométricos* e *modelos de séries temporais*. Os *modelos econométricos* são modelos de previsões que determinam futuros movimentos numa variável relacionando-a com um conjunto de outras variáveis de forma causal. Esses modelos aplicam teorias económicas

para construir modelos que podem incluir várias variáveis. Por exemplo, um modelo econométrico para determinar a taxa de juros, pode utilizar dados de variáveis como PIB, preços, dinheiro em circulação, etc., através dos métodos de regressão como na equação (3.1), onde X_{i1} e X_{i2} são as variáveis explicativas, β_0, β_1, β_2 são os coeficientes da regressão, ε_i o erro e Y_i a variável a ser estimada.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i. \tag{3.1}$$

Em determinadas situações o uso de modelos econométricos pode ser difícil ou impossível. Por exemplo, se não houver dados disponíveis para as variáveis explicativas X_{i1} e X_{i2} que permitam a determinação dos coeficientes β_0, β_1, β_2, ou se os dados disponíveis resultarem em erros ε_i muito grandes.

Outros métodos muito utilizados nas previsões, são os métodos de *séries temporais* que usam dados históricos da variável a ser estimada (podem também utilizar outras variáveis). Modelos de previsão de séries temporais são muitas vezes caracterizados como *método sofisticado de extrapolação* [18]. O foco desse trabalho é justamente previsão com método de séries temporais. Informação adicional sobre previsão usando métodos econométricos pode ser consultado em [18].

3.2 Séries temporais

Uma série temporal é um conjunto de observações ordenadas no tempo de característica quantitativa de interesse e que são usualmente registadas em intervalos regulares. Como exemplo de séries temporais temos:

i. Estimativas trimestrais do PIB de um país;
ii. Valores diários da temperatura em Cabo Verde;
iii. Cotações diárias das ações do Banco Comercial do Atlântico na Bolsa de Valores de Cabo Verde;
iv. Valores mensais de vendas numa determinada empresa;

v. Registo de marés na ilha de Santiago - Cabo Verde.

As séries dos exemplos (i) a (iv) são séries temporais *discretas*, enquanto a série do exemplo (v) é *contínua* [17]. As séries temporais contínuas podem ser transformadas em discretas se for considerado uma amostragem em intervalos de tempos iguais. Dessa forma, para analisar a série (v) seria preciso usar uma amostra, por exemplo, em intervalos de uma hora, convertendo a série contínua numa discreta com N pontos.

Podemos ver na Figura 3.1, a representação de dados em forma de série temporal, relativos ao PIB de Cabo Verde entre 2007 e 2020 (Fonte: INE). Trata-se de uma série discreta, visto termos informações anuais(O gráfico parece ser contínua, pois foi feito a união do pontos, mas é discreta).

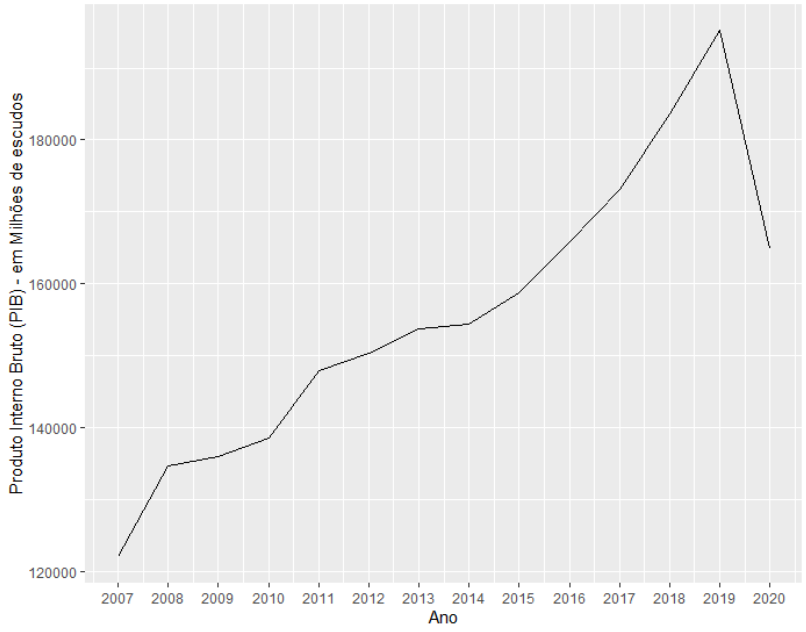

Figura 3.1: PIB de Cabo Verde a preços correntes entre 2007 – 2020

3.2.1 Objetivos da análise de séries temporais

Ao analisar uma série temporal, espera-se que exista uma causa relacionada com o tempo, que influenciou os dados no passado e que continuará a influenciar futuramente [25]. Segundo Morettin [17] o objetivo da análise de séries temporais é de construir modelos para as séries, com propósitos determinados.

Podemos destacar os seguintes objetivos da análise de séries temporais, segundo Xavier [25]:

1. Determinar as propriedades de uma série,tais como, sazonalidades, estacionaridade, padrão de tendência, etc;
2. Desenvolver um modelo estatístico que permita explicar o comportamento da série no período observado;
3. Estimar valores futuros de uma série temporal, com base em dados históricos;
4. Tomar medidas para controlar determinado processo.

De acordo com Morettin [17], existem dois enfoques usados na análise de séries temporais para o alcance dos objetivos acima. No primeiro enfoque, a análise é feita no *domínio temporal* e os modelos propostos são *modelos paramétricos* (com número finito de parâmetros). Pode-se destacar nos modelos paramétricos os modelos ARIMA. No segundo enfoque, a análise é conduzida no *domínio de frequências* e os modelos propostos são *modelos não paramétricos*. No domínio de frequências tem-se a *análise espectral*, que tem aplicações em ciências físicas e engenharia que consiste em decompor a série inicial em componentes de frequência onde a existência de *espectro* é a característica fundamental. Para este trabalho não será estudado este tipo de análise, podendo-se consultar [4] para mais detalhes.

Quer no domínio temporal, quer no domínio de frequências, são construídos *modelos probabilísticos* ou *estocásticos* para atingir os objetivos acima. Os modelos construídos devem ser simples e parcimónicos, ou seja, deve-se ter o menor número possível de parâmetros, e sempre que possível, a utilização dos modelos não deve apresentar dificuldades na sua aplicação [17].

3.3 Processos estocásticos

Os modelos que se utilizam para descrever séries temporais são processos estocásticos, ou seja, processos controlados por leis probabilísticas. Para Chatfield [6], a maioria das séries temporais são estocásticas, ou aleatórias, isto é, o futuro é apenas parcialmente determinado

por valores passados, sendo o modelo para estas séries muitas vezes chamado *processo estocástico*.

Definição 3.1 (Fenómenos Aleatórios)**.** São fenómenos naturais em que se supõe a intervenção do acaso no sentido em que não é possível, a partir do passado prever determinísticamente o futuro.

Os fenómenos aleatórios que se desenrolam no tempo são objeto de estudo dos *processos estocásticos*, isto é, os modelos matemáticos que descrevem os fenómenos aleatórios que evoluem ao longo do tempo.

Definição 3.2 (Processo Estocástico)**.** Dado um espaço de probabilidade (Ω, F, P) e um conjunto T, *processo estocástico* é uma função $X(t, w)$ definida no produto cartesiano $T \times \Omega$, que para cada $t \in T$ é uma variável aleatória. Simbolicamente escreve-se:

$$X = \{X_t, t \in T\}. \tag{3.2}$$

Assim, um processo estocástico é uma família de variáveis aleatórias $X = \{X_t, t \in T\}$ definidas num mesmo espaço de probabilidade (Ω, F, P). Normalmente T é tomado como o conjunto $\mathbb{Z}$ ou o conjunto dos números reais $\mathbb{R}$. Para cada $\omega_0 \in \Omega$, $X(w_0, t) = x(t)$ é uma função não aleatória de t com domínio em T. Assim podemos identificar um *processo estocástico* como sendo um sistema que, a cada ponto $\omega_0 \in \Omega$ faz corresponder uma função de parâmetro t. Cada uma dessas funções chama-se *realização* ou *trajetória* do processo, ou ainda, uma *série temporal* [12]. A totalidade das realizações $X_t^{(1)}$, $X_t^{(2)}$, ... designa-se por *ensamble*.

3.3.1 Classificação dos processos estocásticos

O conjunto dos valores $X_t, t \in T$ é chamado *espaço dos estados*, E, do processo estocástico e os valores de X_t são chamados *estados*, ou seja, E é o conjunto dos possíveis valores das variáveis aleatórias $X_t, t \in T$.

- Se T e E forem conjuntos discretos diz-se que X é um processo estocástico de tempo discreto com espaço de estados discreto;
- Se T for discreto e E contínuo, diz-se que X é um processo estocástico de tempo discreto com espaço de estados contínuo;
- Se T contínuo e E discreto, diz-se que X um processo estocástico de tempo contínuo com espaço de estados discreto;
- Se T e E são contínuos diz-se que X um processo estocástico de tempo contínuo com espaço de estados contínuo.

3.3.2 Lei de probabilidade de um processo estocástico

Sejam $\{t_1, t_2, ..., t_n\}$ elementos quaisquer do conjunto T pode-se determinar a lei de probabilidade conjunta do vector aleatório $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ de dimensão finita n, através da função de distribuição conjunta:

$$F_{(X_{t_1}, X_{t_2}, ..., X_{t_n})}(x_1, x_2, ..., x_n) = P(X_{t_1} \leq x_1, X_{t_2} \leq x_2, ..., X_{t_n} \leq x_n). \tag{3.3}$$

O conjunto de todas as leis de probabilidade, definida por:

$$F_{(X_{t_1}, X_{t_2}, ..., X_{t_n})} : (t_1, t_2, ..., t_n \in T)$$

é conhecida como *família de distribuições de dimensão finita* do processo estocástico. Geralmente, quando t é discreto, o conhecimento das funções de distribuição de dimensão finita permite a determinação da probabilidade de qualquer acontecimento associado ao respetivo acontecimento [12].

Definição 3.3 (Incrementos Independentes)**.** O processo estocástico $X = \{X_t, t \in T\}$ diz-se *processo estocástico com incrementos independentes* sse $\forall n, \forall t_1, t_2, ..., t_n \in T : t_1 < ... < t_n$ as variáveis aleatórias $X_{t_2} - X_{t_1}, ..., X_{t_n} - X_{t_{n-1}}$ são independentes.

A definição anterior é importante do ponto de vista da modelação, visto que permite descrever fenómenos cujos resultados são independentes em intervalos de tempo disjuntos.

Definição 3.4 (Incrementos estacionários)**.** O processo estocástico $X = \{X_t, t \in T\}$ diz-se *processo estocástico com incrementos estacionários* sse $\forall s, t \in T (s < t)$ a distribuição de incrementos $X_t - X_s$ depende apenas da amplitude $t - s$.

Em outras palavras, a distribuição de resultados em qualquer intervalo de tempo depende unicamente da amplitude do intervalo. Quando um processo estocástico tem incrementos independentes e estacionários simultaneamente diz-se *processo com incrementos independentes e estacionários*.

Definição 3.5 (Processo de segunda ordem)**.** O processo estocástico $X = \{X_t, t \in T\}$ diz-se *processo real de segunda ordem* se

$$\forall t \in T, E(X_t^2) < \infty.$$

Exemplos clássicos de um processo de segunda ordem são:

1. **Ruído Branco** ou *white noise* $(X_t, t \in T)$ define-se por:

- $\forall t \in T, E(X_t) = 0$;
- $\forall t \in T, V(X_t) = \sigma^2$;
- $\forall s, t \in T : s \neq t, Cov(X_s, X_t) = 0$.

2. **Processo Gaussiano** $(X_t, t \in T)$ tal que $\forall n \in \mathbb{N}, \forall t_1, ..., t_n \in T$, $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ é vector aleatório Gaussiano.

3.3.3 Processos estacionários

Os modelos de séries temporais a serem desenvolvidos nos próximos capítulos são todos baseados na suposição de que as séries são geradas por um processo estocástico. Por outras palavras, assumimos que cada valor $X_{t_1}, X_{t_2}, ..., X_{t_n}$ é extraído aleatoriamente de uma distribuição de probabilidade. Ao modelar esses processos, tentamos descrever as

características da sua aleatoriedade. Isso nos permite inferir, até certo ponto, sobre as probabilidades associados aos valores futuros alternativos da série.

Se pudéssemos numericamente especificar a função de distribuição de probabilidade das séries, poderíamos determinar resultados futuros. Infelizmente, a determinação de tal função de distribuição é, normalmente, impossível. Entretanto, é possível construir um modelo simplificado de séries temporais que explicam a aleatoriedade, de maneira útil para previsões. O modelo simplificado construído não precisa produzir obrigatoriamente os valores passados conhecidos, uma vez que a série e o modelo são estocásticos. Esse modelo deve simplesmente capturar as características da aleatoriedade da série analisada [18].

Segundo Hamilton [12], poderemos ter as seguintes classes de processos:

- Processos estacionários ou não estacionários, consoante a independência relativamente à origem dos tempos;
- Processos normais (Gaussianos) ou não normais, de acordo com as funções de distribuição de probabilidades (fdp) que caracterizam os processos;
- Processos Markovianos ou não Markovianos, consoante a independência dos valores em instantes precedentes.

Assim um processo X_t é estacionário se desenvolver no tempo de modo que a escolha de uma origem dos tempos não seja importante, ou seja, as características de $X_{t+\delta}$, para todo δ, são as mesmas de X_t.

Em termos gerais, os processos estacionários traduzem situações em que o sistema se apresenta num estado de equilíbrio estatístico em torno de um nível médio fixo, isto é, tem propriedades probabilísticas estáveis ou invariantes ao longo do tempo.

Definição 3.6 (Função de covariância)**.** Uma função real

$$\gamma : T \times T \to \mathbb{R}$$
$$(s,t) \to \gamma(s,t)$$

é uma função de covariância de um processo estocástico real de segunda ordem sse γ é uma função simétrica e definida positiva.

Por outras palavras, γ é uma função de covariância se satisfazer as seguintes condições:

- $\gamma(0) > 0$;
- $\gamma(-h) = \gamma(h)$;
- $|\gamma(t)| \leq \gamma(0)$
- $\gamma(t)$ é **positiva definida**, no sentido que

$$\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \gamma(t_j - t_k) \geq 0, \forall n \in \mathbb{N}, \forall a_j \in \mathbb{R}, j, k = 1, ..., n. \tag{3.4}$$

Definição 3.7 (Função de autocorrelação)**.** A função de autocorrelação é definida por

$$\rho = Corr(X_t, X_{t+h}) = \frac{Cov(X_t, X_{t+h})}{\sqrt{V(X_t)}\sqrt{V(X_{t+h})}} = \frac{\gamma(h)}{\gamma(0)}. \tag{3.5}$$

Como usualmente é quase impossível obter uma descrição completa de um processo estocástico (isto é, especificar a função de distribuição de probabilidade correspondente), a função de autocorrelação demonstra-se muito útil porque ela fornece uma descrição parcial do processo para efeito de modelação. A função autocorrelação indica o nível de correlação existente entre pontos vizinhos de uma série e ela é usada normalmente para identificar um modelo adequado para a série temporal X_t [18].

Definição 3.8 (Estacionaridade forte)**.** Um processo estocástico $X = \{X_t, t \in T\}$ diz-se *fortemente estacionário*, se $\forall n \in \mathbb{N}, t_1, t_2, ..., t_n \in T : t_1 < ... < t_n$,

$$(X_{t_1}, ..., X_{t_n}) \overset{\mathrm{d}}{=} (X_{t_{1+h}}, ..., X_{t_{n+h}}), h > 0, \tag{3.6}$$

isto é,

$$P(X_{t_1} \leq x_1, ..., X_{t_n} \leq x_n) = P(X_{t_{1+h}} \leq x_1, ..., X_{t_{n+h}} \leq x_n). \tag{3.7}$$

De acordo com Shumway [22], a estacionaridade em (3.7) é muito forte para a maioria das aplicações. Em vez de impor condições para todas as possíveis distribuições de uma série

temporal, pode-se usar uma versão mais "mansa" que impõe condições somente nos primeiros dois momentos das séries (momentos de segunda ordem). Teremos então a seguinte definição:

Definição 3.9 (Estacionaridade fraca)**.** Um processo estocástico $X = \{X_t, t \in T\}$ diz-se *fracamente estacionária*, se

- $\forall t \in T, E(X_t^2) < \infty$ (processo de segunda ordem);
- $\forall t \in T, E(X_t) = \mu$ (independente de t);
- $\forall t, s \in T, \Gamma(s,t) = Cov(X_s, X_t) = \gamma(|t-s|)$.

Daqui em diante usaremos o termo **estacionária** em vez de **fracamente estacionária**. Se um processo for fortemente estacionário, usaremos o termo **estritamente estacionária** ou **fortemente estacionária** Os processos estacionários são apropriados para modelar fenómenos aleatórios cujo comportamento parece não mudar muito com o tempo [22].

Existem vários tipos de não estacionaridade, mas nos próximos capítulos veremos que tais processos podem ser transformados em processos estacionários através de diferenças sucessivas.

3.4 Tendência e sazonalidade das séries temporais

Dois dos principais motivos que causam variações na maioria das séries temporais são a *tendência* e a *sazonalidade*. Pode-se definir **tendência** como um comportamento de longo prazo da série temporal. Normalmente esse tipo de variação está presente quando uma série apresenta constante crescimento ou declínio, em sucessivos períodos [25].

Na Figura 3.2 temos a representação da entrada de turistas em Cabo Verde do ano 2000 ao ano 2020 (Fonte: INE). Uma simples análise visual permite observar uma tendência de crescimento até finais do ano 2019. Em 2020 pode-se ver uma grande queda na entrada de turistas, devido ao impacto da COVID-19 que assolou o mundo em 2020.

Uma pequena análise gráfica pode dar uma ideia do comportamento dos dados, mas temos sempre que realizar testes estatísticos específicos para concluir sobre determinado

comportamento dos dados, nesse caso, a tendência. Veremos mais adiante alguns métodos usados para detetar a existência de tendência em séries temporais.

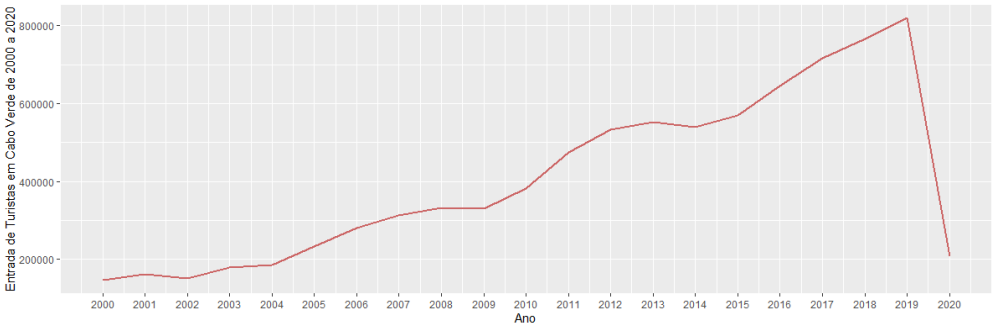

Figura 3.2: Entrada de turistas em Cabo Verde de 2000 – 2020.

Se considerarmos que a série temporal não apresenta a componente sazonal, poderemos escrever o modelo

$$X_t = T_t + \varepsilon_t. \tag{3.8}$$

onde ε_t é um ruído branco (uma variável aleatória com média zero e variância σ^2) e T_t o estimador da tendência. Existem vários métodos para estimar T_t. Segundo Hamilton [12] os mais utilizados são:

1. Ajustar uma função do tempo, como um polinómio, uma exponencial ou outra função suave de t;
2. Suavizar (ou filtrar) os valores da série ao redor de um ponto, para estimar a tendência naquele ponto.

Após estimada a tendência $\hat{T}_t$, podemos considerar a série ajustada para a tendência ou livre da tendência,

$$Y_t = X_t - \hat{T}_t. \tag{3.9}$$

Outro procedimento que é também utilizado para eliminar a tendência de uma série, consiste em tomar sucessivas diferenças da série original até encontrar uma estacionária. Segundo

Morettin [17], em séries económicas, é frequente conseguir a estacionaridade com as primeiras diferenças.

$$\Delta X_t = X_t - X_{t-1}. \tag{3.10}$$

O testes mais utilizados para identificar tendências são:

1. *Teste de Sequências* (Wald-Wolfowitz);
2. *Teste do Sinal* (Cox - Stuart);
3. *Teste baseado no coeficiente de correlação de Spearaman.*

Definição 3.10 (Operador de Diferenças)**.** Define-se operador de diferenças por:

$$Bx_t = x_{t-1} \tag{3.11}$$

e pode ser extendido para outros expoentes $B^2x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$, e assim por diante. Temos então:

$$B^k x_t = x_{t-k}, \tag{3.12}$$

Dado a diferença de primeira ordem $\nabla x_t = x_t - x_{t-1}$, a mesma pode ser escrita como:

$$\nabla x_t = (1 - B)x_t. \tag{3.13}$$

Podemos então estender a notação para ordens superiores. Por exemplo, o operador de diferenças de segunda ordem será:

$$\begin{aligned} \nabla^2 x_t = (1 - B)^2 x_t &= (1 - 2B + B^2)x_t \\ &= x_t - 2x_{t-1} + x_{t-2}. \end{aligned} \tag{3.14}$$

Definição 3.11 (Diferenças de ordem d)**.** São definidas como:

$$\nabla^d x_t = (1 - B)^d \tag{3.15}$$

Pode-se expandir o operador $(1-B)^d$ algebricamente para avaliar o maior valor inteiro de d. Quando $d = 1$ escreve-se simplesmente $(1-B)$.

A primeira diferença, $\nabla x_t = x_t - x_{t-1}$, é um exemplo de filtro linear que pode ser aplicado para eliminar a tendência da série. Outros filtros, formados pela média dos valores perto de x_t podem produzir séries ajustadas que eliminam outros tipos de flutuações indesejáveis [22]. O método da diferença, foi desenvolvido por Box e Jenkins [4] e é um componente importante na aplicação do modelo ARIMA.

A **Sazonalidade** é um tipo de variação onde fenómenos que ocorrem durante o tempo se repetem a cada período idêntico de tempo, por exemplo, o aumento de vendas de uma loja todos os anos no período de Natal. A sazonalidade pode ser dividida em dois tipos, *sazonalidade determinista*, quando o padrão sazonal é regular e estável no tempo, ou *sazonalidade estocástica* quando o padrão sazonal varia como tempo [25].

A Figura 3.3 representa o movimento de passageiros nos aeroportos de Cabo Verde entre o ano 2006 e o ano 2017 (Fonte: INE). Pode-se observar que o mês de julho normalmente têm maior tráfego de passageiros nos aeroportos de Cabo Verde, possivelmente devido ao período de férias escolares e também por coincidir com o início do verão, onde aumenta o fluxo de turistas que visitam Cabo Verde. Trata-se assim de dados que apresentam um padrão sazonal, pois tem comportamento semelhante em certos períodos (nesse caso no mês de julho).

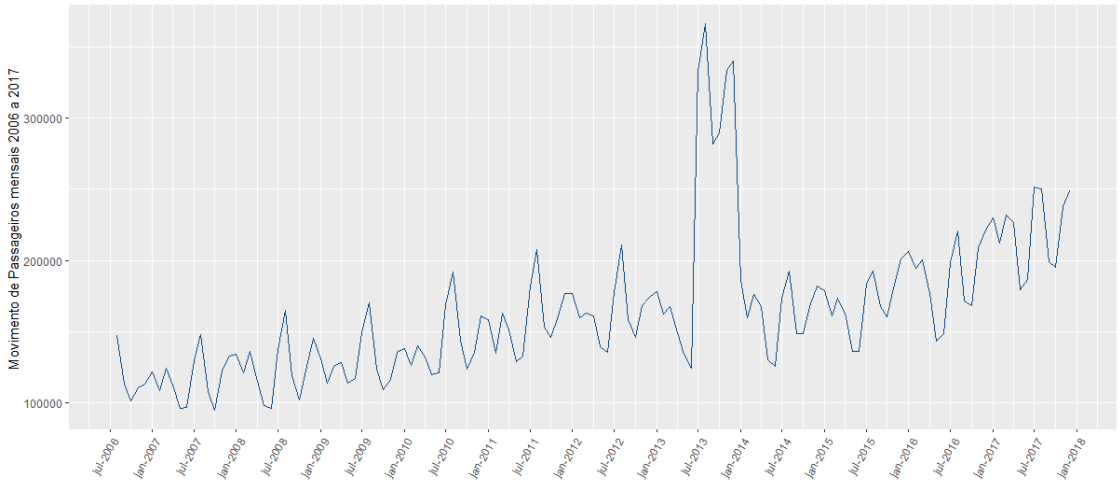

Figura 3.3: Movimento de passageiros nos aeroportos de Cabo Verde de 2006 – 2017.

Existem vários procedimentos para se estimar a sazonalidade, sendo que as mais utilizadas são os *métodos de regressão* e os *métodos de médias móveis*. Os métodos de regressão são mais apropriados para séries que apresentam sazonalidade determinística, ou seja, pode ser prevista normalmente a partir de comportamento dos dados em períodos anteriores [12]. Já o método de médias móveis é mais apropriado para séries temporais cuja componente sazonal varia com o tempo, isto é, séries de sazonalidade estocástica.

Os procedimentos mais utilizados para estimar a sazonalidade são:

1. *Teste de Kruskal-Wallis*, para várias amostras independentes;
2. *Teste de Friedman*, para amostras dependentes ou emparelhadas;
3. *Teste F* para determinar sazonalidade determinística.

Outra abordagem seria incorporar a variação sazonal e a tendência num modelo ARIMA (será estudado no próximo capítulo).

3.4.1 Teste da normalidade

Existem alguns testes utilizados para verificar se os dados de uma série temporal são normalmente distribuídos. Entre esses testes podemos destacar o teste de Shapiro-Wilk, o teste de Doornik-Hansen, o teste de Lillefors e o teste de Jarque-Bera. Nesses testes, parte-se da hipótese nula de que os dados provêm de uma distribuição normal, e a hipótese alternativa é de que os dados não provêm de uma distribuição normal.

3.4.2 Teste da raiz unitária ou teste de estacionaridade

Como explicado anteriormente, uma série temporal em que as propriedades não se alteram ao longo do tempo é chamada de série estacionária. Fazendo a análise de uma representação gráfica de uma série temporal, pode-se ter uma ideia se a série é estacionária ou não. Entretanto, somente a visualização gráfica é insuficiente para concluir sobre a estacionaridade da série. Assim como para a normalidade, existem alguns testes usados para se estudar a

existência de estacionaridade duma série temporal, sendo que a maioria deles baseiam-se em encontrar uma raiz unitária. Alguns testes mais utilizados são os testes de Dickley-Fuller Aumentado, o teste de Phillips-Perron e o teste KPSS [25].

3.4.2.1 Teste de Dickey-Fuller Aumentado

Também é conhecido como teste ADF (Augmented Dickley-Fuller) e tem como base a seguinte regressão:

$$\Delta x_t = \beta_1 + \beta_2 t + \delta x_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta x_{t-1} + \varepsilon_t; \tag{3.16}$$

- β_1 é o ponto de interceto ou *drift* da série;
- β_2 é o coeficiente da tendência;
- δ é o coeficiente de presença de raiz unitária;
- m é o número de desfasagem tomada na série.

No teste ADF, consideram-se as seguintes hipóteses

- $H_0 : \delta = 0$ existe pelo menos uma raiz unitária.
- $H_1 : \delta \neq 0$ a série é estacionária.

Será feita então uma regressão de Δx_t em Δx_{t-1},...,Δx_{t+p-1}, e logo a seguir é calculada a estatística T, que é dada por:

$$T = \frac{\hat{\delta}}{se(\hat{\delta})}, \tag{3.17}$$

onde $\hat{\delta}$ é o estimador para δ e $se(\hat{\delta})$ o estimador para o desvio padrão do erro de δ. Os valores críticos de T são calculados por Dickley e Fuller através da simulação Monte Carlo [25].

3.4.2.2 Teste de Phillips-Perron

Esse teste é a generalização do teste de Dickley-Fuller para os casos em que os erros $\{\varepsilon_t\}$, $t \in \mathbb{Z}$ são correlacionados e, possivelmente, heterocedásticos. Baseando-se na regressão

$$\Delta x_t = \beta_1 + \beta_2 t + \delta x_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta x_{t-1} + \varepsilon_t, \tag{3.18}$$

a estatística Z é calculada por:

$$Z = n\hat{\delta}_n - \frac{n^2\hat{\delta}^2}{2s_n^2}(\hat{\lambda_n^2} - \hat{\gamma_{0,n}}), onde: \tag{3.19}$$

- $\hat{\gamma}_{j,n} = \frac{1}{n}\sum_{i=1+j}^{n} r_i r_{i-j}$;
- $\hat{\lambda}_n^2 = \hat{\gamma}_{0,n} + 2\sum_{j=1}^{q}\left(1 - \frac{j}{q+1}\right)\hat{\gamma}_{j,n}$;
- $s_n^2 = \frac{1}{n-k}\sum_{j=1}^{n} r_i^2$.

com, r_i o resíduo de x_i utilizando estimadores de mínimos quadrados, k o número de covariáveis na regressão e q o número de desfasagem utilizadas para calcular $\hat{\lambda}_n^2$.

Quando o processo for não correlacionado as covariâncias são nulas e então $\hat{\lambda}_n^2 = \hat{\gamma}_{0,n}$. Se o processo não for heterocedástico teremos $se(\delta) = \frac{1}{n}$ e Z será dado por $Z = n\hat{\delta} = \frac{\hat{\delta}}{se(\hat{\delta})}$, ou seja, Z transforma-se numa estatística de Dickley-Fuller, portanto, terá a mesma distribuição do teste ADF.

3.4.2.3 Teste KPSS

Esse teste foi criado por Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt e Youngcheol Shin, e tem por objetivo determinar a estacionaridade duma série temporal. As hipóteses para se determinar a estacionaridade são:

- H_0: A série é estacionária;
- H_1: A série apresenta raiz unitária.

A estatística do teste KPSS é definida por:

$$LM = \sum_{t=1}^{N} \frac{s_t^2}{N^2 \hat{\sigma}_{\varepsilon}^2}, \tag{3.20}$$

onde s_t é a soma parcial dos resíduos e $\hat{\sigma}_{\varepsilon}^2$ o estimador para a variância dos erros da regressão.

Capítulo 4

Alguns modelos para previsão de séries temporais

Um método de previsão, normalmente, tem associado alguns procedimentos que, conforme os dados históricos disponíveis, permite prever o comportamento dos mesmos no futuro. Os métodos de previsão de séries temporais têm como base a suposição de que os dados passados contêm as informações sobre o padrão de comportamento da série.

De acordo com Chatfield [6], podemos classificar os métodos de previsão em três tipos:

- **As previsões por julgamento** (*Judgemental*), que se baseia no julgamento subjetivo, intuição e na experiência sem qualquer outra informação relevante;
- **Métodos univariados**, em que as previsões dependem apenas dos valores passados de uma única variável, podendo ser auxiliada por uma função de tempo ou por uma tendência linear;
- **Métodos multivariados**, onde as previsões de uma variável dependem dos valores de uma ou mais variáveis adicionais, chamados variáveis explicativas.

Existem vários métodos dos mais simples aos mais complexos, mas nem sempre os mais complexos produzem os melhores resultados, por isso será necessário avaliar as vantagens dos métodos antes de se iniciar a previsão. Ao escolher o método, deve-se ter em atenção alguns instrumentos para avaliação do erro.

- Análise gráfica;
- Diagrama de dispersão;
- Coeficiente de correlação;
- Erro quadrático acumulado;
- Raiz do erro médio quadrático (RMSE);
- Erro percentual absoluto (MAPE).

Podem-se classificar os modelos para séries temporais em duas classes, de acordo como o número de parâmetros envolvidos [12]:

- **Modelos paramétricos**, onde o número de parâmetros é finito;
- **Modelos não paramétricos**, o número de parâmetros é infinito.

Nos modelos *paramétricos*, a análise é feita no domínio do tempo. De entre esses modelos, os mais utilizados são os modelos de regressão (ou de erro), os modelos autoregressivos de médias móveis (ARMA) e os modelos autoregressivos integrados de médias móveis (ARIMA). Os modelos não paramétricos mais frequentemente usados são a função de autocovariância (ou autocorrelação) e sua transformada de Fourier, o *espectro*.

Pode-se escrever uma série temporal observada na forma

$$X_t = f(t) + \varepsilon_t, t = 1, ..., n, \tag{4.1}$$

onde $f(t)$ é denominado *sinal* e ε_t o *ruído*.

De acordo com as hipóteses estabelecidas para (4.1), pode-se ter duas classes de modelos, os modelos de regressão ou os modelos ARIMA.

4.1 Modelos de erros ou de regressão

Os modelos de regressão são os mais clássicos e provavelmente os primeiros a serem utilizados [12]. Nesses modelos, o sinal de $f(t)$ em (4.1), é uma função completamente determinada

(parte sistemática determinística) e ε_t é uma sequência aleatória, independente de $f(t)$ [12]. Supõe-se que as variáveis aleatórias ε_t não são correlacionadas, têm média zero e variância constante, isto é,

$$E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma_\varepsilon^2, E(\varepsilon_t, \varepsilon_s) = 0, \text{ para } s \neq t. \tag{4.2}$$

Nessas condições, a série ε_t é chamada *ruído branco*, como mencionado anteriormente.

Desta maneira, qualquer efeito do tempo influencia somente a parte determinística $f(t)$ e os modelos onde X_t depende funcionalmente de $X_{t-1}, X_{t-2}, ...$ não estão incluídos em (4.1) com estas suposições. Seguem-se alguns exemplos desses métodos.

4.1.1 Modelo de tendência linear

$$X_t = \alpha + \beta t + \varepsilon_t, \text{ com } t = 1, ..., n \tag{4.3}$$

com $f(t) = \alpha + \beta t$, que é uma função linear dos parâmetros.

4.1.2 Modelo de regressão:

$$X_t = \alpha + \beta x_t + \varepsilon_t, \text{ com } t = 1, ..., n \tag{4.4}$$

com $f(t) = \alpha + \beta x_t$, sendo x_t uma quantidade observável e $f(t)$ um função linear de parâmetros. Nestes casos os parâmetros podem ser estimados usando o método de mínimos quadrados.

4.1.3 Modelos de curva de crescimento

$$X_t = \alpha + e^{\beta t + \varepsilon_t}, \text{ ou } X_t = \log \alpha + \beta t + \varepsilon_t. \tag{4.5}$$

Neste caso, $f(t)$ não é uma função linear dos parâmetros, embora $log(X_t)$ o seja. Segundo Hamilton [12], normalmente, há dois tipos diferentes de funções para $f(t)$

1. Polinómio em t, em geral, de grau baixo, da forma

$$f(t) = \beta_0 + \beta_1 t + ... + \beta_m t^m. \tag{4.6}$$

de modo que a componente sistemática move-se lentamente, suavemente e progressivamente ao longo do tempo. A função $f(t)$ representa uma *tendência polinomial determinística de grau* m. Daí resulta que o processo X_t será não estacionário, se $m > 0$.

2. Polinómio harmónico, ou seja, uma combinação linear de senos e cossenos com coeficientes constantes, da forma

$$f(t) = \sum_{n=1}^{p} (\alpha_n cos\lambda_n t + \beta_n sen\lambda_n t), \tag{4.7}$$

com $\lambda_n = \frac{2\pi n}{p}$, se $f(t)$ tem período p.

O modelo de erro é clássico para análise de séries económicas, onde $f(t)$ é composta da adição ou multiplicação de ambos os tipos de função, onde (4.6) representará a tendência e (4.7) representará as variações sazonais, isto é, $f(t) = T_t + S_t$, donde

$$X_t = T_t + S_t + \varepsilon_t. \tag{4.8}$$

Normalmente, T_t é a componente da tendência, enquanto S_t é a componente sazonal.

Segundo Morettin [17], algumas vezes, o sinal de $f(t)$ no modelo (4.1), não pode ser aproximado por uma função simples do tempo, como em (4.6). Nesse caso para a estimação da tendência tem de se utilizar procedimentos não paramétricos de suavização. **Suavização ou alisamento** é um processo que transforma a série X_t, no instante t numa outra série dada por

$$X_t^* = \sum_{k=-n}^{n} b_k X_{t+k}, \text{ onde, } t = n+1, ..., N-n. \tag{4.9}$$

Usamos $2n+1$ observações ao redor do instante t para estimar a tendência naquele instante. Nesse caso perdemos n observações no início da série e outras n no final da série (4.9) chama-se um *filtro linear* e normalmente tem-se que $\sum_{k=-n}^{n} b_k = 1$.

4.2 Modelos ARIMA

A hipótese de que os erros não são correlacionados introduz limitações quanto a validade e aplicabilidade dos modelos do tipo (4.1), para descrever o comportamento de séries económicas, onde os erros normalmente são autocorrelacionados e influenciam a evolução do processo [12].

Nessas situações, os modelos ARIMA são muito úteis para a previsão. Os modelos *Autoregressivos Integrados de Médias Móveis* (ARIMA) são modelos fundamentais para séries univariáveis. O modelo ARIMA é composto por três componentes chave:

- *Componente autoregressivo* que é a relação entre a variável dependente atual e a variável dependente em períodos de tempo defasados;
- *Componente integrado* que se refere a transformação dos dados, subtraindo os valores anteriores de uma variável dos valores atuais da mesma variável de modo a tornar os dados estacionários;
- *Componente Média Móvel*, que se refere à dependência entre a variável dependente e os valores anteriores de um termo estocástico.

Podem-se descrever duas classes de processos pelo método ARIMA:

- Processos lineares estacionários.

Os processos lineares estacionários podem ser representados na forma

$$X_t - m = \epsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + ... = \sum_{k=0}^{\infty} \phi_k \varepsilon_{t-k}, \text{ com } \phi_0 = 1. \tag{4.10}$$

Na expressão (4.10), como mencionado anteriormente, ε_t é ruído branco e $m = E(X_t)$; e $\phi_1, \phi_2, ...$ a sequência de parâmetros tal que

$$\sum_{k=0}^{\infty} \phi_k^2 < \infty. \tag{4.11}$$

- Processos lineares não estacionários homogéneos.

Os processos lineares não estacionários constituem uma generalização dos processos lineares estacionários, que pressupõem que o mecanismo gerador da série produz erros autocorrelacionados e que as séries são não estacionárias. Entretanto, estas séries podem se tornar estacionárias através de um número finito de diferenças.

Esses processos são descritos de maneira adequada pelos métodos autoregressivos integrados de médias móveis de ordem p, d, e q $ARIMA(p, d, q)$ que podem ser generalizados pela inclusão de um operador sazonal.

4.2.1 Modelos autoregressivos

Num modelo autoregressivo, faz-se a previsão da varíavel de interesse usando combinação linear dos valores passados dessa variável. O termo *autoregressivo* indica que é uma regressão da variável contra ela mesma [13]. Assim um modelo autoregressivo de ordem p pode ser escrito como:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t \tag{4.12}$$

com ε_t ruído branco e $\phi_0, \phi_1, ..., \phi_p$ parâmetros reais. Refere-se a (4.12) como **modelo** $AR(p)$ e lê-se *modelo autoregressivo de ordem p*.

Para um modelo $AR(1)$, modelo autoregressivo de ordem 1, temos os seguintes casos:

- para $\phi_1 = 0$, x_t é equivalente a um **ruído branco**;
- para $\phi_1 = 1$ e $\phi_0 = 0$, x_t é equivalente a um **caminho aleatório**(*Random Walk*);

- para $\phi_1 = 1$ e $\phi_0 \neq 0$, x_t é equivalente a um caminho aleatório com deriva;
- para $\phi_1 < 0$, x_t tende a oscilar à volta da média.

Normalmente restringem-se modelos autoregressivos aos dados estacionários, implicando algumas restrições nos valores dos parâmetros requeridos [13].

- para um modelo $AR(1)$: $-1 < \phi_1 < 1$;
- para um modelo $AR(2)$: $-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$ e $\phi_2 + \phi_1 < 1$, quando $p \geq 3$, as restrições tornam-se muito mais complexas.

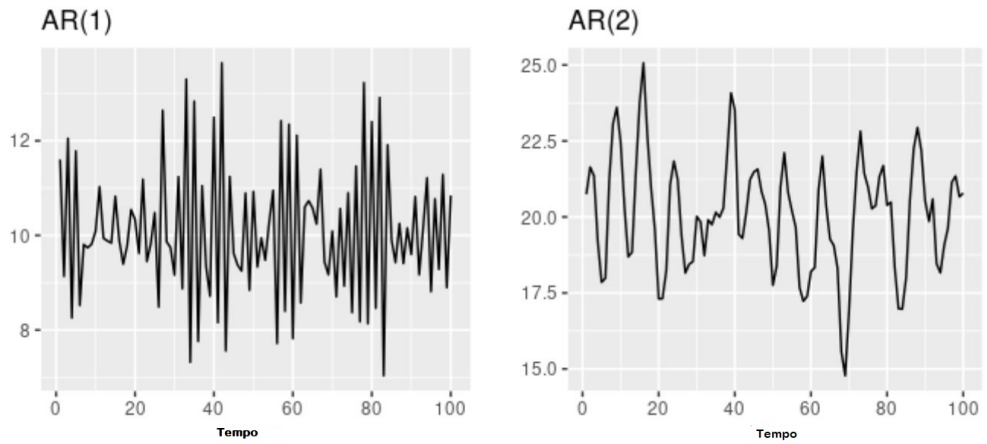

Figura 4.1: Dados de modelos autoregressivos com parâmetros diferentes

A figura 4.1 representa séries de um modelo $AR(1)$ e $AR(2)$ (Fonte: [13]). Alterando os parâmetros $\phi_1, ..., \phi_p$, altera o padrão da série. A variância do termo de erro ε_t somente irá mudar a escala das séries, não o padrão. No exemplo da figura 4.1 para $AR(1)$ têm-se $x_t = 18 - 0.8x_{t-1} + \varepsilon_t$ enquanto para o modelo $AR(2)$ têm-se $x_t = 8 + 1.3x_{t-1} - 0.7x_{t-2} + \varepsilon_t$. Em ambos os casos, ε_t é um ruído branco normalmente distribuído com média 0 e variância 1.

4.2.2 Modelos médias móveis

Ao contrário dos modelos autoregressivos que usam valores passados da variável a ser determinada a previsão, os modelos médias móveis usam erro de previsões anteriores num modelo semelhante a uma regressão

$$x_t = \phi_0 + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2 x_{t-2} + ... + \phi_q x_{t-q} \tag{4.13}$$

onde ε_t é um ruído branco. Denomina-se modelo $MA(q)$, ou seja, modelo média móvel de ordem q. Cada valor de x_t pode ser pensado como uma média móvel ponderada dos últimos erros da previsão.

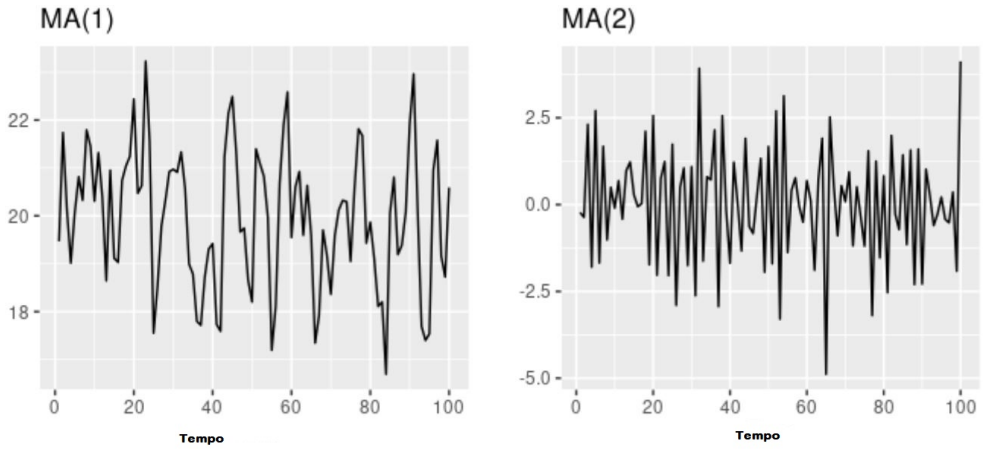

Figura 4.2: Dados de modelos médias móveis com parâmetros diferentes

A Figura 4.2, mostra dados de um modelo $MA(1)$ e de um modelo $MA(2)$ (Fonte: [13]). Alterando os parâmetros $\phi_1, ..., \phi_q$ resulta em diferentes padrões da série. Semelhantemente aos modelos autoregressivos, a variância do termo de erro ε_t somente mudará a escala das séries, não o padrão. Na Figura 4.2 para $MA(1)$ tem-se $x_t = 20 + \varepsilon_t + 0.8\varepsilon_{t-1}$, e para $MA(2)$ tem-se $x_t = \varepsilon_t - \varepsilon_{t-1} + 0.8\varepsilon_{t-2}$. Nos dois casos ε_t é um ruído branco normalmente distribuído com média zero e variância um.

É possível escrever qualquer modelo estacionário $AR(p)$ como um modelo $MA(\infty)$ [13]. Pode-se demonstrar isso para um modelo $AR(1)$ usando substituição.

$$
\begin{aligned}
x_t &= \phi_1 x_{t-1} + \varepsilon_t \\
&= \phi_1(\phi_1 x_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
&= \phi_1^2 x_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\
&= \phi_1^3 x_{t-3} + \phi_1^2 x_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\
etc...
\end{aligned}
$$

Dado $-1 < \phi < 1$, o valor de ϕ_1^k vai ficando menor quando k for aumentando. Finalmente teremos

$$x_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + ... \tag{4.14}$$

um processo $MA(\infty)$.

4.2.3 Função de autocorrelação parcial (FACP)

Um dos problemas na construção de modelos autoregressivos é a identificação da ordem do respetivo processo. Para modelos $MA(q)$ isso não é um problema relevante pois, se o processo for de ordem q, as funções de autocorrelação apresentarão valores próximos de zero para desfasamentos superiores a q. Embora algumas informações sobre a ordem de um processo autoregressivo possam ser obtidas a partir do comportamento oscilatório das funções de autocorrelação, mais informações podem ser obtidas a partir da *função de autocorrelação parcial (FACP)* [18]

Para entender o que é uma função de autorrelação parcial e como pode ser usada, vamos primeiro considerar as funções de covariância e autocorrelação para o processo autoregressivo de ordem p. Primeiro notamos que a covariância com desfasamento k é determinado por:

$$\gamma_k = E[y_{t-k}(\phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t)]. \tag{4.15}$$

Seja $k = 0, 1, ..., p$, obtém-se as seguintes p+1 equações de diferenças que podem ser resolvidas

simultaneamente para $\gamma_0, \gamma_1, ..., \gamma_p$:

$$\begin{aligned}
\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + ... + \phi_p\gamma_p + \sigma_\varepsilon^2 \\
\gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 + ... + \phi_p\gamma_{p-1} \\
&..... \\
\gamma_p &= \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + ... + \phi_p\gamma_0.
\end{aligned} \tag{4.16}$$

Para $k > p$ as covariâncias são determinadas por

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + ... + \phi_p\gamma_{k-p}. \tag{4.17}$$

Dividindo as equações em (4.16) por γ_0, pode-se derivar um conjunto de p equações que juntos determinam os primeiros p valores da função de autocorrelação:

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + ... + \phi_p\rho_{p-1} \\
&..... \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + ... + \phi_p.
\end{aligned} \tag{4.18}$$

Para $k > p$, da equação (4.17), temos que:

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ... + \phi_p\rho_{k-p}. \tag{4.19}$$

As equações em (4.18) são chamadas **Equações de Yule-Walker**. Se $\rho_1, \rho_2, ..., \rho_p$ forem conhecidos, as equações podem ser resolvidas por $\phi_1, \phi_2, ..., \phi_p$.

Resolvem-se as equações de Yule-walker, iniciando em $p = 0$ e determinando a respetiva estimativa de $\phi_i, i = 1, ..., p$, digamos a_i. Assim para $p = 1, 2, ...$ obteremos a série $a_1, a_2, ...$ que é a *função de autocorrelação parcial (FACP).*

4.2.4 Modelos autoregressivos de médias móveis (ARMA)

Os modelos autoregressivos e de médias móveis são representados por ARMA(p,q), e é composto pela junção dos modelos autoregressivos (AR) e dos modelos de médias móveis (MA).

Definição 4.1 (Modelos ARMA)**.** Uma série temporal x_t com $t \in \mathbb{Z}$ é um **ARMA(p,q)** se for estacionária e se

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ...\phi_p x_{t-p} + ... + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}, \tag{4.20}$$

com $\phi_p \neq 0, \phi_q \neq 0$ e $\sigma_w^2 > 0$. Os parâmetros p e q denotam a ordem da componente autoregressiva e a ordem da componente de média móvel, respetivamente. Se x_t tiver média não nula μ, estabelecemos $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$ e escrevemos o modelo como:

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ...\phi_p x_{t-p} + ... + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}, \tag{4.21}$$

onde w_t será considerado um ruído branco Gaussiano, a menos que seja especificado de outra forma.

Quando $q = 0$, o modelo é chamado autoregressivo de ordem p, $AR(p)$ e quando $p = 0$ o modelo é chamado média móvel de ordem q, $MA(q)$. Usando as notações dos modelos AR e MA, podemos escrever (4.20) na seguinte forma:

$$\phi(B)x_t = \theta(B)w_t \tag{4.22}$$

4.2.5 Modelos autoregressivos integrados e de médias móveis (ARIMA)

Na prática, a maioria das séries temporais em economia são *não estacionários*. Nesse caso os modelos ARIMA apresentam-se como modelos muito úteis para esses tipos de séries.

Os modelos $ARIMA(p, d, q)$ são uma generalização de modelos $ARMA$ usados para a previsão de séries temporais que não sendo estacionárias, podem ser transformadas em séries estacionárias por diferenciações, com $p, d, q > 0$ onde:

- p é o número de termos autoregressivos;
- d é o número de diferenciações para que a série se torne estacionária;
- q é o número de termos de médias móveis

Definição 4.2 (Modelos ARIMA)**.** Um processo x_t é um modelo $ARIMA(p, d, q)$ se $\nabla^d x_t = (1 - B)^d x_t$ for um modelo $ARMA(p, q)$. Em geral, escreveremos o modelo como:

$$\phi(B)(1 - B)^d x_t = \theta(B) w_t. \tag{4.23}$$

Se $E(\nabla^d x_t) = \mu$, escrevemos o modelo como:

$$\phi(B)(1 - B)^d x_t = \alpha + \theta(B) w_t, \tag{4.24}$$

onde $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$.

4.2.5.1 Estimação de modelos ARIMA

Usando as funções de autocorrelação e autocorrelação parcial determina-se a ordem d do modelo, ou seja, o número de vezes que a série precisa de ser diferenciada. Uma vez determinada a ordem do modelo, os parâmetros podem ser determinados utilizando:

- Método dos momentos;
- Estimadores de mínimos quadrados;
- Estimadores de máxima verosimilhança.

4.2.5.2 Diagnóstico de modelos ARIMA

Uma vez determinado o grau d e feita a estimação dos parâmetros, deve-se verificar se o modelo está adequado aos dados da série. Essa adequação é verificada recorrendo a análise

de resíduos.

Seja o modelo ajustado

$$\phi(B)w_t = \theta(B)\varepsilon_t, \tag{4.25}$$

onde $w_t = \nabla^d x_t$ com d diferenças nas séries, $\phi(B)$ e $\theta(B)$ polinómios do operador B. Os resíduos são descritos por:

$$\varepsilon_t = \phi^{-1}(B)\phi(B)w_t. \tag{4.26}$$

O modelo será considerado adequado aos dados se os resíduos ε_t apresentarem um comportamento aleatório, com média zero e variância constante, compatível com uma distribuição $N(0,1)$. Para se verificar se os resíduos estimados são *não correlacionados* podemos utilizar os testes de $Box - Pierce$ ou o teste de $Ljung - Box$. Para a escolha do modelo ARIMA podem ser utilizados os critérios AIC, BIC e HDC.

4.2.5.2.1 Testes de Box-Pierce e Ljung-Box Os testes de $Box - Pierce$ e $Ljung - Box$ são usados para testar o ajustamento de um modelo aos dados de uma série temporal. São testes aplicados aos resíduos de uma série temporal, após o ajustamento de um modelo. São analisadas m autocorrelações dos resíduos e quanto menor for esse valor melhor o ajustamento do modelo escolhido [25].

A estatística de $Box - Pierce$ pode ser definida por:

$$Q = N \sum_{k=1}^{m} \hat{r}_k^2, \tag{4.27}$$

onde, N é o tamanho da amostra, m a duração da desfasagem e $\hat{r}_k$ a autocorrelação estimada da série.

Para amostras grandes, a estatística Q será uma distribuição qui-quadrado com m graus de liberdade (gl).

O teste de $Ljung - Box$ é uma generalização do teste de $Box - Pierce$ e apresenta melhores resultados [25].

$$LB = N(N+2)\sum_{k=1}^{m}\left(\frac{\hat{r}_k^2}{N-K}\right). \tag{4.28}$$

Para os dois testes acima, a hipótese nula considera a não existência de autocorrelação conjunta dos resíduos. H_0 será rejeitada se o valor da estatística de teste for superior ao valor crítico da distribuição qui-quadrado para um determinado nível de significância α e graus de liberdade $gl = m$, isto é, as autocorrelações para um determinado número de *lags* (passos) poderão ser significativamente diferentes de zero, mostrando que os valores não são aleatórios e independentes ao longo do tempo.

4.2.5.3 Critério de escolha dos modelos

Existem vários critérios utilizados na escolha do modelo que melhor se ajusta aos dados. Vamos destacar neste trabalho dois desses critérios, o critério de informação de Akaike e o critério de informação Bayesiano.

4.2.5.3.1 Critério de Informação de Akaike (AIC) O critério de AIC baseia-se na existência de um modelo "real" que é desconhecido, mas que permite descrever os dados. A estimativa do AIC para um modelo é dado por $AIC = -2L + 2K$, onde L é a log-verosimilhança máxima e K é o número de parâmetros do modelo. O modelo com menor valor de AIC é considerado o modelo de melhor ajuste [25].

4.2.5.3.2 Critério de Informação Bayesiano (BIC) O critério BIC é definido como a estatística que maximiza a probabilidade de se identificar o melhor modelo entre um grupo de modelos em estudo. A estimativa do BIC para um determinado modelo é dado por $BIC = -2L + 2Kln(n)$, onde L é a log-verosimilhança máxima, K é o número de parâmetros do modelo e n o número de observações. O modelo com menor BIC é considerado o que melhor se ajusta aos dados.

4.2.5.4 Previsão com modelos ARIMA

Após ser estimado e escolhido o modelo, este pode ser utilizado para fazer previsões de valores futuros da série temporal. Dado o modelo ARIMA pretende-se obter a previsão de y_t para o período $T+h, h>1$. Denotamos essa previsão por $\hat{y}_T(h)$, designada por *previsão de origem T e horizonte h* [18].

As previsões podem ser calculadas seguindo os seguintes passos [13]:

1. Expandir a equação ARIMA tal que y_t fique do lado esquerdo da equação e todos os restantes termos do lado direito;
2. Substituir na equação t por $T+h$;
3. No lado direito da equação, substituir as observações futuras pelas respetivas previsões, erros futuros por zero e erros passados pelos resíduos correspondentes.

Iniciando com $h=1$, esses passos são repetidos para $h=2,3,...$ até todas as previsões forem calculadas.

Uma vez determinado o modelo $ARIMA(p,d,q)$, estacionário, invertível e com parâmetros conhecidos, pode-se utilizar uma das seguintes fórmulas para fazer a previsão.

1. Fórmula de equação de diferenças

$$X_{t+h} = \phi X_{t+h-1} + ... + \phi_{p+d} X_{t+h-p-d} - \theta_1 \varepsilon_{t+h-1} - ... - \theta_q \varepsilon_{t+h-q} + \varepsilon_{t+h}. \tag{4.29}$$

2. Fórmula de choques aleatórios

$$X_{t+h} = \sum_{j=-\infty}^{t+h} \psi_{t+h-j} \varepsilon_j = \sum_{j=0}^{\infty} \psi \varepsilon_{t+h-j} \tag{4.30}$$

com $\psi_0 = 1$ e os outros pesos obtidos pela resolução do sistema de operadores $\theta(B) = \phi(B)\psi(B)$.

3. Fórmula invertida

$$X_{t+h} = \sum_{j=1}^{\infty} \pi_j X_{t+h-j} + \varepsilon_{t+h} \tag{4.31}$$

com os π_j obtidos pela resolução do sistema de operadores $\phi(B) = \theta(B)\pi(B)$.

No ajustamento de um modelo ARIMA a um conjunto de dados em forma de série temporal (não sazonal), podemos então utilizar os seguintes passos [13]:

1. Representar graficamente os dados e identificar observações não usuais e identificar o comportamento dos dados;
2. Se necessário, transformar os dados (usando a transformação Box-Cox) para estabilizar a variância;
3. Se as séries forem não estacionários, tomar as primeiras diferenças até que se tornem estacionárias;
4. Examinar a função de autocorrelação e autocorrelação parcial e identificar se um modelo $ARIMA(p, q, 0)$ ou $ARIMA(0, d, q)$ são apropriados;
5. Testar o modelo(s) escolhido(s), e usar o critério AIC para procurar o melhor modelo;
6. Verificar os resíduos do modelo escolhido representando graficamente a função de autocorrelação dos resíduos e fazer teste aos resíduos. Se não aparecerem como um ruído branco, tentar um modelo modificado;
7. Quando os resíduos aparecerem como um ruído branco, calcular as previsões.

4.2.6 Modelos ARIMA sazonais (SARIMA)

Os modelos ARIMA também são capazes de modelar dados sazonais. Um modelo ARIMA sazonal (SARIMA) é formado pela inclusão de termos sazonais no modelo ARIMA. É representado como $SARIMA(p, d, q)(P, D, Q)_m$, onde (p, d, q) é a parte não sazonal do modelo, $(P, D, Q)_m$ a parte sazonal e m o número de observações por ano.

Definição 4.3 (Modelos SARIMA)**.** Os modelos SARIMA são dados por:

$$\phi(B)\Phi(B^s)W_t = \theta(B)\Theta(B^s)\varepsilon_t \tag{4.32}$$

com $W_t = \nabla_s^d \nabla^d X_t$ e com os seguintes operadores

- Autoregressivo não sazonal - $\phi(B) = (1 - \alpha_1 B - ... - \alpha_p B^p)$;
- Autoregressivo sazonal - $\Phi(B^s) = (1 - \phi_s B^s - ... - \phi_p B^{P_s}$;
- Média móvel não sazonal - $\theta(B) = (1 + \beta_1 B + ... + \beta_q B^q)$;
- Média móvel sazonal - $\Theta(B^s) = (1 + \theta_s B^s - ... - \theta_p B^{Q_s})$;
- ∇^d - operador diferença não sazonal de ordem d;
- ∇_s^d - operador diferença sazonal de ordem D.

Capítulo 5

Uma aplicação a dados da economia Cabo-Verdiana

5.1 Produto interno bruto de Cabo Verde

O produto interno bruto (PIB) representa a soma em unidades monetárias de todos os bens e serviços produzidos num determinado país ou região, durante um determinado período. O PIB é um indicador macroeconómico muito utilizado na economia, para analisar a "saúde económica" de um país ou região. O PIB é composto de bens e serviços produzidos para venda no mercado e também inclui produção não destinada ao mercado, tais como educação ou serviços da defesa de um país fornecidos pelo governo. Um conceito alternativo é o produto nacional bruto (PNB), que soma toda a produção dos residentes de um país. Por exemplo, se uma empresa com dono Cabo-Verdiano tiver uma fábrica no Senegal, a produção dessa fábrica será contabilizada como PIB do Senegal e como PNB de Cabo Verde [5].

5.1.1 Breve enquadramento da economia Cabo-Verdiana

A Economia Cabo-Verdiana é norteada pelo turismo, que representa cerca de 25% do PIB. Apesar dos desafios relacionados com a pequena economia insular, Cabo Verde tem

tido um progresso económico notável desde 1990, influenciado principalmente pelo rápido desenvolvimento do turismo.

Antes da pandemia da COVID-19, Cabo Verde experimentava um crescimento económico robusto direcionado por um forte setor turístico e por reformas estruturais firmes. Entre 2016 e 2019 o crescimento médio do PIB foi de 4.7%. Condições globais favoráveis e fortes reformas estruturais, principalmente no setor das empresas do estado, contribuíram para o crescimento. O crescimento robusto e sustentável levou a uma diminuição da pobreza de 24.5% em 2015 para 11.5% em 2019 [3].

Devido à paralisação do turismo, o PIB contraiu 14,8% em 2020 (15,7% em termos per capita) - a maior redução já registada no país e uma das maiores de África em 2020. O país depende muito do turismo, setor que representa 25% do PIB e movimenta quase 40% de toda a atividade económica. A crise reverteu o progresso na redução da pobreza atingida desde 2015. A dívida global aumentou de 1.8% em 2019 para 2.9% em 2020, alavancada principalmente pelo impacto da crise na receita fiscal[3].

5.1.2 Análise descritiva dos dados

Na Figura 5.1 podemos verificar a tendência de crescimento do PIB de Cabo Verde e a queda de 2020 referida acima. Entretanto, neste trabalho usaremos dados do PIB de 1980 a 2018, para análise e faremos previsão de 2019 a 2029, comparando assim com valores reais de 2019 e 2020.

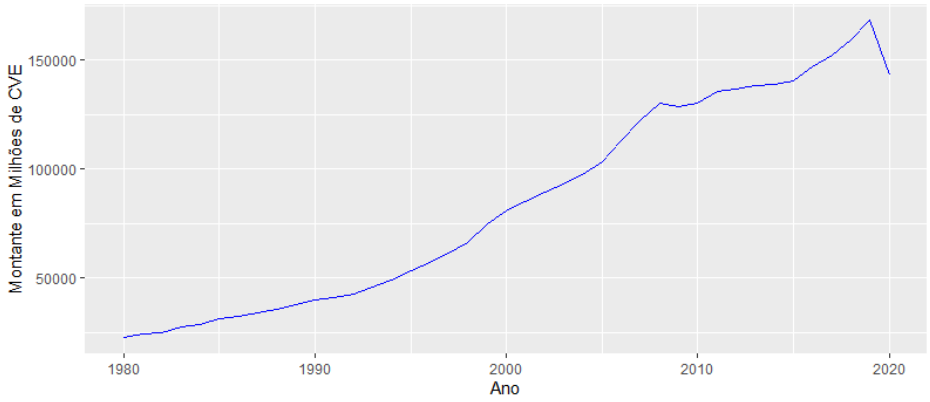

Figura 5.1: Evolução do PIB(em volume) de Cabo Verde de 1980–2020 (Fonte: INE).

No R, usando o *package* fBasics, podemos determinar rapidamente algumas estatísticas básicas dos dados. Na Tabela 5.1 apresentam-se esses valores. A grande evolução nos valores do PIB durante os anos observados está patente nos valores da amplitude amostral e também da amplitude inter-quartis. O valor da mediana mostra que em metade dos anos observados o valor do PIB excedeu os 74595.34 unidades e a assimetria negativa revela haver mais concentração de anos com valores de PIB mais elevados, comparativamente com valores mais baixos.

Tabela 5.1: Análise descritiva dos dados.

nobs	39.00
NAs	0.00
Minimum	22360.47
Maximum	159239.55
1. Quartile	38743.66
3. Quartile	129259.29
Mean	80766.96
Median	74595.34
Sum	3149911.25
SE Mean	7254.05
LCL Mean	66081.89
UCL Mean	95452.02
Variance	2052230316.03
Stdev	45301.55
Skewness	0.24
Kurtosis	-1.54

5.1.3 Normalidade dos dados

Tendo em atenção que muitos dos resultados na modelação de dados dependem das características distribucionais dos dados, é importante perceber o seu comportamento, nomeadamente, se são compatíveis com a proveniência de uma distribuição Normal.

A representação dos dados num histograma (Figura 5.2) não está de acordo com aquilo que seria de esperar se os dados fossem oriundos de uma distribuição Normal. Desta feita vamos aplicar o teste de *Shapiro-Wilk* para análise da normalidade dos dados. Usando a função `shapiro.test` no `R` obtemos um *p-value* $p = 0.001746 < 0.05$, ou seja, rejeitamos a hipótese inicial de que os dados são normalmente distribuídos, considerando um nível de significância de 0.05.

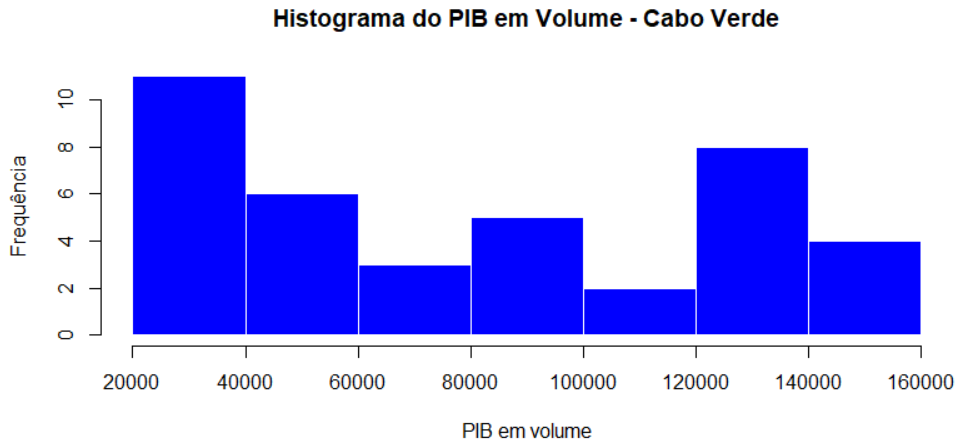

Figura 5.2: Histograma PIB(em volume) de Cabo Verde de 1980–2018 (Fonte: INE).

5.1.4 Teste da estacionaridade

Antes de iniciar qualquer análise é necessário verificar a estacionaridade dos dados, para podermos aplicar o modelo ARIMA. Caso a série não seja estacionária, iremos aplicar a diferenciação para a transformar numa série estacionária.

Na Figura 5.1, pode-se identificar uma tendência, o que traduz a não estacionaridade da série. Realizaremos ainda alguns testes para concluirmos sobre a estacionaridade ou não da série.

Vamos primeiramente verificar se a série original é estacionária ou não, aplicando o teste *Augmented Dickey-Fuller* [8], através da função `adf.test` no *package* `tseries` (valores de 1980 a 2018). Ao aplicar o teste *ADF* encontramos um *p-value* $p = 0.4082 > 0.05$, ou seja, com um nível de significância de 0.05 não rejeitamos a hipótese inicial de que a série é não estacionária, isto é, ela tem uma raiz unitária.

A *função de autocorrelação(ACF)* e *autocorrelação parcial (PACF)*, podem também nos ajudar na análise da estacionaridade. Como podemos verificar na Figura 5.3, a ACF decresce lentamente para zero a cada passo (*lag*), enquanto que o PACF "cai" rapidamente para zero no primeiro *lag*, indicando que a série é não estacionária, ou seja, precisa ser diferenciada pelo menos uma vez ($d = 1$).

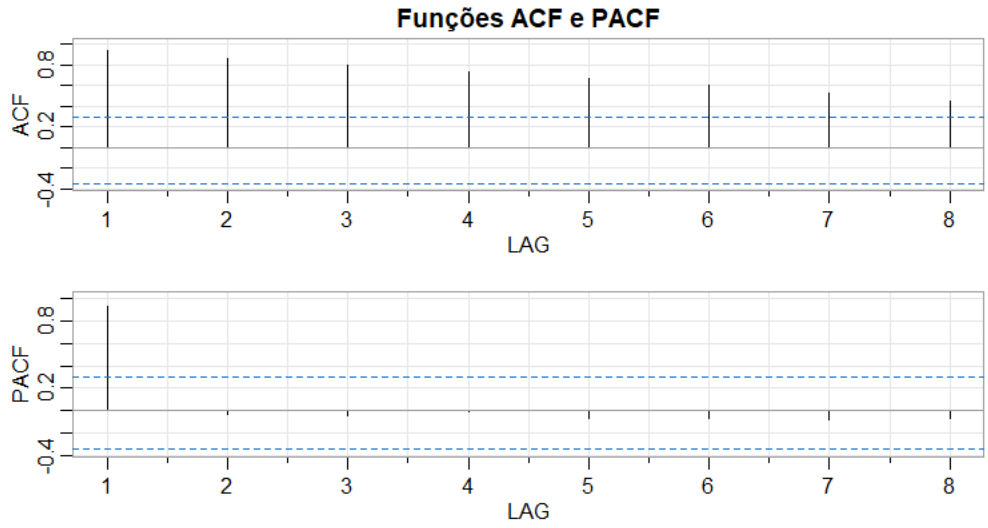

Figura 5.3: ACF e PACF do PIB de Cabo Verde de 1980 a 2018

Sendo que a série não é estacionária, teremos de aplicar a diferenciação para a transformar numa série estacionária.

Para séries económicas e financeiras, é comum o uso do **retorno**, ou **taxa de crescimento** em vez dos valores "brutos" [22]. Dada uma série X_t, a taxa de crescimento é aproximadamente $ln(x_t) - ln(x_{t-1})$, onde x_t é um ponto qualquer da série e x_{t-1} o ponto anterior. Normalmente usando a taxa de crescimento em vez dos dados brutos pode-se transformar a série numa série estacionária, sendo que $ln(X_t)$ estabiliza a variância e a diferença elimina a tendência [22]. No `R`, usando o comando `diff(log())`pode-se

determinar facilmente essa diferenciação.

Na Figura 5.4, encontra-se a representação gráfica da diferença do logaritmo natural da nossa série. Vamos aplicar o *Augmented Dickey-Fuller test*, para verificarmos se a nova série é ou não estacionária.

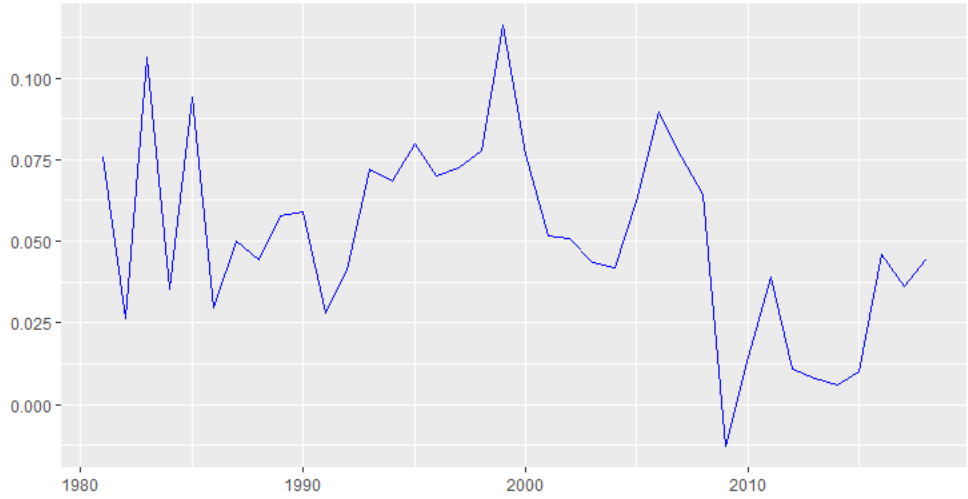

Figura 5.4: Serie log(PIB) aplicado a diferença de ordem 1

Aplicando o teste `adf`, encontramos um *p-value* $p = 0.48 > 0.05$, isto é, com um nível de significância de 0.05 continuamos a não rejeitar a hipótese inicial de que a série é não estacionária, ou seja a primeira diferença não foi suficiente para transformar a série numa estacionária.

Ao aplicar mais uma diferença à nossa série ($d = 2$), e repetindo o teste ADF encontramos um *p-value* $p = 0.01 < 0.05$, ou seja, com um nível de significância de 0.05, rejeitamos a hipótese inicial de que a série é não estacionária.

No `R`, o *package* `forecast` dispõe da função `ndiffs()` que permite identificar qual a ordem de diferenciação necessária para aplicar numa determinada série, para a transformar numa série estacionária. No nosso caso ao aplicar o `ndiffs()` encontramos 2, ou seja, a nossa série precisa de duas diferenciações para ser transformada numa série estacionária.

5.1.5 Identificação do Modelo

Uma vez que $d = 2$, o nosso modelo será do tipo $ARIMA(p, 2, q)$. Assim teremos de substituir valores para p e q e escolher o melhor modelo através do *critério de informação de Akaike(AIC)* e o *critério de informação Bayesiano(BIC)* [10], onde o melhor modelo é o que apresentar menor valor de AIC e BIC.

Na Tabela 5.2 temos os modelos para diferentes valores de p e q. O modelo $ARIMA(1, 2, 1)$ que apresenta menores valores para o AIC, mesmo sendo o BIC ligeiramente superior ao do modelo $ARIMA(0, 2, 0)$. No `R` esses valores são encontrados usando o *package* `astsa` e a função `Arima`. A tarefa de experimentar vários modelos para determinar qual o melhor, com base nos valores de AIC e BIC, fica facilitada no `R`, com a função `auto.arima` do *package* `forecast`.

Tabela 5.2: Comparação dos modelos.

Modelo	AIC	BIC	Posição
ARIMA(1,2,1)	682.15	686.98	1
ARIMA(0,2,0)	684.45	686.06	5
ARIMA(1,2,0)	685.19	688.41	6
ARIMA(0,2,1)	684.14	687.36	4
ARIMA(2,2,2)	685.74	693.80	7
ARIMA(2,2,1)	684.02	690.46	2
ARIMA(1,2,2)	684.03	690.47	3

Assim o modelo $ARIMA(1, 2, 1)$ e os seus respetivos parâmetros encontram-se resumidos na Tabela 5.3. Podemos também observar que os valores do $p - value$ realtivo à significância dos coeficientes $AR(1)$ e $MA(1)$ são pequenos (menores do que o nível de significância de 0.05), ou seja, os coeficientes são significativamente diferentes de zero.

Assim sendo, teremos agora de analisar se os resíduos são correlacionados ou não, antes de usar o modelo para fazer previsões.

Tabela 5.3: Estimativas com modelo ARIMA(1,2,1).

Variável	Coeficiente	Std. Error	t-stat	p-value
AR(1)	0.5361	0.1903	2.8165	0.00790
MA(1)	-0.9239	0.1182	-7.8174	0.00001

5.1.6 Análise dos resíduos

Uma vez identificado o melhor modelo, o próximo passo será a verificação das funções ACF e $PACF$ dos resíduos. Na Figura 5.5, encontra-se a representação gráfica dos resíduos. Para que o modelo produza bons resultados, temos de confirmar que os resíduos são não correlacionados.

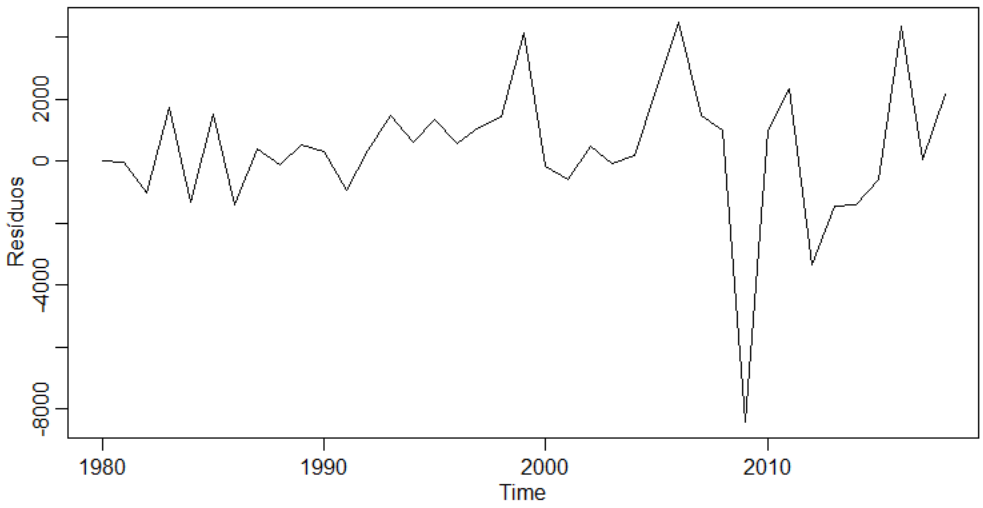

Figura 5.5: Representação gráfica dos resíduos

Na Figura 5.6 encontra-se a representação das funções ACF e $PACF$ dos resíduos. Pode-se ver que não existe autocorrelação dos resíduos, dado que se encontram dentro dos limites de confiabilidade (linha tracejada azul). Além disso, aplicando o teste de *Box-Pierce* ou *Ljung-Box* encontrou-se um $p = 0.9581$, o que indica que os resíduos se comportam como ruído branco. Conclui-se então que o nosso modelo é apropriado para fazer as previsões.

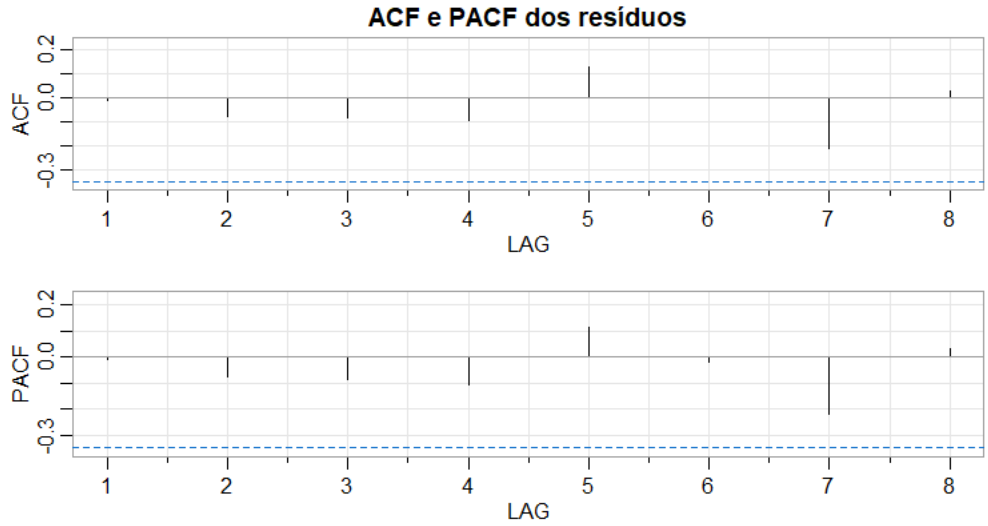

Figura 5.6: Funções ACF e PACF dos resíduos

5.1.7 Previsão

Dado que a análise dos resíduos se mostrou adequada para validação do modelo, este pode ser usado para realizar previsões com base na amostra. Na Figura 5.7 estão representados os valores efetivos de 1980 a 2018 e os valores obtidos através do nosso modelo no mesmo período. Pode-se observar que o modelo ajustado converge para a série original.

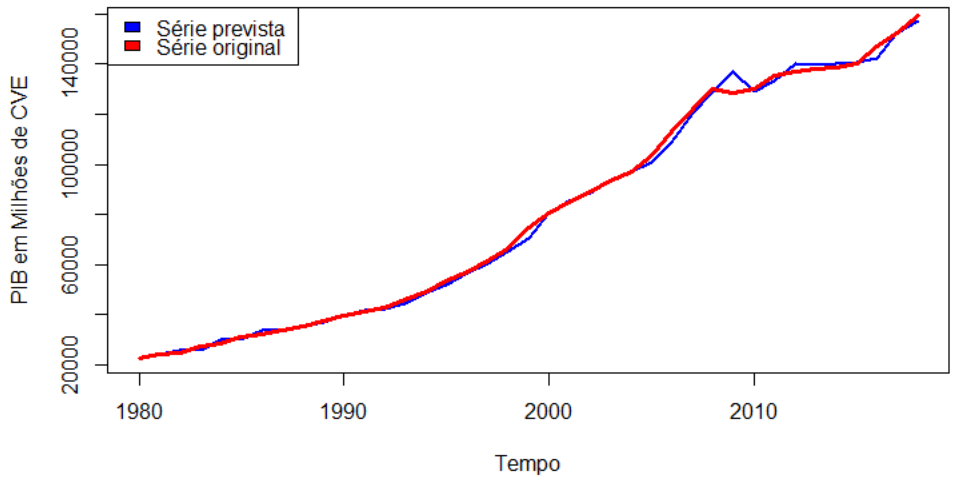

Figura 5.7: Representação gráfica da série original e série ajustada

A Figura 5.8 apresenta os dados reais de 1980 a 2020 e as previsões de 2019 a 2029, usando

o modelo $ARIMA$. Nas previsões, podemos ainda ver os limites de confiança máximo e mínimo dos valores previstos.

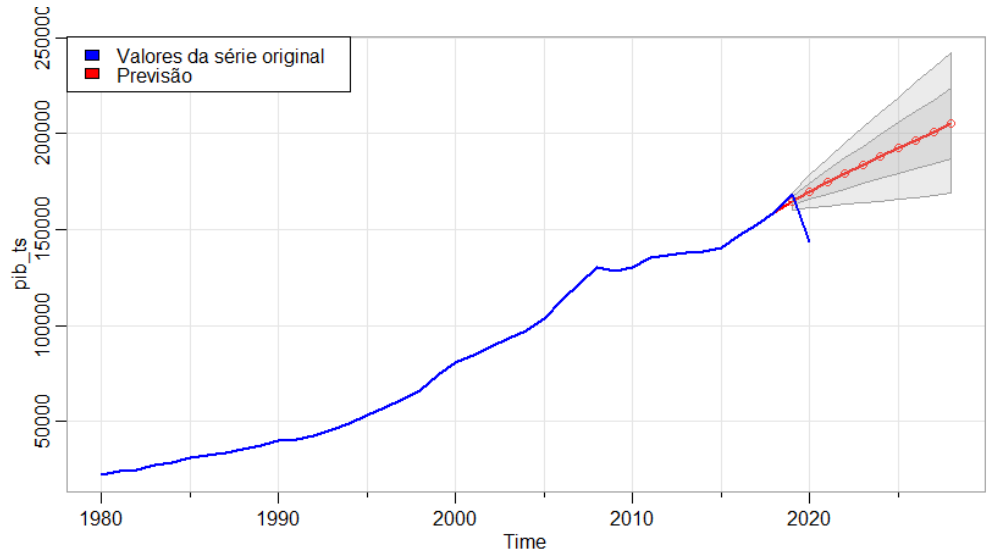

Figura 5.8: Representação gráfica da série original e série ajustada

Na Tabela 5.4, está representada a variação entre o valor previsto e o valor real. Podemos observar que para todos os anos apresentados, os valores previstos estão próximos dos valores reais com exceção do ano 2020, que como explicado anteriormente, foi um ano atípico devido à pandemia da COVID-19 que afetou todas as projeções de crescimento das economias mundiais, desafiando praticamente todos os métodos de previsão.

Tabela 5.4: Previsões com o modelo ARIMA(1,2,1).

Ano	Valores reais (Mesc)	Valores previstos(Mesc)	Var. Prev./Real
2015	140296.6	140897.1	0.42%
2016	146898.7	142548.0	-2.96%
2017	152336.9	152271.5	-0.04%
2018	159239.5	157091.2	-1.35%
2019	168264.9	164942.7	-1.97%
2020	143389.6	170002.9	18.56%
2021	NA	174718.4	NA
2022	NA	179249.1	NA

Ano	Valores reais (Mesc)	Valores previstos(Mesc)	Var. Prev./Real
2023	NA	183680.8	NA
2024	NA	188059.4	NA

5.1.8 Considerações finais

Dos resultados obtidos, é percetível que o modelo $ARIMA$ mostrou-se muito potente na previsão do PIB de Cabo Verde, obtendo valores muito próximos dos valores reais. Também identificamos que uma das fraquezas no modelo ARIMA é a previsão de *outliers*, principalmente quando não houver um histórico frequente de *outliers* nos dados. Na nossa aplicação notamos que o modelo não foi capaz de prever a queda do PIB que aconteceu em 2020, devido à pandemia da COVID-19 (pode-se ver na Figura 5.9 que a variação do PIB em 2020, claramente é um *outlier*). De um modo geral, o modelo é muito potente e capaz de oferecer boas previsões para dados em forma de séries temporais, quando não ocorrerem acontecimentos raros que podem afetar significativamente os *outcomes* futuros.

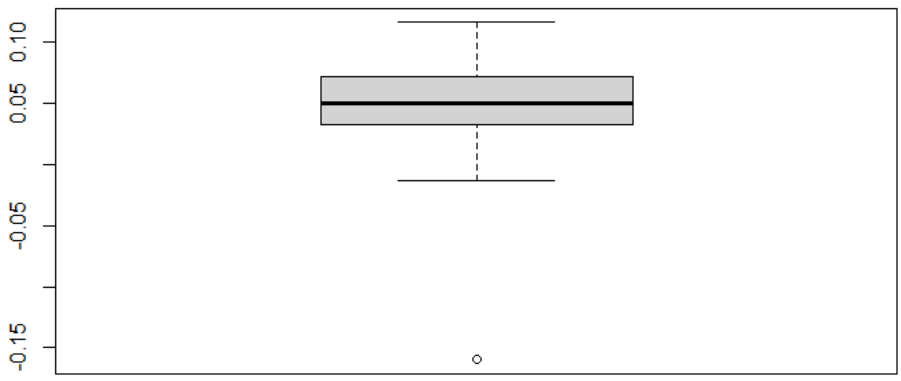

Figura 5.9: Boxplot do PIB real de Cabo Verde 1980 a 2020

Por definição, *outliers* ou observações atípicas, são observações incomuns que se afastam significativamente das observações normalmente observadas. A predição da ocorrência futura

de *outliers* é muito desafiadora, pois são valores que não se observam frequentemente [20]. Com o desenvolvimento da ciência computacional, existem agora alguns métodos aplicados para a previsão de ocorrência de *outiliers*. Modelos que usam o *Machine Learning*, têm sido aplicados com o objetivo de prever *outliers*, mas é desafiador também para esses modelos, pois normalmente não existem dados suficientes para treinar esses modelos de forma a identificarem os *ouliers*. Para mais informações sobre métodos de previsão que tratam de *outliers* consultar Reunanen [20] ou Gerogiadis [11].

Capítulo 6

Conclusão

Uma grande parte das observações de fenómenos económico-financeiros podem ser representados em forma de séries temporais. Essas séries, caracterizadas por pontos observados com igual espaçamento temporal podem ser analisadas e utilizadas para fazer previsões sobre valores futuros. Existem vários métodos usados na análise e previsão de séries temporais, sendo que a maioria usa valores passados da série para criar um modelo usado para representação e previsão da série.

De entre os métodos utilizados na análise de séries temporais, destaca-se o $ARIMA(p, d, q)$, que é a combinação dos modelos autoregressivos de ordem p ($AR(p)$) e dos modelos de médias móveis de ordem q ($MA(q)$) onde é aplicada a diferenciação de ordem d na série original para a transformar numa série estacionária. Esse método é muito utilizado na análise de séries económicas e financeiras.

Neste trabalho, o modelo $ARIMA(1, 2, 1)$ mostrou-se o melhor (entre os modelos $ARIMA(p, d, q)$) para a análise e previsão do Produto Interno Bruto (PIB) de Cabo Verde com dados de 1980 a 2018. A série ajustada obtida produziu valores muito próximas dos valores reais e conseguiu obter previsões muito próximos dos valores reais para o ano 2019. Entretanto, a previsão obtida para o ano 2020 ficou muito longe dos valores efetivos. O ano 2020 foi marcado pela queda da economia mundial devido à pandemia da COVID-19, fazendo com que o PIB da maioria dos países caíssem para valores pouco frequentes. Em 2019 a previsão feita pelo Fundo Monetário Internacional[14] era de que o PIB de

Cabo Verde cresceria em torno dos 5% em 2020, entretanto houve uma recessão de 14,8%, contrariando todas as previsões feitas, incluindo as produzidas pelo modelo $ARIMA$.

Um dos pontos fracos do modelo ARIMA e da maioria dos métodos de previsão é justamente a previsão de *outliers* quando não forem frequentemente observados nos dados históricos (como foi o caso do PIB de Cabo Verde em 2020).

Pela aplicação prática, verifica-se que o modelo $ARIMA(1, 2, 1)$ ajusta-se de forma muito eficientes a dados económico-financeiros, produzindo previsões robustos e podendo ser aplicado a outros dados financeiros da economia Cabo-Verdiana, obtendo assim previsões úteis na tomada de decisão, mas com a ressalva de que a mesma não é muito eficiente na previsão de *outliers* que afastam muito dos valores históricos.

Para mitigar esse problema há que investigar, procurando modelos que, por um lado traduzam fidedignamente a realidade das restantes observações, modelos robustos, mas que, por outro lado, sejam capazes de prever a ocorrência de tais observações, eventualmente, passando pela aplicação da inteligência artificial aos dados.

Neste trabalho foi possível constatar que o modelo ARIMA se ajusta de forma muito adequada aos dados económico-financeiros analisados, podendo ser aplicado nas previsões. A possibilidade de ser aplicado a outros dados financeiros da economia Cabo-Verdiana, obtendo assim previsões úteis na tomada de decisão, constitui uma das direções possíveis para trabalho futuro. A previsível ocorrência de futuras observações atípicas cujas causas são múltiplas e variadas é alvo de preocupação na comunidade académica e profissional. Motiva, por isso, o interesse e a procura de modelos mais apropriados.

Bibliografia

[1] Elena-Adriana Andrei, Elena Bugudui, and Others. 2011. Econometric modeling of GDP time series. *Theoretical and Applied Economics* 10, 10 (2011), 91.

[2] Andreea-Gabriela Baltac and others. 2015. Economic and financial analysis based on time series method. *International Journal of Academic Research in Accounting, Finance and Management Sciences* 5, 3 (2015), 77–82.

[3] The World Bank. 2021. Cabo verde aspectos gerais. Retrieved November 7, 2021 from https://www.worldbank.org/pt/country/caboverde/overview#1

[4] George E P Box, G M Jenkins, and G C Reinsel. 1994. Time series analysis: forecasting and control.

[5] Tim Callen. 2012. Gross domestic product: An economy's all. *International Monetary Fund: Washington, DC, USA* (2012).

[6] Chris Chatfield. 2000. *Time-series forecasting*. CRC press.

[7] Kleyton da Costa and Felipe Silva. 2020. Modelos de previsão de séries temporais aplicados ao produto interno bruto do brasil entre 1996 e 2019.

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

[9] Estatcamp. 4 - Modelos ARIMA - Séries Temporais | Portal Action. Retrieved April 6, 2021 from http://www.portalaction.com.br/series-temporais/modelos-arima

[10] Frank J Fabozzi, Sergio M Focardi, Svetlozar T Rachev, and Bala G Arshanapalli. 2014. *The basics of financial econometrics: Tools, concepts, and asset management applications*. John Wiley & Sons.

[11] Dimitrios Georgiadis, Maria Kontaki, Anastasios Gounaris, Apostolos N Papadopoulos, Kostas Tsichlas, and Yannis Manolopoulos. 2013. Continuous outlier detection in data streams: An extensible framework and state-of-the-art algorithms. In *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, 1061–1064.

[12] James Douglas Hamilton. 2020. *Time series analysis*. Princeton university press.

[13] Rob J Hyndman and George Athanasopoulos. 2018. *Forecasting: Principles and practice*. OTexts.

[14] Fundo Monetário Internacional. 2020. *Cabo Verde: Second Review Under the Policy Coordination Instrument and Request for Modification of Targets–Debt Sustainability Analysis*. IMF. Retrieved November 7, 2021 from https://www.imf.org/en/Search#q=Cape%20Verde%20PCI&sort=relevancy

[15] Zhenwei Li, Jing Han, and Yuping Song. 2020. On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning. *Journal of Forecasting* 39, 7 (2020), 1081–1097.

[16] Nazaré Mendes Lopes. 2008. Séries temporais e decisão dinâmica. *Sociedade Portuguesa de Estatística* (2008), 26–34.

[17] Pedro Alberto Morettin. 1981. *Modelos para previsão de séries temporais*. Instituto de Matemática Pura e Aplicada. Rio de Janeiro.

[18] Robert S Pindyck and Daniel L Rubinfeld. 2008. *Econometric models and economic forecasts*. Irwin, McGraw-Hill.

[19] R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved April 7, 2021 from https://www.R-project.org/

[20] Niko Reunanen, Tomi Räty, Juho, J Jokinen, Tyler Hoyt, and David Culler. 2020. Unsupervised online detection and prediction of outliers in streams of sensor data. *International Journal of Data Science and Analytics* 9, (2020), 285–314. DOI:https://doi.org/10.1007/s41060-019-00191-3

[21] RStudio Team. 2019. *RStudio: Integrated development environment for r*. RStudio, Inc., Boston, MA. Retrieved April 7, 2021 from http://www.rstudio.com/

[22] Robert H Shumway, David S Stoffer, and David S Stoffer. 2000. *Time series analysis and its applications*. Springer.

[23] James.H. Stock. 2015. Time series: Economic forecasting. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. 337–340.

[24] Tao Wang and Others. 2016. Forecast of economic growth by time series and scenario planning method—A case study of Shenzhen. *Modern Economy* 7, 02 (2016), 212.

[25] Jorge Manuel Nunes Xavier. 2016. Análise e previsão de séries temporais com modelos ARIMA e análise espectral singular. Master's thesis. Universidade Aberta.

Apêndice

Códigos Utilizados

PIB de Cabo Verde a preços correntes entre 2007 - 2020 (Figura 3.1)

```
library(ggplot2)
library(readxl)
pib <- read_excel("dados.xls")
ggplot(pib, aes(x = pib$Ano, y = pib$`Milhões de CVE`)) +
geom_line() +
scale_x_continuous(labels=as.character(pib$Ano), breaks= pib$Ano) +
labs(y= "Produto Interno Bruto (PIB)- em Milhões de escudos", x = "Ano")
```

Entrada de Turistas em Cabo Verde de 2000 a 2020 (Figura 3.2)

```
library(ggplot2)
library(readxl)
turi <- read_excel("dados.xlsx")
ggplot(turi, aes(x = turi$ano, y = turi$`entrada_de_turistas`)) +
geom_line(size = 1L, colour = "#CC6666") +
scale_x_continuous(labels=as.character(turi$ano),breaks= turi$ano) +
```

```
scale_y_continuous(labels=function(n){format(n, scientific = FALSE)}) +
labs(y= "Entrada de Turistas em Cabo Verde de 2000 a 2020", x = "Ano")
```

Movimento de Passageiros mensais 2006 a 2017 (Figura 3.3)

```
library(ggplot2)
library(scales)
library(lubridate)
library(readxl)
entrada<- read_excel("R_codes_examples/entrada_turistas.xlsx",
sheet = "mensal", col_types = c("date", "numeric"))
ggplot(entrada, aes(x =`mes_ano`, y = `movimento_de_passageiros`)) +
geom_line(colour = "#0c4c8a") +
scale_x_datetime(date_breaks = "6 months",date_labels = "%b-%Y",
limits = c(as.POSIXct("2006-08-01"), as.POSIXct("2017-12-01"))) +
scale_y_continuous(labels=function(n){format(n, scientific = FALSE)}) +
labs(y= "Movimento de Passageiros mensais 2006 a 2017 ", x = "Mês/Ano") +
xlab("") +
theme(axis.text.x=element_text(angle=60, hjust=1))
```

Evolução do PIB(em volume) de Cabo Verde de 1980-2020 (Figura 5.1).

```
library(readxl)
library(ggfortify)
require(ggfortify)
library(xts)
pib <- read_excel("dados.xls")
pib_1 =pib[ ,2]
pib_ts1 <- ts(pib_1, frequency = 1, start = c(1980))
```

```
autoplot(as.xts(pib_ts1), ts.colour = 'blue',
xlab = "Ano",ylab = "Montante em Milhões de CVE")
```

Análise descritiva dos dados (Tabela 5.1)

```
library(readxl)
suppressMessages(library(timeSeries))
library(fBasics)
pib <- read_excel("dados.xls", sheet = "PIB_em_volume", col_types = c("numeric"))
pib_1 =pib[ ,2]
pib_ts <- ts(pib_1, frequency = 1, start = c(1980,1), end = c(2018,1))
Resumo_pib = data.frame(round(basicStats(pib_ts), digits = 2))
colnames(Resumo_pib)<- NULL
knitr::kable(Resumo_pib, 'pipe',longtable = TRUE, booktabs = TRUE,
  caption = ' Análise descritiva dos dados.'
)
```

Histograma do PIB(em volume) de 1980-2018 (Figura 5.2)

Teste de Shapiro-Wilk para a normalidade

```
library(stats)
SH=shapiro.test(pib_ts)
```

Teste Augmented Dickey-Fuller (ADF), ACF e PACF para a estacionaridade (Figura 5.3)

```r
# Teste adf
adf.test(pib_ts)

# Teste ACF e PACF
acf2(pib_ts, main = "Funções ACF e PACF")
```

Serie log(PIB) aplicado a diferença de ordem 1 (Figura 5.4)

```r
dif_log_pib_ts <- diff(log(pib_ts))
autoplot(as.xts(dif_log_pib_ts), ts.colour = 'blue')
```

Identificação da ordem de diferenciação

```r
library(forecast)
ndiffs(pib_ts)
```

Identificação da modelo ARIMA

```r
# Escolha do modelo com base nos valores de AIC e BIC
library(astsa)
a = Arima(pib_ts, c(1,2,1), include.constant = TRUE, include.mean = TRUE)

# Escolha automática do modelo
library(forecast)
auto.arima(pib_ts, seasonal = FALSE)
```

Análise dos resíduos (Figura 5.5)

```
# Gráfico dos resíduos
plot(a$residuals, ylab= "Resíduos")

# Gráfico AIC e BIC dos resíduos
acf2(a$residuals, main = "ACF e PACF dos resíduos")

# Teste Ljung-Box para ruído branco
Box.test(a$residuals)
```

Previsão usando modelo ARIMA(1,2,1) (Figura 5.7)

```
# Gráfico dos valors efetivos vs valores obtidos
# pelo através do modelo ARIMA (1,2,1)
library(psych)
ts.plot(fitted(a), type = "l", xlab = "Tempo",
ylab ="PIB em Milhões de CVE", lwd =2.5, col ="blue")
lines(pib_ts, col="red", type = "l", lwd = 3)
legend("topleft", c("Série prevista","Série original"),
fill=c("blue","red"))

# Previsões de 2019 a 2029
library(forecast)
forecast(a)

# Representação gráficas das previsões
plot(forecast(a, h= 10), include = 80)
lines(pib_ts1, col=2)
```