Universidade do Minho
Escola de Ciências

Vítor Hugo Araújo da Silva

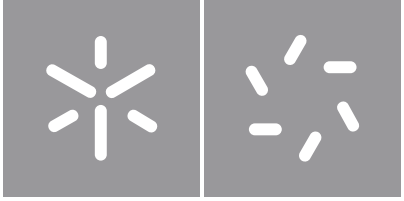
Modelação Estatística: um estudo na Gestão Empresarial Local

outubro de 2019

Modelação Estatística: um estudo na Gestão Empresarial Local

Vítor Hugo Araújo da Silva

UMinho | 2019



Universidade do Minho
Escola de Ciências

Vítor Hugo Araújo da Silva

Modelação Estatística: um estudo na Gestão Empresarial Local

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Arminda Manuela Andrade Pereira Gonçalves
e do
Mestre João Pedro de Oliveira Martins Castro

outubro de 2019

Agradecimentos

"A persistência é o caminho do êxito."
(Charles Chaplin)

À Professora Arminda Manuela, por todos os conhecimentos que me transmitiu desde o primeiro dia. Serão poucas as palavras para lhe agradecer tudo o que fez por mim.

Ao Dr. João Pedro Castro pela ajuda demonstrada ao longo da orientação deste estágio.

Palavras de agradecimento ao Dr. Daniel Pinto e ao Vítor Pinheiro, um enorme obrigado, pelo apoio incansável, por valorizarem e a apostarem meu trabalho e nas minhas capacidades enquanto uma mais valia na empresa.

Agradecimentos à Laura Jota por toda a preocupação e por toda a ajuda prestada nas diversas tarefas, não só na construção dos resultados pretendidos mas também na elaboração de toda a dissertação. Foi uma das melhores companhias deste estágio!

A todos os colegas da VITRUS que diretamente e indiretamente, colaboraram comigo durante o estágio, palavras de gratidão por toda a ajuda demonstrada.

À minha irmã e aos meus pais, por todo o esforço que fazem e fizeram, durante todo o meu percurso, para que nada me faltasse, por toda a compreensão e motivação!

Aos meus avós, por sempre se interessarem e quererem ajudar, sempre e a toda a hora!

Por último, mas não menos importante, agradecimentos à minha família, por todo o apoio incondicional demonstrado desde sempre, por apostarem sempre em mim e por nunca cruzarem os braços nas adversidades.

A todos aqueles que contribuem para o meu crescimento a nível pessoal e profissional e por incentivarem a arriscar!

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

Com o desenvolvimento económico, populacional e social em geral, a quantidade de resíduos, em particular os resíduos urbanos, está a aumentar de forma significativa. Sendo uma das problemáticas a nível nacional e mundial, é necessário adotar medidas de forma que essas quantidades sejam reduzidas e valorizadas. A VITRUS AMBIENTE, EM, S.A. é uma empresa pública que atua a vários níveis na gestão empresarial local, nomeadamente na Gestão de Resíduos Urbanos assegurando a recolha de resíduos no concelho de Guimarães.

Em 2016 foi implementado um projeto pioneiro, denominado de "Pay-As-You-Throw" (PAYT), no Centro Histórico da cidade de Guimarães, cuja entidade gestora é a VITRUS, sendo o Serviço de Higiene Urbana o responsável pela implementação das medidas necessárias para o sucesso deste projeto.

Este trabalho foca-se, especificamente, na Gestão de Resíduos com o objetivo de modelar e prever o comportamento da produção de Resíduos Urbanos nas áreas de atuação da empresa. Assim, são desenvolvidos modelos estatísticos num contexto de Modelos de Regressão Linear (numa abordagem de modelação simples e múltipla) e de análise de Séries Temporais para estimar e prever, nos períodos observados, a produção de resíduos nas zonas dos circuitos de recolha indiferenciada e na zona piloto de implementação do sistema PAYT.

O principal objetivo deste trabalho consiste em avaliar a influência de fatores que estejam relacionados com as quantidades de resíduos recolhidos nas zonas afetas ao Serviço de Higiene Urbana e, também, analisar a evolução das respetivas quantidades produzidas. Desta forma, numa primeira fase, são identificados, com recurso a Modelos de Regressão Linear, os fatores que influenciam, numa perspetiva empresarial, as quantidades de resíduos produzidas na zona piloto de implementação do sistema PAYT e possíveis tendências e padrões sazonais, para uma posterior melhoria das ações de gestão a implementar pela empresa. São também aplicados modelos de previsão em Séries Temporais para a estimação e a previsão da produção de resíduos num horizonte futuro de curto prazo (semanalmente) relativos à recolha indiferenciada em contentores de profundidade, em diversas freguesias do concelho, e na recolha indiferenciada e seletiva no Centro Histórico de Guimarães intramuros.

As metodologias utilizadas servem de apoio para a gestão e processo de tomada de decisões da empresa relativamente à Gestão de Resíduos Urbanos, com o intuito de melhorar os serviços prestados à população tendo sempre como fundamento a preservação do meio ambiente.

Palavras-chave: Gestão de Resíduos, Reciclagem, PAYT, Modelação, Regressão Linear, Previsão, Séries temporais.

Abstract

Due to the economic and social development in general and to population growth, the amount of waste, particularly municipal waste, has been significantly increasing in recent years. It is one of the major problems both at a national and global level, and action is urgently needed to ensure that waste is recovered and its volume reduced. VITRUS AMBIENTE, EM, S.A. is a public company that operates at various levels in local business management, namely in Urban Waste Management, and ensures the collection of waste in the municipality of Guimarães.

In 2016 a pioneer project called "Pay-As-You-Throw" (PAYT) was implemented in the Historic Center of the city of Guimarães. VITRUS is the managing entity and the Urban Hygiene Service is responsible for implementing the necessary measures to ensure the success of this project.

This work focuses specifically on Waste Management and aims at modeling and predicting the behavior of Urban Waste production within the company's areas of activity. Thus, statistical models are developed in the context of Linear Regression Models (in a single and multiple modeling approach) and Time Series analysis to estimate and predict, in the observed periods, the waste production in the areas of the undifferentiated collection circuits and in the PAYT system implementation pilot zone.

The main objective of this work is both to evaluate the influence of factors related to the amount of waste collected in the areas covered by the Urban Hygiene Service and to analyze the evolution of the respective quantities produced. Thus, in the first stage, we use Linear Regression Models to both identify the factors that, from a business perspective, do influence the amount of waste produced in the PAYT system implementation pilot zone and possible seasonal trends and patterns, in order to further improve management actions to be implemented by the company. Time series forecasting models are also applied both for the estimation and forecasting of short-term (weekly) future waste generation regarding undifferentiated waste collection in deep containers in various parishes of the municipality, and for undifferentiated and selective intramural waste collection in the Historic Center of Guimarães.

The methodologies used support the company's management and decision-making process regarding Urban Waste Management, aiming at improving the services provided to the population and having always as its cornerstone the preservation of the environment.

Keywords: Waste Management, Recycling, PAYT, Modeling, Linear Regression, Forecasting, Time Series.

Conteúdo

1 Introdução **1**

1.1 Descrição e caracterização da Instituição de Estágio 1

1.2 Serviço de Higiene Urbana (SHU) 5

1.3 Recolha de Resíduos Urbanos: o caso de Guimarães 6

1.4 Sitema PAYT: o caso de Guimarães 9

1.5 Definição do problema e objetivos 16

1.6 *Software* utilizado 17

1.7 Estrutura do documento 17

2 Enquadramento **19**

2.1 O sistema PAYT (*Pay-As-You-Throw*) 19

2.2 Aplicações 23

3 Modelos de Regressão Linear **25**

3.1 Regressão Linear Múltipla 26

3.1.1 Propriedades dos estimadores 30

3.1.2 Estimação de σ^2 31

3.1.3 Testes de hipóteses sobre os coeficientes de regressão 33

3.1.4 Validação de pressupostos e análise de resíduos 36

3.1.5 Qualidade do modelo e análise de R^2 39

3.1.6 Predição 40

3.2 Modelação da Sazonalidade 41

4 Séries Temporais **43**

4.1 Conceito de Série Temporal 43

4.2 Processos Estocásticos 46

4.2.1 Processos Estocásticos Estacionários 46

4.2.2 Processos Estocásticos não Estacionários 51

5 Métodos de Previsão em Séries Temporais **57**

5.1 A Metodologia Box-Jenkins 58

5.1.1 Modelos de Processos Estacionários 59

5.1.2 Modelos de Processos Não Estacionários 63

5.1.3 Etapas da Metodologia Box-Jenkins 66

5.1.4 Seleção de modelos 69

5.1.5 Previsão 71

5.2 Avaliação de Modelos de Previsão 72

5.2.1 Medidas de Avaliação 73

6 Análise Exploratória de Dados **75**

6.1 Circuitos de recolha indiferenciada 75

6.2 *Pay-as-you-throw*: o caso de Guimarães 81

7 Aplicação de Modelos de Regressão Linear **89**

7.1 Regressão Linear Simples 89

7.2 Regressão Linear Múltipla 91

7.2.1 Resíduos indiferenciados 92

7.2.2 Resíduos seletivos 93

7.2.3 Análise de resíduos 95

7.3 Modelos de Regressão para a sazonalidade 98

7.3.1 Resíduos indiferenciados 98

7.3.2 Resíduos seletivos 102

8 Aplicação de Métodos de Previsão **107**

8.1 Caso I: Recolha de contentores de profundidade 108

8.2 Caso II: Recolha de resíduos em área PAYT 113
8.3 Avaliação dos Modelos de Previsão 123

9 Conclusão **127**

9.1 Sugestões para trabalho futuro 128

A Circuitos de Recolha de Resíduos Urbanos **135**

B Regressão Linear Simples **137**

C Regressão Linear Múltipla **141**

C.1 Resíduos de papel/cartão 141
C.2 Resíduos de vidro 142
C.3 Resíduos de plástico 142
C.4 Modelos de Regressão para a sazonalidade 143

Lista de Figuras

1.1 Sede da VITRUS AMBIENTE, exemplo da reabilitação urbana, situada no centro da cidade de Guimarães (reproduzido de VITRUS AMBIENTE (2019b)) 2

1.2 Impacto de cada serviço na faturação da empresa (adaptado de VITRUS AMBIENTE (2019b)). 3

1.3 Organograma da empresa (reproduzido de VITRUS AMBIENTE (2018)). . 5

1.4 Mapa do concelho de Guimarães, com as respetivas freguesias e união de freguesias representadas (reproduzido de VITRUS AMBIENTE (2018)). . . 8

1.5 Ilustração relativa ao mapa da zona de implementação do projeto piloto PAYT, no Centro Histórico intramuros (reproduzido de VITRUS AMBIENTE (2019a)). 10

1.6 Contentores oferecidos aos utilizadores dos sistema PAYT na zona piloto, como incentivo à separação de resíduos (reproduzido de VITRUS AMBIENTE (2019a)). 11

1.7 Autocolantes utilizados nos sacos não autorizados (reproduzido de VITRUS AMBIENTE (2019a)). 14

1.8 Representação da zona piloto, de alargamento em 2019 e alargamento num futuro próximo do sistema PAYT na cidade de Guimarães (reproduzido de VITRUS AMBIENTE (2019a)). 16

3.1 Representação gráfica dos dados relativos à hereditariedade (Galton, 1889), com a respetiva reta de regressão. 25

4.1 Representação da simulação de um ruído branco e respetivas FAC e FACP empíricas. 51

5.1 Simulação de um processo autorregressivo e respetivas FAC e FACP empíricas. 60
5.2 Simulação de um processo de médias móveis e respetivas FAC e FACP empíricas. 61
5.3 Simulação de um processo autorregressivo e de médias móveis, ARMA$(2, 2)$ e respetivas FAC e FACP empíricas. 63
5.4 Simulação de um processo autorregressivo e de médias móveis integrado, ARIMA$(2, 1, 1)$ e respetivas FAC e FACP empíricas. 64
5.5 Simulação de um processo autorregressivo e de médias móveis integrado sazonal, SARIMA$(2, 1, 1)(1, 1, 1)_{12}$ e respetivas FAC e FACP empíricas. . . 66

6.1 Evolução das quantidades de resíduos indiferenciados, recolhidos por ano. . 77
6.2 Evolução das quantidades de resíduos indiferenciados por tipo de recolha. . 77
6.3 Diagramas em caixa de bigodes da produção de resíduos semanal dos circuitos (1 a 12) de recolha operados pela VITRUS. 79
6.4 Diagramas de dispersão da produção de resíduos semanal dos circuitos (1 a 12) de recolha operados pela VITRUS. 80
6.5 Representação gráfica da evolução da produção de resíduos. 83
6.6 Diagramas de caixa com bigodes relativos às quantidades de resíduos, produzidas mensalmente. 84
6.7 Evolução do número de sacos vendidos, por litragem, conforme o tipo de utilizador (esquerda: UD, direita: UND). 85
6.8 Gráficos relativos à evolução das compras efetuadas pelos utilizadores conforme a litragem do saco. 86
6.9 Evolução do número de deposições ilegais na zona piloto de implementação do sistema PAYT. 87

7.1 Histograma e *QQ plot* dos resíduos do modelo obtido para os resíduos indiferenciados. 96
7.2 Histograma e *QQ plot* dos resíduos do modelo obtido para os resíduos seletivos. 97

8.1 Série dos logaritmos das quantidades recolhidas em contentores de profundidade, após diferenciação de 1.ª ordem ($d = 1$), e respetivas FAC e FACP estimadas. 109

8.2 Série dos logaritmos das quantidades recolhidas em contentores de profundidade, após diferenciação de 1.ª ordem e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas. 110

8.3 Série dos resíduos para a série dos logaritmos das quantidades de resíduos indiferenciados em contentores, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas. 112

8.4 Previsões (no período de teste), pontuais e intervalares (90%), e estimativas pontuais (entre a 16.ª semana de 2016 e a 34.ª semana de 2019) obtidas através do modelo SARIMA, sobrepostas à série das quantidades de resíduos indiferenciados em contentores de profundidade. 113

8.5 Série dos logaritmos das quantidades recolhidas de resíduos seletivos, no CHG, sem aplicação de diferenciação ($d = 0$), e respetivas FAC e FACP estimadas. 114

8.6 Série dos resíduos das quantidades recolhidas de resíduos seletivos, no CHG, sem diferenciação aplicada e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas. 115

8.7 Série dos resíduos para a série dos logaritmos das quantidades de resíduos seletivos no CHG, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas. 117

8.8 Previsões (no período de teste), pontuais e intervalares (90%), e estimativas pontuais (entre a 9.ª semana de 2016 e a 38.ª semana de 2018) obtidas através do modelo SARIMA, sobrepostas à série das quantidades de resíduos seletivos no CHG. 118

8.9 Série dos logaritmos das quantidades recolhidas de resíduos indiferenciados, no CHG, após diferenciação de 1.ª ordem ($d = 1$), e respetivas FAC e FACP estimadas. 119

8.10 Série dos resíduos das quantidades recolhidas de resíduos seletivos, no CHG, após aplicação de uma diferenciação de 1.ª ordem ($d = 1$), e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas. 120
8.11 Série dos resíduos para a série dos logaritmos das quantidades de resíduos indiferenciados no CHG, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas. 122
8.12 Previsões (no período de teste), pontuais e intervalares (90%), e estimativas pontuais (entre a 9.ª semana de 2016 e a 38.ª semana de 2018) obtidas através do modelo SARIMA, sobrepostas à série das quantidades de resíduos indiferenciados no CHG. 123

Lista de Tabelas

1.1 Tipos de recolha efetuados nos circuitos de recolha indiferenciada. 9

1.2 Descrição dos estabelecimentos pertencentes às tipologias estabelecidas nos utilizadores não domésticos (UND). 11

1.3 Litragem dos sacos vendidos para os resíduos respetivos. 12

1.4 Preçário (em euros) estipulado para a diversidade de sacos adquiridos pelos utilizadores, conforme a litragem adquirida (L). 12

3.1 Tabela ANOVA. 36

4.1 Transformações usuais de Box-Cox. 53

5.1 Padrões teóricos das FAC e FACP dos modelos de previsão em séries temporais. 67

6.1 Evolução anual das quantidades de resíduos indiferenciados (em toneladas), por tipo de recolha. 76

6.2 Estatísticas descritivas dos circuitos de recolha indiferenciada no período observado. 78

6.3 Descrição das variáveis em estudo, relativamente ao sistema PAYT, implementado no Centro Histórico intramuros. 81

6.4 Estatísticas descritivas relativas às variáveis em estudo do circuito PAYT (mensais). 83

6.5 Evolução da produção mensal de resíduos no Centro Histórico intramuros, zona piloto do sistema PAYT. 84

6.6 Quantidades de sacos vendidas, anualmente, na globalidade, conforme a litragem de sacos. 85

6.7 Teste de correlação de Spearman para avaliar a associação entre as variáveis (n.º de sacos vendidos) e o número de deposições ilegais. 88

7.1 Regressão Linear Simples, tendo como variável resposta `SELETIVO` e `INDIFERENCIADO`, respetivamente. 90

7.2 Modelo de regressão linear múltipla para a produção de resíduos indiferenciados. 92

7.3 Modelo de regressão linear múltipla para a produção de resíduos seletivos. . 94

7.4 Valores obtidos após modelação do modelo sazonal final, dos resíduos indiferenciados. 99

7.5 Modelo de regressão linear múltipla para a produção de resíduos indiferenciados com combinação das variáveis sazonais. 100

7.6 Modelo de regressão linear múltipla para a produção de resíduos indiferenciados com combinação das variáveis sazonais e variáveis obtidas por regressão linear simples. 101

7.7 Valores obtidos após modelação do modelo sazonal final, dos resíduos seletivos. 103

7.8 Modelo de regressão linear múltipla para a produção de resíduos seletivos com combinação das variáveis sazonais. 104

7.9 Modelo de regressão linear múltipla para a produção de resíduos seletivos com combinação das variáveis sazonais e variáveis obtidas por regressão linear simples. 105

8.1 Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos das quantidades recolhidas em contentores de profundidade. 110

8.2 Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos das quantidades de resíduos indiferenciados em contentores de profundidade. 111

8.3 Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos das quantidades de resíduos indiferenciados em contentores de profundidade. 111
8.4 Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos das quantidades recolhidas em contentores de profundidade. 115
8.5 Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos das quantidades de resíduos seletivos no CHG. 116
8.6 Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos das quantidades de resíduos seletivos no CHG. 116
8.7 Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos das quantidades recolhidas de resíduos indiferenciados no CHG. 120
8.8 Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos das quantidades de resíduos indiferenciados no CHG. 121
8.9 Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos das quantidades de resíduos indiferenciados no CHG. 121
8.10 Medidas de avaliação calculadas para as séries estudadas, no período de treino e no período de teste respetivo, com base nos resultados obtidos na aplicação do método de previsão. 124
8.11 Tabela com os respetivos intervalos de previsão, valores previstos e valores reais para cada série em estudo, respetivamente. 125

A.1 Distribuição dos circuitos de recolha indiferenciada pelas freguesias. 136

B.1 Regressão Linear Simples, tendo como variável resposta `papel`. 137
B.2 Regressão Linear Simples, tendo como variável resposta `plastico`. 138
B.3 Regressão Linear Simples, tendo como variável resposta `vidro`. 139

C.1 Modelo de regressão linear múltipla para a produção de resíduos de papel/-cartão. 141

C.2 Modelo de regressão linear múltipla para a produção de resíduos de vidro. . 142
C.3 Modelo de regressão linear múltipla para a produção de resíduos de plástico. 142
C.4 Valores calculados a partir do modelo sazonal inicial, dos resíduos indiferenciados e seletivos, respetivamente. 143

Lista de abreviaturas

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

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

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

AR – *Autoregressive* (em português, Autorregressivo)

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

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

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

BLUE – *Best Linear Unbiased Estimators* (em português, Melhor Estimador Não Enviesado)

CHG – Centro Histórico de Guimarães

CMG – Câmara Municipal de Guimarães

DF – Dickey-Fuller

EAM – Erro Absoluto Médio

EAMN – Erro Absoluto Médio *Naïve*

EAMNS – Erro Absoluto Médio *Naïve* Sazonal

EEAM – Erro Escalado Absoluto Médio

EMQ – Estimadores de Mínimos Quadrados

EPA – *Environmental Protection Agency*

EPAM – Erro Percentual Absoluto Médio

EQM – Erro Quadrático Médio

ERSAR – Entidade Reguladora do Serviço de Águas e Resíduos

EUA – Estados Unidos da América

FAC – Função de Autocorrelação

FACP – Função de Autocorrelação Parcial

INE – Instituto Nacional de Estatística

KPSS – Kwiatkowski-Phillips-Schmidt-Shin

MA – Médias Móveis

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

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

MQE – Média dos Quadrados dos Resíduos

MQR – Média dos Quadrados da Regressão

MQT – Média dos Quadrados Total

NA – Não Aplicável

PAYT – *Pay-As-You-Throw*

PPP – Princípio Poluidor Pagador

REQM – Raiz do Erro Quadrático Médio

RI – Resíduos Indiferenciados

RS – Resíduos Seletivos

RSU – Resíduos Sólidos Urbanos

RU – Resíduos Urbanos

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

SHU – Serviço de Higiene Urbana

SQE – Soma dos Quadrados dos Resíduos

SQR – Soma dos Quadrados da Regressão

SQT – Soma dos Quadrados Total

UD – Utilizadores Domésticos

UND – Utilizadores Não-Domésticos

ZEDL – Zona de Estacionamento de Duração Limitada

Capítulo 1

Introdução

No âmbito curricular do Mestrado em Estatística foi realizado um estágio curricular na VITRUS AMBIENTE, EM, S.A. (de agora em diante VITRUS), situada na cidade de Guimarães, com início a 14 de janeiro e término a 13 de setembro de 2019. Com esta parceria e sob a orientação Professora Doutora Arminda Manuela Gonçalves e do Mestre João Pedro Castro, foi possível realizar o trabalho que serviu de mote para o presente documento. O trabalho foi elaborado no Serviço de Higiene Urbana (SHU) com a colaboração e auxílio de Laura Jota (Técnica Superior no SHU) e Vítor Pinheiro (Responsável do SHU).

Neste Capítulo será feita uma breve descrição da instituição de estágio, missão, visão e valores. É, também, apresentado o Serviço de Higiene Urbana onde foi elaborado o trabalho apresentado, assim como as noções preliminares da temática a abordar. Estas breves apresentações e descrições foram adaptadas e fundamentadas de documentação interna (VITRUS AMBIENTE (2018)) e do *website* (VITRUS AMBIENTE (2019b)) da instituição de estágio.

Serão mencionados ainda os objetivos definidos, o *software* utilizado e a respetiva descrição da organização do documento.

1.1 Descrição e caracterização da Instituição de Estágio

A VITRUS AMBIENTE, EM, S.A., atua no âmbito geográfico do concelho de Guimarães, cuja cobertura territorial tem vindo a registar um crescimento significativo e sustentável, desde a sua criação e início de atividade. Inicialmente, a VITRUS localizava-se na Praça Colónia de Sacramento, perto do Complexo Desportivo do Vitória *Sport Club*. Em 2013, em acordo com a Câmara Municipal de Guimarães, assumiu a responsabilidade de reabilitar uma casa abandonada situada no coração da cidade de Guimarães, mais precisamente na Avenida Cónego Gaspar Estaço, número 606, na freguesia de Oliveira do Castelo (Figura 1.1). A reabilitação foi conseguida devido à dedicação dos colaboradores da empresa ao longo de 9 meses. Em junho do mesmo ano, a VITRUS passou a estar

mais próxima da comunidade a quem serve diariamente. A VITRUS emprega cerca de 143 colaboradores distribuídos pelos diferentes serviços.

Figura 1.1: Sede da VITRUS AMBIENTE, exemplo da reabilitação urbana, situada no centro da cidade de Guimarães (reproduzido de VITRUS AMBIENTE (2019b))

Constituída por escritura pública em 8 de setembro de 2010 e com início da sua atividade no dia 1 de outubro do mesmo ano, a VITRUS assume, por delegação de competências do Município de Guimarães, a gestão de serviços de interesse geral nas seguintes áreas:

- **Serviços de Recolha de Resíduos Sólidos Urbanos e Higiene Urbana**
  A recolha e transporte ao destino final de Resíduos Sólidos Urbanos (RSU) foi a primeira atividade da VITRUS. É uma das tarefas de grande responsabilidade e de fundamental importância para garantir as melhores condições de higiene urbana e pública no concelho de Guimarães. Para que a satisfação do cliente final atinja níveis elevados, são desenvolvidos todos os esforços por forma a garantir um serviço capaz e eficiente.

- **Seviços de Limpeza**
  O serviço de Higiene e Limpeza Urbana, com início no ano de 2011, passa fundamentalmente pela limpeza de edifícios públicos que estão sob a responsabilidade ou ocupados pela Câmara Municipal de Guimarães. Alguns dos espaços submetidos a estes serviços são: Mercado Municipal, Central de Camionagem, Feira Retalhista, CPCJ, Loja Ponto JÁ, Espaço Saúde Jovem, Pontos de Turismo, Casa da Memória, Centro Ciência Viva, entre outros.

- **Estacionamento Público Urbano**
  VITRUS assumiu, por delegação de competências do Município de Guimarães, a partir do dia 1 de janeiro de 2012, a gestão e fiscalização das zonas de estacionamento de duração limitada, vulgarmente conhecidas como parcómetros. No âmbito dessas competências delegadas, há uma constante preocupação em fazer cumprir o Regula-

mento das Zonas de Estacionamento de Duração Limitada, aprovado pela Câmara e Assembleia Municipal, onde constam as zonas intervencionadas, os horários de funcionamento, a classificação de veículos, as taxas a aplicar, as isenções previstas, as contraordenações aplicáveis, entre outros; A VITRUS também efetua a gestão de cinco parques públicos, nomeadamente o Parque Condessa Mumadona, o Parque Central do Estádio D. Afonso Henriques, o Parque do Mercado Municipal, o Parque do Centro Cultural Vila Flor e, por último, o Parque da Plataforma das Artes e Criatividade.

A empresa, a nível de faturação, tem um grande impacto a nível de serviços de resíduos urbanos, estacionamento público e serviços de limpeza (Figura 1.2).

Figura 1.2: Impacto de cada serviço na faturação da empresa (adaptado de VITRUS AMBIENTE (2019b)).

A VITRUS tem como missão a recolha e transporte dos resíduos sólidos urbanos contentorizados, a limpeza e higiene dos edifícios públicos ou onde estão instalados serviços municipais e a gestão do estacionamento urbano estabelecendo o serviço público de qualidade como referência e tomando como base orientadora os princípios básicos da gestão: elevada eficácia e eficiência. Um dos seus objetivos enquanto empresa é ser uma organização de referência local e nacional, nas respetivas áreas de atuação, onde o reconhecimento pela capacidade de adaptação às constantes mudanças é uma das diversas ambições. Como tal, a VITRUS assenta toda a sua ação nos seguintes valores:

- **Veracidade**: As ações e decisões serem sempre exatas e verdadeiras;
- **Inovação**: Com a intenção de criar valor onde atua e naquilo que faz;
- **Transparência**: Ligação clara e inequívoca na relação com os *stakeholders*;
- **Responsabilidade**: Assumir as funções e implicações das respetivas ações;
- **Utilidade**: Sentir que as funções que desempenha são importantes e imprescindíveis;
- **Sustentabilidade**: Priorizar as necessidades em função dos recursos disponíveis.

No exercício das suas funções, a VITRUS pretende obter elevados níveis de eficácia e eficiência, melhorando continuamente o seu desempenho de forma a alcançar a satisfação plena do seu acionista, clientes, fornecedores, colaboradores e demais interessados. Assente nos valores, a VITRUS quer assumir e demonstrar o importante papel que a atividade desenvolvida representa no desenvolvimento local e na melhor qualidade de vida que proporciona aos munícipes do concelho vimaranense. Em todas as ações a levar à prática, a VITRUS pauta sempre o seu trabalho com um grande sentido de responsabilidade, exigência e rigor, tornando assim o seu crescimento sustentado e alicerçado em premissas sólidas e perenes.

Todos os serviços estão mutuamente interligados, sendo que todos são importantes. O trabalho em equipa permite conseguir muito mais do que aquilo que alguma se conseguiria de forma isolada. Para ser colocado em prática, no espírito de equipa é necessário conhecer não só as suas, mas também as responsabilidades dos seus colegas. Seguidamente são apresentadas, as funções de cada departamento:

- **Administrativo e Financeiro**
  Controlar todo o circuito e movimentação das receitas e despesas, garantir a fiabilidade dos registos e procedimentos contabilísticos, acompanhar a gestão económico-financeira, organizar todos os procedimentos de aquisição ou alienação de bens e serviços, gerir os *stocks* existentes na empresa, gerir a carteira de seguros da empresa, fazer a gestão patrimonial, entre outros;

- **Recursos Humanos**
  Controlar a assiduidade e pontualidade dos colaboradores, elaborar mapas de assiduidade, pontualidade, trabalho extraordinário e das férias, criar e atualizar o banco de horas, planear e executar os processos de recrutamento e seleção, elaborar o plano de formação anual, planear e executar a avaliação de desempenho e a compensação, zelar pela satisfação e motivação de todos os colaboradores, bem como por um elevado espírito de equipa, entre outros;

- **Higiene Urbana**
  A nível da Gestão de Resíduos Urbanos é efetuado o planeamento e coordenação das atividades de recolha, transporte e destino final dos resíduos, efetuando para o efeito a monitorização de circuitos de recolha e implementação dos melhores critérios para os mesmos. Desenvolver um plano efetivo de higienização dos equipamentos afetos à recolha de resíduos, entre outros. Na vertente da Limpeza Pública são construídos mapas de reposição de materiais consumíveis nas respetivas instalações para proceder aos serviços de limpeza de ruas, praças, mercados, recintos desportivos, espaços, instalações, entre outros;

- **Estacionamento Público**
  Este serviço, em particular na vertente da Gestão de Parques Públicos garante que os parques são espaços seguros, asseados e com bom ambiente, garantir a boa circulação

dos parques, controlar as entradas e saídas de clientes, realizar a manutenção dos equipamentos, receber os pagamentos dos avençados e dos rotativos, zelar por um atendimento de excelência aos clientes, entre outros. Na vertente das Zonas de Estacionamento de Duração Limitada, gere e fiscaliza as zonas de estacionamento de duração limitada, cumprindo o regulamento das Zonas de Estacionamento de Duração Limitada (ZEDL), realizar a manutenção dos equipamentos, entre outros.

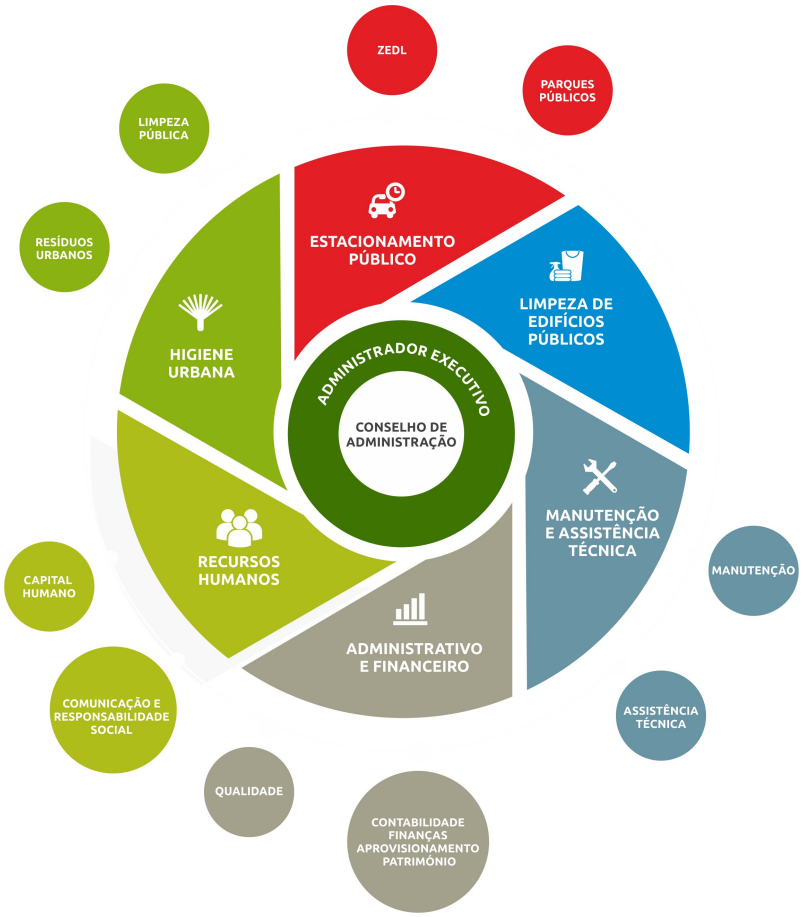


Figura 1.3: Organograma da empresa (reproduzido de VITRUS AMBIENTE (2018)).

1.2 Serviço de Higiene Urbana (SHU)

O Serviço de Higiene Urbana (SHU) iniciou a sua atividade em 2010, com a operacionalização de um sistema de recolha de resíduos urbanos (RU) indiferenciados acondicionados em 202 contentores de profundidade. Hoje em dia, o serviço é responsável por 430 conten-

tores localizados por todo o concelho, operando em cerca de 44 freguesias no município de Guimarães. Pouco tempo após o início da sua atividade, foi adjudicado ao SHU o serviço de limpeza pública, nomeadamente a limpeza de caminhos pedestres e da pista de cicloturismo.

Em 2016 o SHU, ficou também responsável pela recolha de RU no Centro Histórico de Guimarães (CHG), circuito este onde é efetuada a recolha seletiva de resíduos e é designado de Circuito PAYT (*Pay-As-You-Throw*).

Tendo ainda em conta o organograma da VITRUS (Figura 1.3), o SHU encontra-se segmentado em dois setores principais: o serviço de recolha de resíduos e a limpeza pública.

Para a concretização destes serviços existe um quadro de colaboradores alargado, com as seguintes distinções:

- Equipas que trabalham diretamente nos serviços, motoristas e assistentes operacionais: 71 colaboradores diretos;
- Cargos de coordenação e chefia: 2 colaboradores diretos;
- Cargos administrativos e técnicos: 1 colaborador direto.

O SHU possui ainda um estaleiro de apoio às atividades realizadas diariamente, que serve de garagem e armazém e, ainda, agrega a oficina e o parque de lavagem. No serviço de oficina é realizada toda a reparação e manutenção dos equipamentos, ferramentas e materiais e é no parque de lavagem que ocorre a higienização dos veículos, equipamentos e materiais. No estaleiro também existe o serviço de lavandaria onde é efetuada a higienização e manutenção dos uniformes onde, porém, este serviço está afeto a todos os serviços da VITRUS.

Mais se acrescenta que o SHU possui uma frota de viaturas pesadas e ligeiras, de carga, viaturas de apoio à recolha e outras auxiliares. Não tendo sido realizada nenhuma aquisição no ano de 2018, o número que compõe a frota sob alçada da VITRUS é de 17 viaturas.

1.3 Recolha de Resíduos Urbanos: o caso de Guimarães

O modelo de gestão de resíduos em Portugal é da responsabilidade dos municípios, ao abrigo do Decreto-Lei n.º73/2011, de 17 de junho. A gestão de resíduos é uma área onde urge a criação de políticas de redução de custos e de diminuição de produção de resíduos nas áreas de implementação.

O concelho de Guimarães, situado no Norte de Portugal, na sub-região do Vale do Ave, apresenta uma área total de 241 km^2 distribuída por um total de 48 freguesias (Figura 1.4). De acordo com o Instituto Nacional de Estatística (INE), a população deste concelho é composta por 158 124 habitante sendo que, também, é composto por 66 790 alojamentos.

No município de Guimarães, a entidade responsável pelo serviço em alta é a Resinorte, sendo o sistema em baixa tutelado pelo município de Guimarães. A Resinorte providencia a recolha seletiva em toda a área do concelho, com exceção do CHG, o tratamento e a valorização dos resíduos recolhidos. Esta entidade gere, ainda, os ecocentros existentes no concelho uma vez que é nestes locais que qualquer munícipe, assim como o SHU, pode depositar materiais em fim de vida para posterior valorização.

A prestação de serviços, por parte do SHU, consiste essencialmente na recolha e transporte de RU integrando o sistema em baixa referido, anteriormente. Nos locais onde opera como entidade gestora, o SHU presta também um serviço de recolha a pedido, que consiste na recolha de resíduos que não podem ser recolhidos nos circuitos normais, devido à sua forma, volume ou caraterísticas. À recolha deste tipo de resíduos (volumosos, resíduos verdes, etc.) denomina-se de recolha de monstros.

A VITRUS opera, por delegação de competências do Município de Guimarães, na recolha de resíduos urbanos em 38 freguesias do concelho onde as recolhas são distribuídas por doze circuitos de recolha de resíduos indiferenciados e um circuito de recolha indiferenciada e recolha seletiva, respetivamente. O serviço afeto à recolha de RU é o Serviço de Higiene Urbana que, à data, apresenta cerca de 60 colaboradores afetos a este tipo de serviço. As recolhas nos circuitos de recolha de RU indiferenciados são efetuadas em contentores de profundidade ou por recolha porta a porta com deposição de saco perdido. Na Tabela A.1 (Apêndice A) é possível analisar, de forma detalhada, as freguesias afetas a cada circuito de recolha de RU indiferenciado.

O circuito onde é efetuada a recolha seletiva de resíduos é designado de Circuito PAYT. Em 2016, este sistema, implementado no Centro Histórico intramuros, é caraterizado pela recolha de resíduos porta a porta em saco apropriado e fornecido pela VITRUS.

Circuitos de recolha indiferenciada

Para Martinho & Gonçalves (1999), a escolha do sistema de deposição a adotar é condicionada por diversos fatores, desde o clima, os aspetos geográficos, o volume e tipo de resíduos a recolher, o tipo de habitação e urbanização, a frequência e celeridade da recolha, a distância e o tipo de tratamento, valorização ou eliminação que se pretende para os resíduos, os hábitos, as atitudes e as caraterísticas dos produtores de resíduos, o tipo de recipientes e veículos a utilizar e os recursos financeiros e humanos disponíveis. Neste caso, realça-se que por diversos fatores é impossível efetuar a recolha de resíduos com um só sistema o que leva à criação de uma variedade de sistemas de recolha que permitam que cada comunidade utilize a que mais se adequa às suas necessidades tendo em conta os fatores a reter.

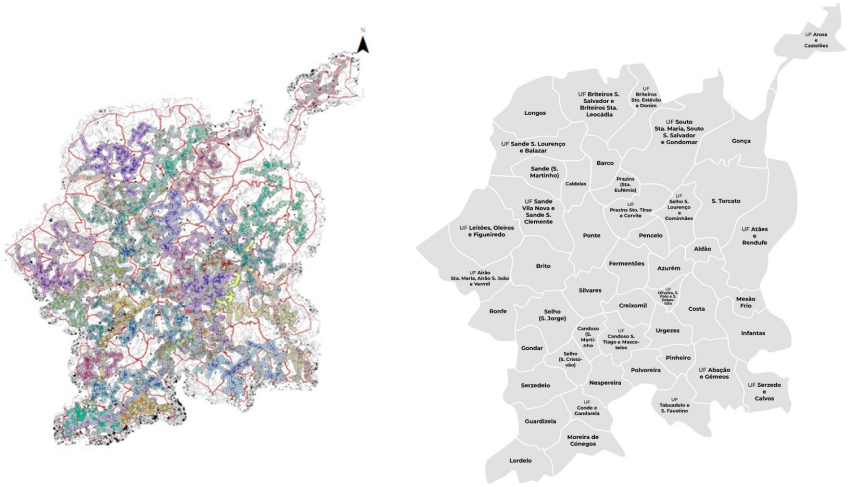


Figura 1.4: Mapa do concelho de Guimarães, com as respetivas freguesias e união de freguesias representadas (reproduzido de VITRUS AMBIENTE (2018)).

Tipos de recolha

O sistema de recolha porta a porta através de saco perdido traduz-se na deposição na via pública, por parte do produtor, do saco, em frente à sua habitação, para posterior recolha por parte do SHU. Neste caso, os produtores, só podem depositar os sacos na via pública a determinadas horas e, devidamente acondicionados, de forma a permitir uma higienização da via pública. Neste estudo, serão considerados/analisados quatro circuitos onde é efetuada a recolha porta a porta: Circuitos 9, 10, 11 e 12.

A recolha em contentores de profundidade é realizada em contentores, colocados em locais estratégicos, próximos das habitações onde a população se desloca para depositar os seus resíduos. Usualmente, este método é utilizado na impossibilidade de efetuar uma recolha porta a porta. Neste estudo, serão estudados cinco circuitos cuja metodologia adotada é, exclusivamente, a recolha em contentores de profundidade: Circuitos 1, 2, 4, 5 e 6.

A recolha mista consiste na combinação das duas anteriores, uma vez que o circuito afeta diversas freguesias cujas necessidades divergem e é então necessária uma alternância na metodologia a adotar para a recolha dos resíduos. Neste caso são três circuitos com esse tipo de recolha: Circuitos 3, 7 e 8.

A Tabela 1.1 resume os tipos de recolha, anteriormente discriminados, nos circuitos de recolha indiferenciada em estudo.

Tabela 1.1: Tipos de recolha efetuados nos circuitos de recolha indiferenciada.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Contentores de profundidade	×	×		×	×	×						
Porta a porta									×	×	×	×
Mista			×				×	×				

1.4 Sitema PAYT: o caso de Guimarães

Em abril de 2016 foi implementado, em Guimarães, um projeto piloto denominado de projeto PAYT (*Pay-as-you-Throw*). Este projeto consiste na implementação de um tarifário calculado de forma proporcional à quantidade de resíduos produzida que segundo Freitas (2013) pode ser uma medida eficaz para os objetivos da política de gestão de resíduos, na medida em que constitui um claro incentivo, por via financeira, para promover a separação na origem e aumentar as taxas de Resíduos Sólidos Urbanos.

A zona de implementação do sistema está adjacente à zona intramuros classificada como património mundial da humanidade, inserida na União de Freguesias de Oliveira, São Paio e São Sebastião, que constitui a principal e mais central freguesia do Município de Guimarães. A área de implementação, ilustrada na Figura 1.5, tem uma área igual a 0,170 km^2 e é caraterizada por uma elevada densidade de construção predominada pela habitação e pelo comércio. Inicialmente a área era constituída por um total de 32 arruamentos mas, após a entrada em vigor da tarifa PAYT e por se situarem numa zona fronteira que divide o Centro Histórico intra do extra muro, alguns utilizadores requisitaram a adesão ao sistema tarifário PAYT e o número de arruamentos foi aumentado para 34. Esta inclusão foi considerada após pedido especial de 10 lojistas do centro comercial do Toural, com morada fiscal no Largo do Toural e 2 utilizadores domésticos do Largo Navarro de Andrade.

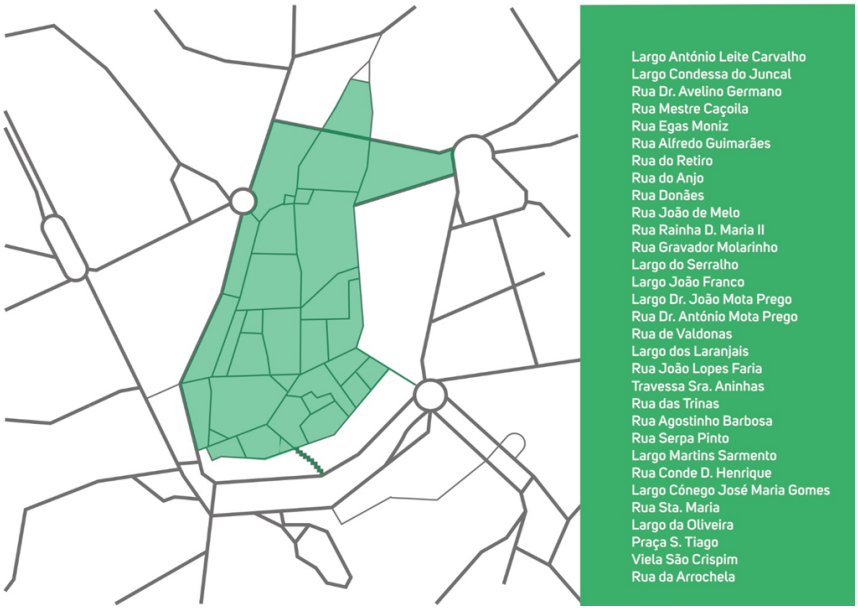


Figura 1.5: Ilustração relativa ao mapa da zona de implementação do projeto piloto PAYT, no Centro Histórico intramuros (reproduzido de VITRUS AMBIENTE (2019a)).

A implementação deste sistema imprimiu uma nova dinâmica da gestão de resíduos, no qual o utilizador é tratado de acordo com a sua efetiva produção de resíduos, vendo isso repercutido na tarifa a pagar.

Este sistema tarifário tem como princípio o conceito poluidor-pagador e, por isso, penaliza a produção de resíduos indiferenciados e incentiva a redução, reutilização e o aumento da separação da fração reciclável na origem. Neste contexto, a operacionalização do sistema PAYT alterou o tarifário associado ao sistema de recolha de resíduos e vai de encontro às novas Diretivas Europeias para a gestão de resíduos, o que contribui para uma atuação "mais verde" e ambientalmente sustentável do munícipe conduzindo para um sistema de gestão de resíduos urbanos mais sustentável e próximo das metas muito específicas e ambiciosas na área dos resíduos.

A implementação deste projeto compreendeu uma monitorização contínua e uma relação de proximidade com a população, com resoluções céleres das necessidades apontadas por todos os utilizadores. Desde a sua implementação, o projeto PAYT foi também auxiliado por fortes ações/campanhas de sensibilização e educação ambiental e pela agilização do processo de controlo e fiscalização para posterior recolha do número de deposições ilegais na zona de intervenção. A distinção dos utilizadores é efetuada por Utilizadores Domésticos (UD) e Utilizadores Não-Domésticos (UND). Na Tabela 1.2 são apresentados os estabelecimentos que compõem as diferentes tipologias constantes dos UND.

Tabela 1.2: Descrição dos estabelecimentos pertencentes às tipologias estabelecidas nos utilizadores não domésticos (UND).

Tipologia	**Estabelecimentos**
Tipologia A	Café, bar e padaria
Tipologia B	Restaurantes
Tipologia C	Lojas de venda a retalho (roupa, sapatos e outos artigos) e prestador de serviços
Tipologia D	Hotel, hostel, alojamento local
Tipologia E	Instituições sociais, institutos e associações locais.

A VITRUS, nomeadamente o SHU é responsável pela deposição, recolha e transporte dos resíduos urbanos produzidos no Centro Histórico. A Resinorte é a empresa responsável pelo tratamento e valorização deste resíduo. Atendendo à morfologia urbana, ao tipo de produtores e características da área, a VITRUS dispõe de diferentes soluções para a deposição e recolha de resíduos no Centro Histórico:

- Recolha seletiva e indiferenciada porta a porta em todas as residências/mistas e junto a todas as entidades (restaurantes, bares, hotéis, mercados, comércio, serviços, etc):
  - Disponibilização gratuita de contentores de pequena capacidade (25 ou 45 litros (Figura 1.6));
  - Disponibilização gratuita de sacos próprios para deposição dos resíduos seletivos (sacos de 50 e 100 litros);
  - Sacos PAYT para venda com diferentes litragens (15 litros, 30 litros, 50 litros e 100 litros);
- Recolha pontual de resíduos a pedido: recolha efetuada mediante pedido prévio do munícipe, de caráter ocasional e realizada em local e data acordada.

Figura 1.6: Contentores oferecidos aos utilizadores dos sistema PAYT na zona piloto, como incentivo à separação de resíduos (reproduzido de VITRUS AMBIENTE (2019a)).

Para deposição de resíduos no CHG, atendendo à obrigatoriedade das normas do sistema PAYT e a todas as necessidades dos utilizadores, a VITRUS dispõe de diferentes soluções para deposição dos resíduos, por exemplo sacos para deposição de resíduos com diferentes volumetrias (Tabela 1.3).

Tabela 1.3: Litragem dos sacos vendidos para os resíduos respetivos.

Litragem	**Indiferenciado**	**Papel/ Cartão**	**Plástico**	**Vidro**
15 litros	×			
30 litros	×	×	×	×
50 litros	×	×	×	×
100 litros	×	×	×	

O sistema tarifário PAYT em Guimarães foi elaborado segundo as recomendações da Entidade Reguladora do Serviço de Águas e Resíduos (ERSAR), que sugere uma tarifa fixa e uma variável, de forma a repercutirem os custos por todos os utilizadores. A tarifa fixa é aplicada baseada nos custos fixos da operação e pela disponibilidade do serviço e a tarifa variável assenta na produção de resíduos.

Desta forma, é aplicado aos utilizadores do CHG uma tarifa de disponibilidade (componente fixa) mais tarifa variável. A tarifa de disponibilidade continuou a ser faturada juntamente com a fatura da água e a cobrança da tarifa variável passou a taxar o volume de resíduos indiferenciados produzidos segundo um sistema de sacos pré-pago.

Os residentes e comerciantes passaram a estar obrigados a adquirirem sacos para os Resíduos Indiferenciados (RI) e apenas os RI que são colocados nestes sacos são recolhidos normalmente pelos colaboradores. A utilização de outro saco que não o autorizado para deposição desta fração de resíduos é recolhido após ser iniciado o processo de fiscalização. Os sacos têm diferentes capacidades, entre 15 e 100 litros, e o preço do saco corresponde à porção dos custos de transporte e tratamento envolvido na eliminação desse resíduo. A Tabela 1.4 apresenta os preços em vigor dos diferentes sacos, colocados ao dispor, para posterior compra dos utilizadores com substituição da tarifa variável.

Tabela 1.4: Preçário (em euros) estipulado para a diversidade de sacos adquiridos pelos utilizadores, conforme a litragem adquirida (L).

Tipo de utilizador		**Preço por capacidade - sacos PAYT**			
		15 L	**30 L**	**50 L**	**100 L**
Utilizador Doméstico	$L \leq 240$	0,173	0,345	0,575	1,150
	$240 < L < 720$	0,174	0,348	0,580	1,160
	$720 \leq L < 1200$	0,177	0,354	0,590	1,180
	$L \geq 1200$	0,182	0,360	0,600	1,200
Utilizador Não Doméstico		0,174	0,348	0,580	1,160

A nível do controlo e fiscalização, segundo a legislação em vigor, com base no artigo 69.º, n.º 2, do Regulamento do Serviço de Gestão de Resíduos Urbanos, publicado no Diário da

República, 2.ª série, n.º 52, de 15 de março de 2016, sob o Edital n.º 248/2016, compete à VITRUS a fiscalização e a instrução dos processos de contraordenação cabendo à Entidade Titular, a Câmara Municipal de Guimarães (CMG), o processamento e a aplicação das coimas.

O serviço de fiscalização ambiental atua junto da população numa atitude preventiva e, em casos de reincidência, de forma coerciva, no sentido de fazer cumprir o Regulamento Municipal do Serviço de Gestão de Resíduos Urbanos. Em caso de reincidência e após advertência verbal, é enviado um ofício personalizado ao infrator identificado a informar que se encontra a infringir e a solicitar que proceda à regularização da situação.

As deposições ilegais consistem no acondicionamento de resíduos em sacos não autorizados, dentro do CHG ou fora da área de controlo. Também se considera como deposição ilegal a incorreta separação dos resíduos e a deposição, mesmo sendo em saco autorizado, fora do horário estipulado, especificamente, a equipa afeta a esta função cumpre o seguinte procedimento:

1. Observar e registar a situação anómala;
2. Identificação do autor da deposição ilegal;
3. Deslocação à morada do infrator;
4. Ação pedagógica e advertência verbal: informar e avisar o prevaricador do ato ilícito praticado e as devidas coimas previstas;
5. Em caso de reincidência, elaboração de ofício;
6. Envio de ofício ao infrator identificado;
7. Acompanhamento e verificação da alteração comportamental.

Na sequência da implementação do sistema PAYT, vários autores defendem que as deposições ilegais e a migração de resíduos são a principal problemática e merecem especial atenção. Desde a implementação do sistema PAYT no CHG, verificam-se alguns comportamentos desviantes praticados pelos utilizadores domésticos e não-domésticos, sendo as mais relevantesa deposição ilegal em sacos não autorizados e a deposição ilegal fora do Centro Histórico.

Assim, para este tipo de comportamentos, inicialmente é afixado um autocolante de "Saco Não Autorizado". Esta medida revelou-se eficaz, na medida que explicava o motivo da não recolha do saco. Foram registadas situações em que, posteriormente à não recolha de resíduos indiferenciados e respetiva colocação do autocolante, o saco não conforme era colocado num saco PAYT para ser recolhido, ou seja, a mensagem era passada. Paralelamente procedia-se ao registo mensal dos sacos não autorizados no momento da recolha.

Desta forma, para estes comportamentos considerados desviantes, foram desenvolvidos outros mecanismos, com o objetivo de aferir o máximo de informação, nomeadamente os

locais, a parte do dia e os dias da semana onde se verificam. Com isto, em janeiro de 2017, criou-se e distribuiu-se pelas equipas de recolha três autocolantes distintos (Figura 1.7) e folhas de registo para anotar os autocolantes usados diariamente, possibilitando um registo a montante mais fiável.

Figura 1.7: Autocolantes utilizados nos sacos não autorizados (reproduzido de VITRUS AMBIENTE (2019a)).

Com todos estes dados recolhidos e analisados é possível conhecer de forma detalhada o comportamento dos prevaricadores, para desta forma atuar, nomeadamente com o policiamento estratégico da Polícia Municipal e com rondas mais intensas.

Tendo em consideração as problemáticas identificadas e com base num plano estruturado, é um dos focos a melhoria gradual do projeto PAYT, através de um trabalho contínuo para garantir a total eficácia e universalidade do serviço. Assim, ao longo do tempo, são aplicadas várias estratégias e mecanismos para identificar, fiscalizar e controlar os comportamentos PAYT.

Rondas diárias na área intramuros do CHG: têm como principal intuito acompanhar de forma muito próxima os comportamentos dos utilizadores PAYT. Para além de possibilitar um acompanhamento diário das ocorrências do CHG, designadamente, identificação de prevaricadores e dados sobre as deposições ilegais.

***Empowerment* do cidadão**: Esta metodologia consiste num conjunto de técnicas sociopedagógicas que colocam em prática dois instrumentos fundamentais, os mecanismos de "escuta" da população e os mecanismos de "participação":

- O mecanismo de "escuta"refere-se às técnicas que são necessárias para ouvir a população, observar o seu comportamento e obter dados sobre a eficácia de funcionamento do projeto;
- O mecanismo de "participação"refere-se às técnicas que são necessárias para pôr a população a participar ativamente no próprio projeto, sobretudo no que diz respeito à sua adaptação à realidade social em que este se deseja implantar.

Esta estratégia tem-se revelado bastante profícua e traduziu-se na elaboração de inquéritos aos diferentes utilizadores, com o objetivo de percecionar a opinião dos moradores e comerciantes do CHG em relação à implementação do sistema. Também foram efetuadas ações de sensibilização individuais com explicações sobre o funcionamento do sistema , a realização de *workshops* para promover a redução, reutilização e reciclagem de resíduos e a distribuição de *newsletters* de forma a criar uma linha direta de comunicação entre o projeto e os seus utilizadores.

Ações corretivas: Para os utilizadores não cumpridores identificados, em primeira instância dirige-se de imediato ao infrator para advertir e perceber o contexto do comportamento observado. Esta ação é transversal ao longo do tempo e é colocada em prática a partir do momento, em que se identifica um utilizador infrator.

Policiamente estratégico: No âmbito da parceria estabelecida entre a VITRUS e a Polícia Municipal são desenvolvidas várias ações de policiamento estratégico nos locais considerados mais problemáticos. Estas ações têm como principal objetivo o visível controlo e fiscalização por parte da figura de autoridade e acessoriamente a possível identificação de infratores em flagrante delito.

Relativamente à recolha de sacos não autorizados, no ano de 2016, a recolha de sacos não autorizados, de uma forma geral, era feita no período máximo de 48 horas desde a sua deposição. Ao longo do mês de fevereiro de 2017, colocaram-se em prática algumas experiências em relação à recolha dos sacos não autorizados. Inicialmente defendia-se que quanto mais imediata for a recolha pior é a reação do autor da infração, compreendendo que o comportamento ilegal tinha exatamente o mesmo tratamento do que o comportamento legal, ou seja, era igualmente recolhido quer estivesse a cumprir ou não.

Desta forma, as entidades competentes decidiram transformar o saco de resíduos indiferenciados num objeto incomodativo para a população, não recolhendo durante alguns dias. Com a implementação desta medida alguns utilizadores queixaram-se e denunciavam os autores dos sacos não autorizados, outros recolhiam o seu saco não autorizado de forma discreta, outros ainda contribuíam para aumentar o monte de sacos não autorizados, de forma dolosa ou negligente.

Posto isto, foi alterada, de forma radical, a estratégia de recolha no mesmo *timing* do envio dos primeiros ofícios e do policiamento estratégico. No sentido de transparecer ainda mais a mudança comportamental verificada. Tendo por base a conclusão retirada de que comportamento negativo gera comportamento negativo, da mesma forma, comportamento positivo gera comportamento positivo, e para tal os sacos não autorizados passaram a ser recolhidos de forma mais célere, tornando o local visivelmente mais educado.

Por último, em relação à atribuição de um reforço positivo e tendo como princípio advertir os não cumpridores e parabenizar os cumpridores, foi desenvolvida a iniciativa de atribuir um dístico aos UND e um certificado aos UD acompanhado de um vale com a

oferta do número de sacos correspondente a 10% das compras em 2016.

Pelo facto de se terem alcançado todas as metas estabelecidas e pelo sucesso do projeto ao longo destes anos, a CMG em parceria com a VITRUS, pretende alargar o projeto PAYT a uma nova zona da cidade, nomeadamente às ruas adjacentes do CHG intramuros.

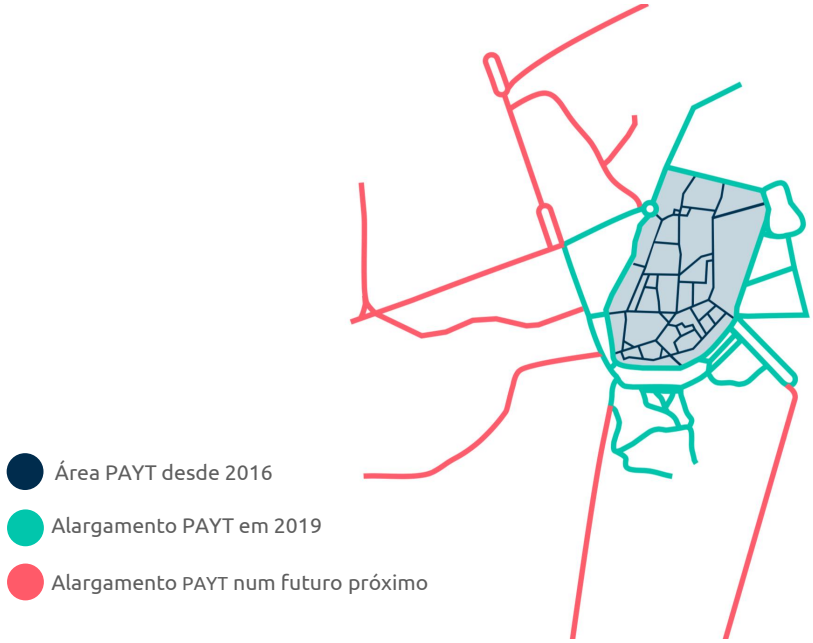


Figura 1.8: Representação da zona piloto, de alargamento em 2019 e alargamento num futuro próximo do sistema PAYT na cidade de Guimarães (reproduzido de VITRUS AMBIENTE (2019a)).

Este alargamento teve início em janeiro de 2019, ainda numa fase preliminar para posterior implementação da tarifa e demais metodologias, também, aplicadas na zona piloto de implementação. A Figura 1.8 ilustra as ruas abrangidas pelo alargamento PAYT na zona piloto (CHG intramuros), zona de implementação em 2019 e a zona de implementação num futuro próximo, respetivamente.

1.5 Definição do problema e objetivos

Segundo Freitas (2013), numa época em que as políticas nacionais e europeias obrigam à otimização dos recursos, recuperação de custos e a incutir comportamentos mais sustentáveis nos cidadãos ao nível da política de resíduos, é fundamental criar meios de planeamento e de gestão mais eficazes. Desta forma, os objetivos deste estudo são:

– efetuar uma análise preliminar dos dados fornecidos pela entidade gestora de forma a adquirir os conceitos e demais conhecimentos em relação às variáveis em estudo, analisar os seus comportamentos e identificar possíveis relações entre si. Numa primeira etapa, foi realizada uma análise exploratória da informação recolhida no período observado para os diversos circuitos de recolha de resíduos afetos ao SHU da VITRUS;

– aplicar conceitos de Inferência Estatística de forma a avaliar determinados fenómenos ocorridos no período de tempo observado;

– formular modelos de Regressão Linear com o objetivo de identificar covariáveis estatisticamente significativas na explicação das quantidades de resíduos indiferenciados e seletivos, respetivamente, nas zonas de atuação do SHU;

– estabelecer modelos de Séries Temporais de forma a prever as quantidades de resíduos recolhidos e/ou produzidos nas áreas de atuação do SHU, nomeadamente nos circuitos de recolha indiferenciada em contentores de profundidade e, também, no circuito de recolha porta a porta no CHG onde é efetuada a recolha indiferenciada e seletiva.

1.6 *Software* utilizado

No decorrer do estudo, foram utilizados três *softwares*:

– *Microsoft Office Excel* como ferramente de suporte para inserção de dados e respetiva criação das bases de dados utilizadas;

– `R` como ferramenta para tratamento dos dados fornecidos e respetiva aplicação das metodologias apresentadas;

– *Software* de gestão Primavera BSS: para extrair bases de dados relativas às vendas efetuadas para posterior análise.

1.7 Estrutura do documento

A dissertação está dividida em 9 Capítulos, que vão desarticulando os diversos momentos deste estudo aplicado ao Setor Empresarial Local.

No Capítulo 1 é descrita uma breve introdução ao tema em análise. É apresentada, também, uma referência à empresa e ao respetivo serviço nela articulado, na qual foi realizado o presente estudo.

No Capítulo 2 é elaborada uma breve revisão da literatura, isto é, um enquadramento à temática da Gestão de Resíduos Urbanos. São apresentados exemplos de aplicação da temática principal do estudo e, também de aplicação da Regressão Linear e dos Métodos de Previsão em Séries Temporais.

Nos Capítulos 3, 4 e 5 são descritos os conteúdos teóricos relacionados com a Regressão Linear, Séries Temporais e os Métodos de Previsão em Séries Temporais. São abordados os conceitos fundamentais das metodologias aplicadas e as medidas de avaliação para avaliar os modelos em estudo.

A análise exploratória dos dados em estudo é apresentada no Capítulo 6. Neste Capítulo são analisadas as bases de dados correspondentes aos circuitos de recolha indiferenciada e do sistema PAYT implementado no CHG.

A aplicação da Regressão Linear ao tema em estudo é apresentada no Capítulo 7. Numa fase inicial são formulados modelos de Regressão Linear Simples de forma a inferir sobre quais variáveis que poderão ter um maior poder explicativo sobre as duas variáveis resposta de interesse, correspondentes às quantidades de resíduos indiferenciados e seletivos na zona de implementação do sistema PAYT em Guimarães. Numa fase seguinte, tendo em conta o princípio da parcimónia, são formulados modelos de Regressão Linear Múltipla aplicando o método de seleção regressiva.

Numa última fase, são formulados modelos de Regressão para incorporar a componente da sazonalidade das variáveis com recurso a variáveis indicatrizes. São também formulados modelos de Regressão Linear Múltipla resultantes da combinação das variáveis indicatrizes com as variáveis selecionadas nos modelos de Regressão Linear Simples e de Regressão Linear Múltipla, respetivamente, após aplicação do método regressivo de seleção de variáveis.

O Capítulo 8 apresenta a aplicação dos Métodos de Previsão em Séries Temporais, nomeadamente à série temporal correspondente à recolha de resíduos em contentores de profundidade e às séries temporais dos resíduos indiferenciados e seletivos produzidos no CHG, respetivamente. São também calculadas as medidas de avaliação dos modelos e respetivas taxas de cobertura com o objetivo de determinar o desempenho da metodologia utilizada.

As principais conclusões do trabalho desenvolvido e sobre os resultados obtidos são descritas no Capítulo 9, assim como algumas sugestões de investigação para trabalho futuro.

Capítulo 2

Enquadramento

2.1 O sistema PAYT (*Pay-As-You-Throw*)

Atualmente, os Resíduos Sólidos Urbanos (RSU) são um dos principais problemas ambientais a nível mundial, onde é notório o crescimento na sua produção. Apesar da separação destes materiais, por parte da população, verifica-se um crescimento significativo, mas ainda insuficiente para que esta temática deixe de ser um problema. Em Portugal, as tarifas cobradas aos cidadãos para recolha e tratamento de resíduos não estão diretamente relacionadas com a sua produção, sendo esta uma das principais causas para as estratégias adotadas para a redução da produção de RSU não obterem os efeitos desejados. Desta forma, é cada vez mais importante definir metodologias que permitam atingir as metas que foram definidas no Plano Estratégico para os Resíduos Sólidos Urbanos (PERSU).

Uma das formas eficazes com o intuito de reduzir a produção de resíduos, passa pela aplicação de incentivos ou penalizações financeiras, recorrendo à implementação de um novo sistema que permita o cálculo de uma tarifa mais equitativa, com base nas quantidades de resíduos produzidos. Este sistema já existe e denomina-se *Pay-As-You-Throw* (PAYT), ou seja, "paga o que produzes".

Para Batllevell & Hanf (2008), um sistema PAYT assenta em duas diretrizes de planeamento: o princípio poluidor pagador (PPP) e o conceito da responsabilidade partilhada. De acordo com o PPP, os cidadãos devem pagar os custos da sua parte de responsabilidade no que toca à produção de resíduos. Os sistemas PAYT são aplicados sob forma de um incentivo financeiro, já que com este sistema o cidadão apenas paga a porção de resíduos indiferenciados que produz, enquanto a deposição seletiva não entra para o cálculo da tarifa.

Esta forma de cálculo pode-se traduzir num incentivo, mas também pode servir como penalização já que quanto maior for a quantidade de resíduos produzida, mais o cidadão terá de pagar. Com a aplicação destes sistemas espera-se promover a separação na origem e aumentar as taxas de recolha seletiva. Vários países já têm implementados sistemas PAYT. Como existem vários modelos deste sistema, a sua implementação pode adotar

várias formas. Estes modelos dependem da forma como a identificação do produtor de resíduos é feita, bem como da medição dos mesmos resíduos.

Bilitewski (2008) afirma que os sistemas PAYT permitem uma reinvenção dos Serviços de Gestão de Resíduos Urbanos, visto que irão distinguir todos os seus utilizadores e que cada um deles irá pagar pelo que realmente produz. Com isto, os utilizadores que efetuarem reciclagem ou reutilizarem os seus resíduos terão um custo menor com os seus RI, sendo assim possível criar um sistema de faturação transparente, ao contrário do sistema aplicado atualmente.

Batllevell & Hanf (2008), Bilitewski (2008), Karagiannidis (2008) e demais autores concluem que existe uma enorme diversidade de sistemas PAYT e que cada um deles pode ser aplicado consoante as necessidades dos cidadãos. Contudo, convergem na opinião que não existe uma resposta única na implementação destes sistemas. Acrescentam ainda que os sistemas PAYT têm que ser flexíveis e moldáveis às necessidades dos utilizadores. Existem, na grande maioria, dois tipos distintos de sistemas PAYT, por volume e por peso, desta forma o tarifário vai ser diferente e necessitam de um período mínimo para estudo para posterior implementação do sistema. Schindler et al. (2012) e Canterbury & Newill (2003) referem que a tarifa baseada no volume dos resíduos pode ser calculada de duas formas:

- Através da capacidade dos contentores em que a tarifa é aplicada por cada vez que o utilizador vai ao contentor na via pública;

- Através de sacos com tara perdida, que incluem uma tarifa de resíduos.

Nos sistemas por volume existem várias opções que podem ser influenciadas pela tipologia de habitações, rapidez de deposição e os custos de implementação e manutenção dos equipamentos. Nos contentores de proximidade, o volume é dado pela existência de uma tômbola, de um volume fixo e com acesso controlado. Desta forma, é possível identificar que utilizadores vão a estes equipamentos e aplicar uma tarifa com base no número de deposições que o utilizador irá fazer.

Nos sistemas de contentorização, os utilizadores escolhem o número e o volume dos contentores para depositar os seus resíduos. Os cidadãos serão taxados conforme o número de recolhas que pretendem e os volumes dos contentores adquiridos inicialmente. Estes sistemas podem ser equiparados aos sistemas porta a porta. No caso dos sacos com tara perdida, os utilizadores adquirem previamente os sacos às entidades gestoras do sistema, que já incorporam o valor da tarifa por saco adquirido. É um método muito aplicado na Europa, porque é de fácil implementação. Segundo o trabalho de Skumatz & Green (2002) é possível concluir que todos os sistemas PAYT apresentam pontos fortes, que devem ser examinados de forma a que estes sistemas se tornem sustentáveis e pontos fracos que devem ser estudados e minimizados.

Casos de estudo

Estados Unidos, EUA

De acordo com Skumatz (2008) os EUA foram os pioneiros no desenvolvimento dos sistemas PAYT. Desde 1980 que têm sido implementados estes sistemas, passando de 100 comunidades para 1000 registadas no início dos anos 90 e, em 2001, verificou-se um posterior acréscimo para 5200 comunidades. Atualmente estão referenciadas 7100 comunidades que implementaram este sistema tarifário, representando 25% do total da população dos EUA. Este crescimento resultou de um grande apoio legislativo onde muitos estados tiveram de alterar os seus regulamentos e políticas, como por exemplo o estado do Minnesota que tem este sistema implementado em 100% das suas comunidades. Já o estado de Washington obriga à implementação destes sistemas em comunidades que estejam certificadas ambientalmente. Resumidamente, nos municípios com maiores dimensões predominam os esquemas de pagamento por recolha de contentor, sem subscrição enquanto que nos municípios mais pequenos são mais usados os esquemas de pagamento por saco ou por contentor com sistemas de identificação dos produtores. Alguns projetos, infra descritos, são casos onde se verifica que, apesar de alguma resistência por parte dos utilizadores, foi possível implementar este sistema com a criação de boas campanhas de sensibilização e sistemas de tarifários justos. Canterbury & Newill (2003) aproveitando o trabalho desenvolvido pela *Environmental Protection Agency* (EPA), elaboraram uma lista onde é possível verificar os maiores casos de sucesso no país:

- **Vancouver, Washington**: No sistema PAYT, implementado em 1990, a recolha do segundo contentor era 84% mais cara do que o primeiro contentor. Com isto, os utilizadores deixaram de utilizar dois contentores onde, depois de algum tempo de implementação, foi criado um serviço de minicontentor mais económico que recebeu uma grande adesão. O resultado desta implementação traduziu-se na redução de resíduos e num aumento da produção de resíduos seletivos na ordem dos 50%, em 1995;

- **Mount Vernon, Iowa**: Este município, em 1991, implementou um sistema de etiquetas, que tinham o custo de 1,75 u.m., que eram colocadas em contentores com uma limitação de 114 litros ou 18 quilos. Para outras tipologias de resíduos, i.e. resíduos volumosos, eram utilizados outros sistemas. Para além da compra dos selos, era cobrado a cada utilizador uma tarifa fixa mensal de 7 u.m. Como consequência, reduziu-se em cerca de 40% os resíduos enviados para aterro;

- **Falmouth, Maine**: Em 1992 foi implementado um tarifário que tinha por base a compra de sacos que, posteriormente, seriam utilizados para a deposição de resíduos indiferenciados. Estes tinham o custo de 0,91 u.m. para sacos de 125 litros e 0,64 u.m. para sacos de 75 litros. Com este tarifário, as taxas de reciclagem aumentaram para além dos 50% e a deposição de RU decresceu cerca de 35%;

- **South Kingstown, Rhode**: Neste município, em 1994, foi implementado um sistema de etiquetas que eram adquiridas pelos cidadãos a 1 u.m., com a finalidade de as colocar nos sacos de deposição com um limite de 125 litros por saco. Os resultados traduziram-se numa taxa de reciclagem de 40%;

- **Fort Collins, Colorado**: Em 1995, foi implementado um sistema tarifário baseado no volume de resíduos recolhidos onde, a deposição de resíduos recicláveis era gratuita. A reciclagem aumentou para os 79% em residências unifamiliares comparativamente com os 53% do ano anterior.

Europa

Os sistemas PAYT já são bastante usados no Norte e Centro da Europa, nomeadamente na Suíça, Áustria, Alemanha, Itália, Dinamarca e Holanda. Utilizando como referência o estudo *Association of Cities & Regions for Recycling and for sustainable Resource Management*, elaborado por Dohogne (2016), são apresentados de seguida os casos com maior relevância na implementação de sistemas PAYT na Europa:

- **Interza, Bélgica**: Em 2004, foi implementado em Interza, na Bélgica, com uma população de 82 425 habitantes e 33 235 habitações, um sistema de sacos pré-pagos e preço por volume dos resíduos recicláveis. Esta metodologia carateriza-se pelo pagamento de sacos de resíduos indiferenciados com um valor muito superior ao estipulado para os sacos destinados aos resíduos seletivos. Este sistema permitiu uma redução de 25% nos RU ou equiparáveis;

- **Maastricht, Holanda**: No Município de Maastricht, na Holanda, com uma população abrangida de 122 481 habitantes e um número de habitações igual a 67 281, inaugurou-se em 2001 o projeto com um sistema de sacos pré-pagos com um preço estipulado para sacos de 50 litros e a utilização de contentores para deposição de resíduos seletivos a título gratuito. Verificou-se um aumento significativo nos resíduos seletivos e mais de 50% de separação nos resíduos biodegradáveis;

- **Umeå, Suécia**: Para uma população de 119 613 habitantes e 55 943 habitações, em 1996, foi delineado e implementado um sistema de frequência, volume e peso que se caraterizava pela variação de um preço para uma recolha bimensal de contentor de 4 m^3 até aos contentores de 8 m^3. As tarifas são aplicadas a condomínios ou conjunto de habitações térreas. Não é aplicada qualquer taxação para resíduos seletivos. Verificou-se uma diminuição de 23% dos RU e aumento de 25% na reciclagem nos primeiros dois anos de implementação. Comparando 1996 e 2014, houve uma diminuição de 44% dos RU e um aumento de 360% na separação de resíduos;

- **Zollernalbkeries/Zollernalbdistrict, Alemanha**: Na Alemanha, no município de Zollernallbkeries, iniciou-se um projeto PAYT em 1998 para uma população igual a 184 611 habitantes e 80 123 habitações. Baseado num sistema por frequência e,

posteriormente, em 2001, num sistema por peso, é cobrada uma taxa por unidade de medida (kg) e isenção de pagamento tratando-se de resíduos seletivos. Verificou-se então uma diminuição dos RU no ano de implementação e aumento dos materiais recicláveis recolhidos nos ecopontos;

- **Município de Treviso, Itália**: No ano de 2014, no município italiano de Tresivo, foi aplicado um sistema baseado na frequência de recolha dos contentores de RU, para uma população de 83 652 habitantes e de 41 951 habitações. É aplicada uma tarifa para contentores de RU de 30 litros e isenção no caso de resíduos seletivos. Verificou-se uma diminuição de 80% na produção de RU entre 2012 e 3014. A taxa de reciclagem aumentou ligeiramente;

- **Município de Besançon, França**: Em 2012, no município de Besançon, em França, foi adotado um sistema baseado no volume, peso e frequência. Para uma população de 176 339 habitantes e 84 873 habitações, a tarifa varia conforme a localização do utilizador e a tipologia de contentor, isto é, a tarifa consiste no somatório de um valor fixo pela recolha quinzenal com o valor variável consoante o peso de resíduos depositados e um valor fixo respeitante à taxa de recolha. De realçar que para os resíduos seletivos não há qualquer custo associado. Os RU tiveram uma redução de 23% e os recicláveis um aumento de 17%;

- **Município de Innsbruck, Áustria**: Para 127 944 habitantes distribuídos em 60 234 habitações, em 1995, foi aplicado um sistema baseado no volume, onde o utilizador paga um determinado valor por litro nos RU ou, então, os utilizadores podem comprar sacos de 60 litros por um determinado valor fixado pela entidade gestora. Neste caso, os RU reduziram 13% e os recicláveis aumentaram 38%.

Por todo o planeta, existem comunidades onde os tarifários PAYT se encontram implementados ou estão em fase de implementação. Para Reichenbach (2008), os últimos 20 anos foram fundamentais para o desenvolvimento técnico da implementação de soluções em sistemas PAYT, onde começaram a ser agrupadas condições de melhoramento que permitem a diminuição na produção de resíduos e o aumento da recolha seletiva, que levou a um acréscimo no número de países europeus e mundiais a adotar este sistema inovador na área da Gestão de Resíduos Urbanos.

2.2 Aplicações

A nível da Gestão de Resíduos Urbanos, vários autores aplicam metodologias estatísticas com vista à tomada de decisão. No estudo de Rimaitytė et al. (2012) são selecionados métodos de previsão para prever a produção de Resíduos Sólidos Urbanos. Os dados deste estudo dizem respeito à cidade de Kaunas, na Lituânia. As previsões relativas à produção de RSU foram baseadas na atividade económica da cidade, por modelos de regressão e

por modelos de séries temporais. Relativamente aos modelos de séries temporais foram utilizados os modelos SARIMA (Modelo Autorregressivo Integrado de Médias Móveis Sazonal), modelos de alisamento exponencial e a combinação dos dois referidos. Estes métodos foram aplicados a dados semanais, i.e., à produção semanal de resíduos, e a previsão foi feita para um horizonte de um ano. As conclusões retiradas deste estudo basearam-se na precisão das previsões obtidas para cada uma das metodologias e para a combinação das duas, respetivamente. Concluiu-se que a combinação de um modelo SARIMA com o modelo de alisamento exponencial apresenta uma alta precisão devido ao valor do seu erro percentual absoluto médio (MAPE). A combinação destas metodologias incorporou a influência de valores aleatórios e uma tendência crescente. Os autores defendem que este foi um modelo útil na medida que se obtiveram boas previsões nos valores semanais da produção de resíduos.

Já Song & He (2014) propõem modelos de séries temporais de forma a efetuar previsões para a produção diária de RSU na cidade de Seattle no estado americano de Washington. Os dados para modelação completavam cerca de 1001 observações das quais 901 para treino e 100 para teste. Neste estudo, foram aplicadas três metodologias, para posterior comparação. De forma a comparar a eficácia das três metodologias, os autores recorreram à REQM (Raiz do Erro Quadrático Médio) e ao MAPE, concluindo que a mais eficaz foi o modelo SARIMA de sazonalidade diária.

Em Navarro-Esbrí (2002) são apresentados dois métodos de previsão, um método baseado no modelo SARIMA, segundo a metodologia de Box & Jenkins. No outro método é utilizada a análise de sistemas não lineares, onde se assume que o sistema de produção de RSU é um sistema dinâmico discreto e extrai as taxas de produção de resíduos para uma posterior modelação. As técnicas utilizadas permitem a previsão para um horizonte temporal pretendido. Nesse estudo os dados para modelação dizem respeito às cidades de Thessaloniki (Grécia), Valencia (Espanha) e Castellón (Espanha), que foram analisados e modelados, conforme o planeamento e a posterior recolha de resíduos. De forma a avaliar as previsões dos dois métodos utilizados, os autores recorrem ao REQM e ao MAPE. Os modelos SARIMA foram os mais adequados uma vez que a técnica com recurso a sistemas não lineares, de uma forma geral, obtém um pior ajustamento. Assim, os modelos SARIMA apresentam um melhor ajustamento para este tipo de modelação.

Capítulo 3

Modelos de Regressão Linear

A análise de regressão consiste numa técnica estatística utilizada para analisar o comportamento de uma variável de interesse, designada de variável resposta ou variável dependente, como função de outras covariáveis, designadas de variáveis explicativas, variáveis independentes ou covariáveis. Esta técnica tem como fundamento a descrição de relações entre variáveis e a estimação ou previsão de valores da variável de interesse para valores, por vezes, não observados das covariáveis em estudo. O termo "regressão" remonta a Galton (1889), que o empregou pela primeira vez num estudo que relacionava a altura entre pais e filhos (Figura 3.1). Nesse estudo, Galton concluiu que embora existisse uma tendência para pais altos terem filhos altos e pais baixos terem filhos baixos, os filhos de pais excecionalmente altos (baixos) não eram tão altos (baixos) como os seus pais.

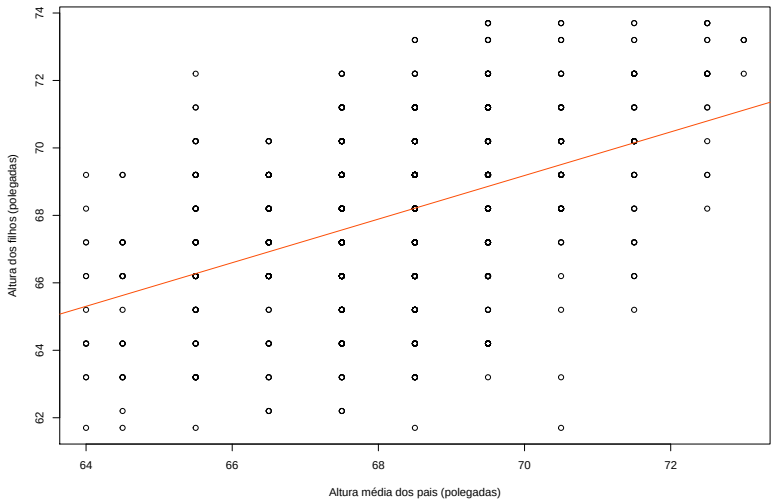


Figura 3.1: Representação gráfica dos dados relativos à hereditariedade (Galton, 1889), com a respetiva reta de regressão.

Foi desta forma, que o cientista, primo de Charles Darwin, descobriu que a altura os filhos tendia para a altura média da população. Mais tarde, a teoria de Galton foi confir-

mada por um dos seus discípulos, Karl Pearson, que denominou o fenómeno descoberto por "regressão para a média".

Desde essa época até hoje, muitos estudos foram realizados, muitas descobertas e adaptações foram feitas, mas por questões históricas, o termo "regressão" permaneceu.

Hoje em dia o termo regressão é comummente utilizado quando surge a necessidade de estudar uma relação funcional entre uma ou mais variáveis covariáveis e a variável resposta. Esta relação é representada por um modelo estocástico, isto é, por uma equação que associa a variável dependente ou resposta com a(s) variável(eis) independente(s).

A uma equação de regressão que contenha apenas um preditor chama-se equação de regressão simples ou univariada. A uma equação que contenha mais do que um preditor dá-se o nome de equação de regressão múltipla.

Independentemente do modelo ser simples ou múltiplo, pode ainda ser linear (equação da reta ou do plano) ou não linear (equação exponencial, logarítmica, etc.).

Esta técnica é utilizada nas diversas áreas científicas nas quais se pode encontrar aplicações da mesma, desde a Agricultura, Medicina, Biologia, Economia, Sociologia, Psicologia, Engenharia e demais áreas.

Os conteúdos apresentados nas diversas secções deste Capítulo têm como principal suporte os contributos de Sen & Srivastava (2012) e Fahrmeir & Kneib (2013).

3.1 Regressão Linear Múltipla

Considere-se um modelo de regressão que contenha p covariáveis e cuja função de regressão seja linear, ou seja, um modelo da forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ip} + \epsilon_i, i = 1, \ldots, n \tag{3.1}$$

onde y_i é a variável resposta do i-ésimo elemento da amostra, $x_{ij}, j = 1, \ldots, p$, são os correspondentes valores (fixos) das covariáveis, $\beta_0, \beta_1, \ldots, \beta_p$ são os parâmetros desconhecidos e ϵ_i é o erro aleatório associado ao elemento i da amostra.

O modelo 3.1 é denominado por modelo de regressão linear múltipla e pode ser visto como a soma de duas componentes, uma determinística dada por $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \sum_{j=0}^{p} \beta_j x_{ij}$ com $x_{i0} = 1, \forall i = 1, \ldots, n$ e outra aleatória dada por ϵ_i. A componente determinística, mesmo dependendo de parâmetros desconhecidos, é considerada fixa, enquanto que a componente aleatória admite uma distribuição de probabilidade que usualmente se supõe ser Normal, sendo esta uma condição que permite a elaboração de testes de hipóteses e obtenção de intervalos de confiança.

Recorrendo à notação matricial é possível reescrever o modelo de regressão linear múltipla definido na equação 3.1 como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

onde,

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{3.3}$$

Na forma matricial, representa-se por $\boldsymbol{Y}$ o vetor $n \times 1$ das observações da variável resposta, por $\boldsymbol{X}$ a matriz de planeamento $n \times (p+1)$ constituída pelas observações das covariáveis, por $\boldsymbol{\beta}$ o vetor $(p+1) \times 1$ dos coeficientes de regressão e por $\boldsymbol{\epsilon}$ o correspondente vetor $n \times 1$ dos termos do erro.

Segundo Fahrmeir & Kneib (2013), o modelo da equação 3.1 pressupõe, para além da existência de uma relação matemática linear entre as variáveis, da verificação de um conjunto de condições, denominadas "Condições de Gauss Markov":

1. Os erros são variáveis aleatórias de valor médio nulo, isto é, $E(\boldsymbol{\epsilon}) = \mathbf{0}_{n\times 1}$, fazendo com que $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$;

2. Os erros são variáveis aleatórias não correlacionadas de variância constante, isto é, $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}) = \sigma^2 \boldsymbol{I}_{n\times n}$, o que implica que $Cov(\boldsymbol{Y}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\boldsymbol{T}}) = \sigma^2 I_{n\times n}$.

Nestes modelos os parâmetros são de fácil interpretação, onde o parâmetro β_0 representa o valor esperado da variável resposta (Y) quando todas as variáveis explicativas $(x_j, j = \dots, p)$ são nulas e cada um dos parâmetros β_j indica qual a variação do valor esperado de Y por cada incremento unitário da variável x_j quanto todas as outras covariáveis se mantêm constantes. No entanto, apesar das fáceis interpretações, os parâmetros são desconhecidos, pelo que se torna de extrema importância conhecer uma técnica para os estimar.

Uma forma de obter essas estimativas é o Método de Mínimos Quadrados, desenvolvido por Legendre (1805), que consiste na minimização da soma do quadrado dos erros, isto é, minimizar a função:

$$SQE = SQE(\beta_0, \dots, \beta_p) = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2. \tag{3.4}$$

A resolução deste problema de otimização, numa fase inicial, passa por encontrar os valores dos parâmetros que anulam as derivadas parciais da função 3.4, ou seja, os valores de β_j com $j = 0, \dots, p$ que são solução da equação:

$$\begin{aligned} \frac{\partial SQE}{\partial \beta_k} = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} 2(y_i - \sum_{j=0}^{p} \beta_j x_{ij})(-x_{ik}) = 0 \\ &\Leftrightarrow \quad \sum_{i=1}^{n} y_i x_{ik} = \sum_{j=0}^{p} \beta_j \sum_{i=1}^{n} x_{ij} x_{ik}, \text{com } k = 0, \dots, p. \end{aligned} \tag{3.5}$$

O sistema de $p+1$ equações lineares a $p+1$ incógnitas, obtido na equação 3.5, pode ser reescrito na forma matricial, da seguinte forma

$$(\boldsymbol{X^T X})\boldsymbol{\beta} = \boldsymbol{X^T Y}. \tag{3.6}$$

Neste caso, se a matriz $\boldsymbol{X^T X}$ for invertível, o vetor dos estimadores de mínimos quadrados dos coeficientes de regressão linear é dado por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T Y}). \tag{3.7}$$

Caso a matriz $\boldsymbol{X^T X}$ não seja invertível, verifica-se a existência de multicolinearidade, i.e, uma ou mais variáveis que são combinação linear entre si. Nestes casos, as variáveis que resultam da combinação linear de outras devem ser retiradas do modelo.

Para a verificação do valor do zero da derivada com um mínimo de SQE, são descritas as definições de valores ajustados e de resíduos, respetivamente.

Os valores ajustados são os valores que, em cada $(p+1)$-úplo observado $(x_{i0}, x_{i1}, \ldots, x_{ip})$, se encontram sobre o hiperplano $(p+1)$-dimensional ajustado, isto é, os valores que verificam $\hat{y}_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij}, i = 1, \ldots, n$. Numa perspetiva matricial, tem-se

$$\hat{\boldsymbol{Y}} = \begin{bmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{bmatrix}. \tag{3.8}$$

Os resíduos correspondem às estimativas dos termos de erro e são dados pelas diferenças entre os valores observados e os valores ajustados $e_i = y_i - \hat{y}_i$, para $i = 1, \ldots, n$, o que em notação matricial pode ser reescrito como

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{3.9}$$

Com estas definições prova-se que os resíduos são ortogonais à matriz de planeamento, isto é, que se verifica a igualdade $\boldsymbol{X^T e} = \boldsymbol{0_{(p+1)\times 1}}$, pois tem-se que

$$\boldsymbol{X^T e} = \boldsymbol{X^T}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{X^T}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X^T Y} - \boldsymbol{X^T X}\hat{\boldsymbol{\beta}} \tag{3.10}$$

e $\boldsymbol{X^T Y} - \boldsymbol{X^T X}\hat{\boldsymbol{\beta}}$ corresponde ao vetor nulo de dimensão $(p+1) \times 1$, pois corresponde ao conjunto das equações normais. Note-se que se todas as colunas da matriz de planeamento são ortogonais ao vetor dos resíduos e se a coluna do termo constante tem todos os elementos iguais à unidade, então facilmente se conclui que a soma dos resíduos é nula,

ou seja,

$$\sum_{i=1}^{n} e_i = 0. \tag{3.11}$$

A ortogonalidade entre as colunas da matriz $\boldsymbol{X}$ e o vetor de resíduos $\boldsymbol{e}$ tem como consequência a ortogonalidade entre estes e o vetor de valores ajustados, uma vez que

$$\hat{\boldsymbol{Y}}\boldsymbol{e} = \hat{\boldsymbol{\beta}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{e} = \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\mathbf{0}_{(\boldsymbol{p+1})\times\mathbf{1}} = 0. \tag{3.12}$$

Uma vez comprovada a ortogonalidade dos resíduos em relação à matriz de planeamento, torna-se possível a verificação de que os estimadores de máxima verosimilhança correspondem a um mínimo da função representada em 3.4. Para tal, tendo como base a equação 3.4 em notação matricial:

$$\begin{aligned} SQE &= (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + \mathbf{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &+ (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}). \end{aligned} \tag{3.13}$$

Com o resultado acima demonstrado de que os resíduos são ortogonais à matriz de planeamento verifica-se a igualdade

$$2(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) = 2(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{e} = 0,$$

pelo que SQE pode ser simplificada em

$$SQE = (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}). \tag{3.14}$$

O primeiro termo da expressão 3.14, $(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$, não depende de $\boldsymbol{\beta}$ e o segundo termo, $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ é não negativo uma vez que não é mais do que a soma de quadrados. Assim, o mínimo da função é atingido no ponto que anular o segundo termo da soma, isto é, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

A combinação das condições de "Gauss-Markov"com o pressuposto da normalidade dos erros tem-se que $\boldsymbol{Y}$ tem distribuição Normal com $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ e $Cov(\boldsymbol{Y}) = \sigma^2\mathbf{I}[n\times n]$, pelo que a verosimilhança da amostra é dada por:

$$L(y_1,\ldots,y_n;\beta,\sigma^2) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2\sigma^2}(y_i-\sum_{j=0}^{p}\beta_j x_{ij})^2\right)\right]. \tag{3.15}$$

Linearizando e simplificando a equação 3.15 é obtida uma função simétrica à função obtida em 3.4. Uma vez que a maximização de $-SQE(\boldsymbol{\beta})$ é equivalente à minimização de $SQE(\boldsymbol{\beta})$, os estimadores de mínimos quadrados para $\boldsymbol{\beta}$ são também os seus estimadores de máxima

verosimilhança, quando as observações têm distribuição Normal.

3.1.1 Propriedades dos estimadores

Os estimadores de mínimos quadrados, sob a validade das "Condições de Gauss Markov", gozam de boas propriedades estatísticas (Fahrmeir & Kneib, 2013).

Considere-se um estimador $\hat{\theta}$ para um parâmetro θ. Este é considerado centrado, quando $E(\hat{\theta}) = \theta$. Na Regressão Linear, verificada a primeira condição constante das "Condições de Gauss Markov", é estabelecida a nulidade do valor esperado dos erros, para que se chegue à conclusão que $\hat{\boldsymbol{\beta}}$ são estimadores centrados para $\boldsymbol{\beta}$, uma vez que

$$E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T})\boldsymbol{E(Y)} = (\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T X})\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{3.16}$$

Considerando que os erros para além de verificarem a primeira condição, verificam também a segunda, isto é, admitindo que os erros, apresentam valor médio nulo e são variáveis aleatórias não correlacionadas de variância constante (σ^2), então conclui-se que a matriz de covariâncias dos estimadores de mínimos quadrados é dada por:

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}) &= Cov[(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}] = [(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}]Cov(Y)[(\boldsymbol{X^T X^{-1} X^T})]^{\boldsymbol{T}} \\ &= [(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}]\sigma^2\boldsymbol{I_{n\times n}}[(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}]^{\boldsymbol{T}} = \\ &= \sigma^2(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T X}(\boldsymbol{X^T X})^{-1} = \sigma^2(\boldsymbol{X^T X})^{-1}. \end{aligned} \tag{3.17}$$

É ainda garantida a consistência dos estimadores de mínimos quadrados sempre que a soma dos elementos da diagonal principal da matriz $(X^T X)^{-1}$ tenda para zero, à medida que a dimensão amostral se aproxima do infinito, o que é equivalente a escrever

$$\lim_{n\to+\infty} \text{tr}((X^T X)^{-1}) = 0.$$

O Teorema de Gauss-Markov garante que os estimadores de mínimos quadrados (EMQ) são os estimadores lineares centrados de variância mínima. Desta forma, os EMQ do modelo de regressão múltipla são estimadores BLUE (*Best Linear Unbiased Estimators*) ou seja, de entre todos os estimadores lineares centrados são aqueles que possuem variância mínima.

De realçar que as propriedades apresentadas são válidas independentemente da matriz de planeamento e da distribuição de probabilidade dos erros aleatórios, que reforça o facto de que o método dos mínimos quadrados produz bons estimadores em condições muito gerais. Então, sem a verificação do pressuposto da normalidade dos erros aleatórios considera-se que os EMQ são estimadores BLUE, ou seja, dentro da classe dos estimadores lineares são centrados e de variância mínima.

Se se admitir a normalidade dos resíduos, ou seja, os erros aleatórios são variáveis aleatórias independentes e indenticamente distribuídas à distribuição Normal, onde $\epsilon \sim N(0, \sigma^2)$,

pode-se concluir que os EMQ são estimadores de variância mínima não só dentro da classe dos estimadores lineares, mas também dentro da classe dos estimadors não lineares.

3.1.2 Estimação de σ^2

O resultado obtido na equação 3.17, na Subsecção 3.1.1, depende do valor de σ^2, que geralmente é desconhecido e que necessita de ser estimado. A estimação deste parâmetro pode ser realizada com recurso à soma dos quadrados dos resíduos $\sum_{i=1}^{n} e_i^2$, usualmente denominada como soma de quadrados dos erros e denominada por SQE.

Para a obtenção dos resultados pretendidos defina-se a matriz $\mathbf{H}$ e a matriz $\boldsymbol{M}$, respetivamente, que têm um papel fundamental na dedução de algumas das propriedades que se apresentam.

A matriz $\boldsymbol{H}$ é uma matriz $n \times n$, tal que

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}. \tag{3.18}$$

Esta matriz permite enfatizar o facto de que cada um dos valores ajustados pode ser escrito como função linear do valores observados, uma vez que:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y} = \boldsymbol{HY} \tag{3.19}$$

Sendo $\boldsymbol{H}$ a matriz que transforma $\boldsymbol{Y}$ em $\hat{\boldsymbol{Y}}$ é usualmente designada por matriz *hat*.

A matriz $\boldsymbol{M}$ é também uma matriz de dimensões $n \times n$ definida como

$$\boldsymbol{M} = \boldsymbol{I_n} - \boldsymbol{H} \tag{3.20}$$

e que, consequentemente, verifica

$$\boldsymbol{MX} = (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{0}_{\boldsymbol{n\times(p+1)}}. \tag{3.21}$$

Esta matriz, à semelhança de $\boldsymbol{H}$, permite enfatizar o facto de que cada um dos resíduos se pode escrever como função linear dos erros aleatórios, uma vez que

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = \boldsymbol{MY} = \boldsymbol{M}(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) = \boldsymbol{M\epsilon}. \tag{3.22}$$

Acrescente-se que qualquer uma das duas matrizes acima definidas, $\boldsymbol{H}$ e $\boldsymbol{M}$, é uma matriz simétrica e idempotente.

Durante a estimação de σ^2, é necessário evidenciar que se $\boldsymbol{M} = (m_{ij})$ é a matriz simétrica e idempotente definida na expressão 3.20, então verifica-se

$$\sum_{i=1}^{n} e_i^2 = \boldsymbol{\epsilon^T \epsilon} = \boldsymbol{\epsilon^T M^T M \epsilon} = \boldsymbol{\epsilon^T M \epsilon} = \sum_{i=1}^{n} m_{ii}\epsilon_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} m_{ij}\epsilon_i\epsilon_j.$$

E desta forma,

$$E\left[\sum_{i=1}^{n} e_i^2\right] = \sum_{i=1}^{n} m_{ii}E[\epsilon_i^2] + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} m_{ij}E[\epsilon_i\epsilon_j].$$

Considerando que os erros são não correlacionados com variância comum σ^2, a expressão anterior pode ser simplificada em

$$\begin{aligned} E\left[\sum_{i=1}^{n} \epsilon_i^2\right] &= \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma tr(\boldsymbol{M}) = \sigma^2 tr(\boldsymbol{I}_n - \boldsymbol{H}) = \sigma^2(n - tr(\boldsymbol{H})) \\ &= \sigma^2(n - tr(\boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})) = \sigma^2(n - tr(\boldsymbol{X^T X}(\boldsymbol{X^T X})^{-1})) \quad (3.23) \\ &= \sigma^2(n - tr(\boldsymbol{I}_{p+1})) = \sigma^2(n - p - 1), \end{aligned}$$

uma vez que o traço de qualquer matriz identidade de ordem k é dado pela soma dos elementos da sua diagonal, que como são todos unitários somam k e que, dadas duas quaisquer matrizes $\boldsymbol{A}$ e $\boldsymbol{B}$, tais que $\boldsymbol{A}$ é $m \times l$ e $\boldsymbol{B}$ é $l \times m$ se tem $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$. Assim sendo, o estimador dado por

$$S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p - 1} \quad (3.24)$$

é um estimador centrado e consistente para σ^2.

Resumidamente se afirma que a análise de resíduos é a metodologia que permite a validade das "Condições de Gauss Markov".

Dado o papel fundamental que os resíduos desempenham, é descrita a Partição da Soma de Quadrados, ou seja, o resultado que permite decompor a variabilidade total de Y- Soma dos Quadrados Total: SQT - na soma de duas parcelas: uma representando a variabilidade não explicada pelo modelo de regressão, ou seja, a variabilidade existente na presença de erros aleatórios - Soma dos Quadrados dos Resíduos: SQE - e outra, a variabilidade de Y que o modelo consegue explicar - Soma dos Quadrados da Regressão: SQR. O resultado da Partição da Soma de Quadrados é de fácil demostração, como se pode verificar:

$$\begin{aligned} SQT &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i\hat{Y}_i)(\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \\ &= SQE + 2\sum_{i=1}^{n}\epsilon_i(\hat{Y}_i - \overline{Y}) + SQR = SQE + 2\sum_{i=1}^{n}\epsilon_i\hat{Y}_i - 2\overline{Y}\sum_{i=1}^{n}\epsilon_i + SQR \\ &= SQE + SQR. \quad (3.25) \end{aligned}$$

3.1.3 Testes de hipóteses sobre os coeficientes de regressão

A componente aleatória que constitui o modelo de regressão linear múltipla, $\epsilon_i, \forall i = 1, \ldots, n$, admite uma distribuição de probabilidade que usualmente, para efeitos de inferência estatística, se supõe ser Normal, sendo este um dos pressupostos de elevada importância uma vez que a sua violação não permite a obtenção da distribuição de probabilidade de um conjunto de variáveis aleatórias que constituem a base do processo de construção de intervalos de confiança e/ou testes de hipóteses para os parâmetros do modelo.

Em Alpuim (2013) a elaboração dos testes de hipóteses é baseada nas "Condições de Gauss Markov", da normalidade dos termos de erro e, também, no seguinte teorema:

Teorema 1. *Seja* $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ *um modelo linear em que* $\epsilon = [\epsilon_1 \ldots \epsilon_n]^T$ *é um vetor de variáveis aleatórias independentes e identicamente distribuídas, com distribuição Normal* $N(0, \sigma^2)$*. Então,*

1. *O estimador de mínimos quarados do vetor de parâmetros* $\boldsymbol{\beta}$*, isto é,* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}$ *tem distribuição Normal multivariada,* $N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X^T X})^{-1})$*;*

2. *A variável aleatória*
$$\frac{(n-p-1)S^2}{\sigma^2} = \frac{\boldsymbol{\epsilon^T \epsilon}}{\sigma^2} \tag{3.26}$$
tem distribuição Qui-quadrado com $n-p-1$ *graus de liberdade;*

3. $\hat{\boldsymbol{\beta}}$ *e* S^2 *são independentes.*

De notar que com este resultado é possível construir qualquer uma das estatísticas de teste subjacentes aos testes de hipóteses utilizados no presente estudo.

Nulidade de um coeficiente: Teste t

Tendo em conta o princípio da parcimónia na elaboração de um modelo preditivo, ou seja, num cenário em que existam dois modelos distintos, que não difiram significativamente no que diz respeito à qualidade do ajustamento, deve optar-se sempre pelo mais simples. Ao adicionar uma variável a um modelo de regressão são esperadas as seguintes situações: aumenta a soma de quadrados da regressão e a soma de quadrados dos erros diminui, por outro lado a variância dos valores ajustados aumenta. Assim sendo, deve-se ter em conta que apenas se devem incluir num modelo as covariáveis que explicam a variável resposta, ou seja, aquelas cujo coeficiente é estatisticamente diferente de zero.

Desta forma, nesta Secção será explorado o problema de avaliar a significância de cada uma das covariáveis *per si*, ou seja, serão realizados testes à nulidade de cada um dos parâmetros $\beta_j, j = 1, \ldots, p$, de forma isolada. As hipóteses em teste serão então dadas por

$$H_0 : \beta_j = 0 \ vs \ H_1 : \beta_j \neq 0, j \in \{1, \ldots, p\}. \tag{3.27}$$

A não rejeição de H_0 significa, neste caso, não rejeitar a nulidade do parâmetro em teste, o que equivale a dizer que a variável x_j que lhe está subjacente não influencia significativamente a variável resposta Y e, consequentemente, não deve ser incluída no modelo.

No ponto 1 do Teorema 1 é referido que $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X^T X})^{-1})$, pelo que se se considerar z_{ij} como o j-ésimo elemento da diagonal principal da matriz $(\boldsymbol{X^T X})^{-1}$ tem-se que $Var(\hat{\boldsymbol{\beta_j}}) = \sigma^2 z_{ij}$. Utilizando os pontos do Teorema 1 e tendo em conta que o quociente entre uma Normal padrão e a raíz de uma Qui-quadrado a dividir pelo seu número de graus de liberdade tem distribuição t de Student com esses mesmos graus de liberdade, conclui-se que, sob a validade da hipótese nula, uma estatística de teste possível para testar as hipóteses mencionadas é

$$T = \frac{\frac{\hat{\boldsymbol{\beta_j}}}{\sqrt{\sigma^2 z_{jj}}}}{\sqrt{\frac{\frac{(n-p-1)S^2}{\sigma^2}}{n-p-1}}} = \frac{\hat{\boldsymbol{\beta_j}}}{S\sqrt{z_{jj}}} \sim t_{n-p-1}, j \in \{1, \dots p\}.$$

A região de rejeição do teste bilateral, cujas hipóteses foram definidas em 3.27, é dada por

$$\frac{|\hat{\boldsymbol{\beta_j}}|}{S\sqrt{z_{jj}}} > t_{n-p-1;1-\frac{\alpha}{2}},$$

em que $t_{n-p-1;1-\frac{\alpha}{2}}$ representa o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição t de Student com $n-p-1$ graus de liberdade.

Em qualquer modelo preditivo é imprescindível reconhecer quais as variáveis que melhor explicam o acontecimento que se pretende modelar e podendo existir um grande número de combinações de variáveis a considerar, é necessário determinar qual o subconjunto destas que melhor explica a variável resposta entre todas as covariáveis disponíveis.

Note-se que a seleção de apenas um subconjunto de variáveis implica uma equidade entre o compromisso de obtenção do máximo de informação possível e o da obtenção de estimativas com variância o mais reduzida possível. Desta forma, surgiram vários métodos de seleção de variáveis, uma vez que a escolha destas é um dos dilemas na análise de regressão. Salientam-se os métodos de seleção de covariáveis considerados neste estudo: o método regressivo (*backward elimination*), o método progressivo (*forward selection*) e o método passo a passo (*stepwise method*), que foram verificados em Chatterjee & Hadi (2009) e em Chatterjee (2000):

- **Método de seleção regressiva (*Backward elimination*)**- é um método de exclusão de variáveis, uma vez que se inicia com o modelo completo e em cada iteração é eliminada a variável menos significativa, até que se obtenha um modelo em que todas as variáveis que o constituem são significativas;

- **Método de seleção progressiva (*Forward selection*)**- é um método de inclusão

de variáveis, uma vez que o procedimento se inicia com o modelo nulo e as variáveis vão sendo adicionadas, uma a uma, conforme a importância que têm na explicação da variável resposta. O processo termina quando se chega a uma variável que já não acrescenta valor ao modelo;

– **Método de seleção (*Stepwise method*)**- é um método que não segue uma direção única no que toca à seleção das variáveis. Este método consiste na inclusão e remoção sequencial de variáveis independentes, até que não existam mais variáveis significativas a incluir no modelo, ou que todas as variáveis já incluídas no modelo sejam significativas.

Nulidade de todos os coeficientes: Teste F

Nos modelos de Regressão Linear Múltipla, caso todos os coeficientes de regressão sejam nulos, com exceção daquele que corresponde ao termo constante (β_0), a equação do modelo reduz-se a $Y = \beta_0 + \epsilon$. Logo, admitindo que todas as covariáveis têm coeficiente nulo, assume-se que nenhuma dessas variáveis tem poder explicativo sobre a variável resposta e, consequentemente, o ajustamento de um modelo linear ao conjunto de observações em consideração não é adequado. Desta forma, para averiguar a utilidade do modelo testa-se a hipótese de que todos os coeficientes de regressão sejam simultaneamente nulos, ou seja, realizar um teste sob as hipóteses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_j = 0 \ vs \ H_1 : \exists j \in \{1, \ldots, p\} : \beta_j \neq 0.$$

A não rejeição da hipótese nula indica que não existe razão para que não se considerem nulos todos os coeficientes de regressão e, portanto, o modelo é inútil. Todavia, a não rejeição da hipótese nula não implica uma possível melhoria do modelo, apenas indica que pelo menos uma das covariáveis contribui significativamente para explicar a variável resposta

Um teste para esta hipótese baseia-se na Partição da Soma de Quadrados que se apresentou em 3.25. A divisão dos membros dessa igualdade por σ^2 resulta em:

$$\frac{SQT}{\sigma^2} = \frac{SQE}{\sigma^2} + \frac{SQR}{\sigma^2} \Leftrightarrow \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sigma^2} + \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sigma^2}. \quad (3.28)$$

De forma de avaliar a utilidade do ajuste do modelo de regressão aos dados é efetuada a comparação da fração da variância explicada pelo modelo de regressão (SQR) com a da variância atribuída aos resíduos (SQE). Caso a primeira seja significativamente superior à segunda, conclui-se que o modelo é significativo. Esta comparação é efetuada com base na distribuição estatística da razão entre estas duas variâncias.

Ora, pelo Teorema 1 sabe-se que $\frac{SQE}{\sigma^2} = \frac{\boldsymbol{e^T e}}{\sigma^2}$ tem distribuição Qui-quadrado com $n - p - 1$ graus de liberdade. Sob a validade de H_0, o modelo resume-se ao modelo nulo e, portanto, $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ coincide com $\sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}(\epsilon_i - \overline{\epsilon})^2$, ou seja, com a

soma dos quadrados dos desvios à media de variáveis aleatórias Normais independentes e identicamente distribuídas com valor médio nulo e variância σ^2, que se sabe ter distribuição Qui-quadrado com $n-1$ graus de liberdade.

Isolando $\frac{SQR}{\sigma^2}$ do lado esquerdo da expressão 3.28 rapidamente se conclui que a parcela em epígrafe tem também distribuição com $p = (n-1)-(n-p-1)$ graus de liberdade, uma vez que resulta da diferença entre duas covariáveis ambas com distribuição Qui-quadrado.

Como $\frac{SQR}{\sigma^2}$ e $\frac{SQE}{\sigma^2}$ são duas quantidades independentes, a estatística de teste, que corresponde ao quociente entre elas, sob a validade da hipótese nula, segue uma distribuição F com p e $n-p-1$ graus de liberdade, ou seja,

$$F = \frac{\frac{SQR}{p}}{\frac{SQE}{n-p-1}} \sim F_{p;n-p-1}.$$

Obtendo valores elevados da estatística de teste, leva a concluir que pelo menos uma das covariáveis é significativa na explicação da variabilidade das observações e, por isso, nesses casos rejeita-se H_0.

É usual apresentar os resultados de uma análise como a que se acabou de descrever numa tabela ANOVA (*Analysis Of Variance*) conforme se apresenta na Tabela 3.1.

Tabela 3.1: Tabela ANOVA.

Origem da variação	Soma de quadrados	Graus de liberdade	Média de quadrados	Estatística de teste
Regressão	SQR	p	MQR	$F_0 = \frac{MQR}{MQE}$
Resíduos	SQE	$n-p-1$	MQE	
Total	SQT	$n-1$		

3.1.4 Validação de pressupostos e análise de resíduos

No decorrer da corrente Secção assumiu-se que os erros verificavam as "Condições de Gauss Markov". Além disso, em algumas secções foi ainda necessário adicionar o pressuposto da normalidade dos erros. Na prática, estes pressupostos não são sempre garantidos. De facto, é bastante frequente que pelo menos um deles seja violado. Desta forma, nesta Secção apresentar-se-ão algumas sugestões sobre como verificar cada um dos pressupostos e o que fazer quando algum destes falha.

A validação dos pressupostos do modelo de regressão linear baseia-se numa análise pormenorizada dos resíduos do modelo, uma vez que estes representam as diferenças entre aquilo que foi realmente observado e o que foi estimado através da equação de regressão. Assim, se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do modelo.

A verificação dos diversos pressupostos baseia-se essencialmente em métodos gráficos, sendo corroborada com a aplicação de testes estatístico sempre que necessário.

Nesta Secção, presume-se que quando se verifica se um pressuposto está a ser violado, os demais são válidos. Todavia, sabe-se, pela experiência de muitos autores, que a falha

das "Condições de Gauss Markov" afeta mais os testes à normalidade do que a falha da normalidade afeta o diagnóstico às "Condições de Gauss Markov" (Sen & Srivastava, 2012).

A verificação das "Condições de Gauss Markov" pressupõe que os resíduos, por serem estimativas dos termos de erro, sejam independentes e apresentem média nula e variância constante. Estes pressupostos podem ser verificados graficamente, representando os resíduos em função dos valores estimados da variável dependente ou em função de cada uma das covariáveis. Os pontos desse gráfico devem distribuir-se de forma aleatória em torno da reta que corresponde ao resíduo zero, formando uma mancha de largura uniforme.

Quando os resíduos não se comportam de forma aleatória, ou seja, seguem um padrão, a condição de independência não é satisfeita, o que pode indicar que não existe uma relação linear entre as variáveis, ou que não constam do modelo uma ou mais covariáveis que influenciam estatisticamente a variável dependente e portanto também os erros.

O pressuposto da independência pode ser, também, avaliado através da observação da Função de Autocorrelação (FAC) dos resíduos. Uma metodologia alternativa para avaliação da validade do pressuposto de independência envolve o uso da estatística de Durbin-Watson, que testa a hipótese nula de independência (sem autocorrelação). A estatística de Durbin-Watson é dada por

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \tag{3.29}$$

onde e_i corresponde ao resíduo da observação i e $0 \leq DW \leq 4$. Quando $DW \approx 2$, não se rejeita a hipótese de independência. Além disso, Wheelwright (1998) refere que valores inferiores a 2 indicam a existência de autocorrelação positiva, enquanto que valores superiores a 2 revelam autocorrelação negativa.

Se os resíduos apresentam um comportamento tendencialmente crescente ou decrescente com os valores das covariáveis, ou com os valores estimados da variável dependente, deve ser posta em causa a hipótese de variância constante dos resíduos. Quando o pressuposto da homocedasticidade é violado pode recorrer-se a uma transformação na variável dependente de forma a estabilizar a variância. Note-se que nem sempre os dados dão indícios de qual a transformação adequada a utilizar e, por isso, sempre que não é possível escolher empiricamente a transformação, o melhor é optar por uma técnica mais objetiva. Um dos procedimentos que permite escolher transformações de maneira relativamente automática é um procedimento da família de transformações potência denominado Box-Cox. Esta metodologia aplica-se quando a variável resposta assume valores positivos e a transformação da variável Y é dada por

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}. \tag{3.30}$$

Usualmente, o parâmetro λ é estimado com base no método da máxima verosimilhança,

assumindo-se que a transformação das respostas $\mathbf{Y}(\lambda)$ tem distribuição Normal multivariada com matriz de valor médio $X\beta$ e matriz de covariâncias $\sigma^2\mathbf{I}_n$. Com este pressuposto, facilmente se verifica que a função de densidade $\mathbf{Y}(\lambda)$ é dada por

$$f(\mathbf{Y}(\lambda))f(\mathbf{Y}(\lambda)) = \frac{\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}(\lambda) - X\beta)^T(\mathbf{Y}(\lambda) - X\beta)\right\}}{(2\pi\sigma^2)^{\frac{n}{2}}}.$$

Se se denotar por $J(\lambda, \mathbf{Y})$ o jacobiano da transformação de $\mathbf{Y}$ em $\mathbf{Y}(\lambda)$, a densidade de $\mathbf{Y}$ é dada por

$$L(\lambda, \beta, \sigma^2|\mathbf{Y}, \mathbf{X}) = f(\mathbf{Y}) = \frac{\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}(\lambda) - X\beta)^T(\mathbf{Y}(\lambda) - X\beta)\right\}}{(2\pi\sigma^2)^{\frac{n}{2}}} J(\lambda, \mathbf{Y}). \tag{3.31}$$

Assumindo que λ é fixo, os estimadores de máxima verosimilhança de $(\beta(\lambda), \sigma^2(\lambda))$ são dados por

$$\hat{\beta}(\lambda) = (X^TX)^{-1}XY(\lambda) \text{ e } \hat{\sigma^2}(\lambda) = \frac{Y(\lambda)^THY(\lambda)}{n}, \tag{3.32}$$

em que H é a matriz *hat*.

Substituindo na expressão 3.31 os valores $(\beta(\lambda), \sigma^2(\lambda)$ pelos valores obtidos na expressão 3.32 e tendo em conta que $J(\lambda, Y) = \prod_{i=1}^n \lambda - 1$, obtém-se a função de log-verosimilhança maximizada sobre $(\beta(\lambda), \sigma^2(\lambda)$, com λ fixo, dada por:

$$\log\left(L(\lambda|Y, X, \hat{\beta}(\lambda), \hat{\sigma^2}(\lambda)\right) = C - \frac{n}{2}\log\hat{\sigma^2}(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i). \tag{3.33}$$

Basta agora maximizar a função 3.33 para se obter uma estimativa para λ. A função 3.30 pode ainda ser modificada de forma a acomodar valores não positivos de Y:

$$y(\lambda) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \lambda_1 \neq 0 \\ \log(y + \lambda_2), & \lambda_1 = 0 \end{cases} \tag{3.34}$$

em que $\lambda = (\lambda_1, \lambda_2)^T$. Na prática, escolhe-se para λ_2 o valor que garante que $y + \lambda_2 > 0$ qualquer que seja y e para λ_1 o valor que maximiza a função 3.33.

Uma vez estabilizada a variância e verificados os restantes pressupostos das "Condições de Gauss Markov", falta ainda verificar o pressuposto da normalidade dos resíduos, visto que toda a inferência estatística por detrás do modelo de regressão linear se baseia também neste pressuposto.

As representações gráficas mais usuais para a validação deste pressuposto são o histograma e o papel de probabilidade. A primeira é a imagem estatística da função densidade, pelo que a sua representação sugere a da função densidade da população subjacente à amostra, neste caso a distribuição dos resíduos. A segunda, usualmente denominada por *QQ-plot*, consiste na representação dos quantis teóricos da distribuição que se assume para os resíduos contra os quantis empíricos destes. Assim, no primeiro caso, quanto mais pró-

ximo da forma de sino, caraterística ds distribuição Normal, estiver a representação gráfica, menor a probabilidade de que o pressuposto em causa não se verifique. Já no segundo caso, espera-se a não violação do pressuposto quando os pares de pontos do gráfico se posicionam em torno da bissetriz dos quadrantes ímpares.

Complementando a análise gráfica, existem os testes de ajustamento. Estes permitem uma verificação menos subjetiva do pressuposto. Existem vários testes de ajustamento, sendo os mais utilizados o de Shapiro-Wilk e o de Kolmogorov-Smirnov.

Note-se que nem sempre as representações gráficas permitem detetar a violação de determinados pressupostos. Por outro lado, os testes de ajustamento quando aplicados a amostras de dimensão elevada podem conduzir à rejeição da normalidade, mesmo quando a distribuição subjacente aos dados é muito próxima da Normal. Assim sendo, o mais adequado, quando se pretende verificar o pressuposto da normalidade dos resíduos, é combinar uma verificação visual com uma verificação analítica.

A falha do pressuposto da normalidade não é condição suficiente para que as inferências realizadas no decorrer da construção do modelo linear não sejam válidas. Na verdade, Sen & Srivastava (2012) demonstram que uma vez garantidas as "Condições de Gauss-Markov", se se verificar que o máximo da diagonal da matriz *hat* é próximo de zero, então as distribuições das estatísticas de teste t e F, apresentadas na Secção 3.1.3, mantêm-se e, consequentemente, os resultados dos referidos testes não devem ser postos em causa. Na prática não é fácil definir o valor $\underset{i=1,\ldots,n}{max}(\mathbf{H}_{ii})$ a partir do qual podemos afirmar que os resultados dos testes são ainda válidos, mesmo que a normalidade dos resíduos falhe. No entanto, considera-se que 0,2 é um valor suficientemente pequeno para garantir que os resultados são ainda válidos, pelo que mesmo que os resíduos não sejam normais se a desigualdade $\underset{i=1,\ldots,n}{max}(\mathbf{H}_{ii}) < 0,2$ se verificar, os resultados obtidos mantém-se fiáveis.

Note-se que a transformação Box-Cox anteriormente mencionada não serve apenas para estabilizar a variância, mas também para estandardizar os resíduos. Todavia, esta transformação pode afetar a relação existente entre as variáveis dependentes e independentes, fazendo com que esta deixe de ser linear, situações em que o mais adequado é transformar também as covariáveis. Como esta situação não se verificou no decorrer do presente trabalho, não se irá detalhar aqui.

3.1.5 Qualidade do modelo e análise de R^2

Na Secção 3.1.3, já se viu que é possível decompor a variabilidade total da amostra (SQT) na soma de quadrados residual (SQE), com a soma de quadrados devida à regressão (SQR). Com base nesta decomposição, expressa na equação 3.25, define-se coeficiente de determinação, (R^2), como a percentagem de variação da amostra que é explicada pelo modelo de regressão, o que se traduz na expressão

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}. \tag{3.35}$$

Se a soma de quadrados residual toma o valor mínimo de zero ($SQE = 0$), então o coeficiente de determinação é unitário ($R^2 = 1$) e estamos perante um ajustamento perfeito. Já quando é a soma de quadrados de regressão que se anula ($SQR = 0$), o valor da soma de quadrados residual é máximo, coincidindo com o valor da variabilidade total da amostra e, consequentemente, o valor do coeficiente de determinação é nulo. Neste cenário, o modelo linear em nada contribui para explicar a variabilidade das observações. Assim sendo, é intuitivo compreender que o coeficiente de determinação está compreendido entre 0 e 1 e que quanto mais próximo estiver da unidade, melhor é o ajustamento do modelo. Contudo, um valor elevado de R^2 não implica necessariamente que o modelo de regressão esteja bem ajustado, uma vez que a adição de uma nova variável aumenta sempre o seu valor, mesmo que essa variável não seja estatisticamente significativa. Por outro lado, se a variância dos termos de erro for de facto elevada porque faltam variáveis no modelo, este coeficiente tende a ser reduzido, o que não significa necessariamente que o modelo esteja mal ajustado. Desta forma, o coeficiente de determinação não só deve ser utilizado como precaução como deve ser encarado como uma medida da utilidade do modelo e não como medida da qualidade do seu ajustamente.

Note-se que existem outros indicadores que medem a qualidade do ajustamento, como o coeficiente de determinação ajustado ou o Critério de Informação de Akaike. Qualquer um deles tem a vantagem de levar em conta o número de covariáveis utilizadas, mas ambos têm a desvantagem da perda do compromisso entre a soma dos quadrados dos erros e a soma dos quadrados da regressão. Assim, estas duas medidas apenas são úteis quando se pretende comparar dois ou mais modelos, não evidenciando qualquer informação quando calculadas para um único modelo.

3.1.6 Predição

Uma aplicação muito importante de um modelo de regressão é a previsão de novas ou futuras observações de $\boldsymbol{Y}$, correspondentes a determinadas combinações das covariáveis, ou seja, estimar o valor da variável $\boldsymbol{Y^\star}$ quando o conjunto de covariáveis toma valores até então desconhecidos $\boldsymbol{Y^\star = x^\star \beta + \epsilon^\star}$, em que $\boldsymbol{\epsilon^\star}$ tem distribuição Normal, $N(0, \sigma^2)$. No entanto, o mais usual não é se queira uma estimativa pontual para a predição de novas observações, mas sim uma estimativa intervalar.

A construção da variável fulcral que está na base de cálculo do intervalo de confiança pretendido baseia-se na estandardização do erro de predição, variável que é combinação linear do vetor dos estimadores de mínimos quadrados e do termo de erro $\boldsymbol{\epsilon^\star}$, ambos normalmente distribuídos e independentes.

O erro padrão, dado por $\boldsymbol{\hat{Y}^\star - Y^\star = x^{\star T}(\hat{\beta} - \beta) - \epsilon^\star}$, é uma variável aleatória com distribuição Normal de valor médio nulo, já que se tem:

$$E\left(\boldsymbol{\hat{Y}^\star - Y^\star}\right) = E\left(\boldsymbol{x^{\star T}(\hat{\beta} - \beta) - \epsilon^\star}\right) = \boldsymbol{x^{\star T} E\,(\beta - \beta)} = 0 \tag{3.36}$$

e variância igual ao erro quadrático médio de $\hat{\mathbf{Y}}^\star$, ou seja, variância igual a:

$$\begin{aligned}
Var \quad & \left(\hat{\boldsymbol{Y}}^\star - \boldsymbol{Y}^\star\right) = EQM(\hat{\boldsymbol{Y}}^\star) = E\left[(\hat{\boldsymbol{Y}}^\star - \boldsymbol{Y}^\star)^2\right] = E\left[\left(\hat{\boldsymbol{Y}}^\star - \boldsymbol{Y}^\star\right)\left(\hat{\boldsymbol{Y}}^\star - \boldsymbol{Y}^\star\right)^{\boldsymbol{T}}\right] \\
= \quad & E\left[\left(\boldsymbol{x}^{\star\boldsymbol{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{\epsilon}^\star\right)\left(\boldsymbol{x}^{\star\boldsymbol{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{\epsilon}^\star\right)^{\boldsymbol{T}}\right] = \\
= \quad & E\left[\left(\boldsymbol{x}^{\star\boldsymbol{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{\epsilon}^\star\right)\left(\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\boldsymbol{T}}\boldsymbol{x}^\star - \boldsymbol{\epsilon}^{\star\boldsymbol{T}}\right)\right] = \\
= \quad & E\left[\boldsymbol{x}^{\star\boldsymbol{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{x}^\star - \boldsymbol{x}^{\star\boldsymbol{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\epsilon}^{\star\boldsymbol{T}} - \boldsymbol{\epsilon}^\star(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{x}^\star + \boldsymbol{\epsilon}^\star\boldsymbol{\epsilon}^{\star\boldsymbol{T}}\right] \\
= \quad & E\left[\boldsymbol{x}^{\star\boldsymbol{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{x}^\star - \boldsymbol{x}^{\star\boldsymbol{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\epsilon}^{\star\boldsymbol{T}} - \boldsymbol{\epsilon}^\star(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}\boldsymbol{x}^\star\right] + E\left[\boldsymbol{\epsilon}^\star\boldsymbol{\epsilon}^{\star\boldsymbol{T}}\right] = \\
= \quad & \boldsymbol{x}^{\star\boldsymbol{T}}Cov(\boldsymbol{\beta}^{\boldsymbol{T}})\boldsymbol{x}^{\star\boldsymbol{T}} + Cov(\boldsymbol{\epsilon}^\star) = \\
= \quad & \boldsymbol{x}^{\star\boldsymbol{T}}\boldsymbol{\sigma}^{\boldsymbol{2}}(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-\boldsymbol{1}}\boldsymbol{x}^\star + \boldsymbol{\sigma}^{\boldsymbol{2}} = \boldsymbol{\sigma}^{\boldsymbol{2}}\left[\boldsymbol{x}^{\star\boldsymbol{T}}(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-\boldsymbol{1}}\boldsymbol{x}^\star + 1\right]. \qquad (3.37)
\end{aligned}$$

Desta forma, a variável

$$\frac{\frac{\hat{\mathbf{Y}}^\star - \mathbf{Y}^\star}{\sqrt{\sigma^2[x^{\star T}(XX^T)^{-1}x^\star + 1]}}}{\sqrt{\frac{\frac{(n-p-1)S^2}{\sigma^2}}{n-p-1}}} = \frac{\hat{\mathbf{Y}}^\star - \mathbf{Y}^\star}{S\sqrt{x^{\star T}(XX^T)^{-1}x^\star + 1}} \qquad (3.38)$$

tem distribuição t de Student com $n - p - 1$ graus de liberdade e, consequentemente, o intervalo equilibrado de $(1 - \alpha)100\%$ de confiança para $\mathbf{Y}^\star$ é dado por

$$\left(\hat{\boldsymbol{Y}}^\star - t_{n-p-q,1-\frac{\alpha}{2}}S\sqrt{x^{\star T}(XX^T)^{-1}x^\star + 1}; \hat{\boldsymbol{Y}}^\star + t_{n-p-q,1-\frac{\alpha}{2}}S\sqrt{x^{\star T}(XX^T)^{-1}x^\star + 1}\right) \qquad (3.39)$$

em que $t_{n-p-1;1-\frac{\alpha}{2}}$ designa, como habitualmente, o quantil de ordem $\frac{\alpha}{2}$ da distribuição t de Student com $n - p - 1$ graus de liberdade.

3.2 Modelação da Sazonalidade

A componente de sazonalidade (S_t) incorporada num modelo de regressão pretende representar uma variabilidade periódica. Tal pode corresponder a um aumento/decréscimo que ocorre regularmente em determinados períodos do ano, originando oscilações que se repetem. Muitos dados são recolhidos mensalmente, tendo usualmente a série temporal associada uma forte componente sazonal, podendo esta ser explicada, por exemplo, por causas naturais, tais como as estações do ano ou outros fatores que influenciam de forma direta os valores obtidos

Na abordagem descrita por Gonçalves & Alpuim (2011), baseada nos modelos lineares, a componente sazonal γ_t, toma doze valores diferentes, λ_i, $i = 1, \ldots, 12$, cada um associado a um mês e expressam o desvio positivo ou negativo dos dados derivado ao efeito do mês. Este efeito é descrito com o auxílio de onze variáveis indicatrizes e a soma dos coeficientes

deve, no total, ser igual a zero. A componente sazonal é representada pela combinação linear de onze covariáveis, γ_t definidas por:

$$\gamma_t = \begin{cases} 1 & \text{se os dados no tempo } t \text{ correspondem ao mês } i,\ i = 2, \ldots, 12 \\ -1 & \text{se os dados no tempo } t \text{ correspondem ao mês 1} \\ 0 & \text{caso contrário.} \end{cases} \tag{3.40}$$

A componente sazonal relativa ao mês 1 pode ser obtida a partir da seguinte fórmula

$$\hat{\lambda_\perp} = -\sum_{i=2}^{12} \hat{\lambda_i}. \tag{3.41}$$

O modelo de regressão integra duas componentes: uma relativa à componente com variáveis indicatrizes, S, associadas à componente sazonal e, por fim, um erro estocástico. Com isto, o modelo com indicadores sazonais pode ser escrito como

$$Y_t = T_t + \beta_1 D_1 + \ldots + \beta_s D_s + \epsilon_t, \tag{3.42}$$

onde T_t representa a tendência (em função de t sem o termo constante, β_0), $\beta_1, \ldots, \beta_s$ são os coeficientes que refletem os s efeitos sazonais e D_i $(i = 1, \ldots, s)$ são as s variáveis indicatrizes que representam os diferentes períodos sazonais: tomam o valor 1 quando o tempo t pertence ao período i e 0 nos casos restantes, então β_i só é tido em consideração para observações registadas nesse mês.

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

Capítulo 4

Séries Temporais

Neste Capítulo são abordadas algumas das noções sobre Séries Temporais e Processos Estocásticos para compreensão dos tópicos abordados.

4.1 Conceito de Série Temporal

Designa-se de série temporal um conjunto de observações medidas de forma ordenada no tempo. Essas medições podem ser feitas continuamente no tempo ou apenas em momentos específicos, geralmente igualmente espaçados: dias, meses, trimestres ou anos.

Definição 1. *Uma série temporal consiste num conjunto de observações medidas sequencialmente no tempo. Desta forma, considerem-se o conjunto de observações* $Y_{t1}, Y_{t2}, \ldots, Y_{tn}$ *nos períodos* $t_1, t_2, \ldots, t_n$ *contados a partir de uma determinada origem.*

As séries temporais podem ser contínuas ($Y_t, t \in \mathbb{R}$) ou discretas ($Y_t, t = 1, 2, \ldots, n$) e, segundo Chatfield (2000, 2004), as séries são assim designadas independentemente da natureza da variável medida. Acrescente-se que, as séries temporais podem também ser classificadas em univariadas, se são constituídas por observações de uma só variável, e em multivariadas, se se observarem mais variáveis em cada instante.

Na análise de séries temporais contínuas, de uma forma geral, estas são transformadas em séries discretas com intervalos de tempo iguais. Este procedimento, de uma forma geral, não resulta numa perda significativa de informação sob a condição de o intervalo de amostragem ser suficientemente pequeno.

Regra geral, os dados de séries temporais têm uma propriedade muito peculiar: observações sucessivas são correlacionadas e, então, a análise desses dados deve ter em conta a ordem em que as observações são recolhidas. De facto, no caso de séries temporais univariadas, cada observação pode ser vista como bivariada, considerando que a segunda variável corresponde ao tempo em que esta é observada (Chatfield, 2000).

Conforme Box (2013) e Chatfield (2000), o estudo de séries temporais série tem como objetivos principais:

1. descrever os dados usando estatísticas descritivas e/ou métodos gráficos e compreender o mecanismo gerador da série, ou seja, procurar encontrar razões que justifiquem o comportamento da série, monitorizar a sua trajetória, analisar periodicidades relevantes nos dados, etc.;

2. encontrar um modelo estatístico adequado para descrever a evolução da série temporal. São vários os modelos propostos, entre os quais se destacam os modelos ARIMA;

3. prever o comportamento futuro da série, o que pode revelar-se extremamente útil na construção e execução de planos a curto, médio ou longo prazos e/ou no controlo de um determinado processo. Estas previsões podem ser a *1-passo*, se realizadas apenas para a observação seguinte, ou *multi-passos*, se englobam várias observações futuras.

Para Persons (1919), a variação de uma série temporal pode ser decomposta em 4 componentes: a tendência (T), a componente sazonal (S), a componente cíclica (C) e a componente irregular/residual (E). Por sua vez, o autora Alpuim (1998) descreve-as como:

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

Sazonalidade (S) corresponde a um padrão de aumento e diminuição que ocorre regularmente na série em períodos específicos, originando oscilações que se repetem. O movimento dentro de um período tem, então, duração fixa e é atribuído a fatores "sazonais", i.e., relacionados com aspetos do calendário (e.g., os meses ou trimestres de um ano ou os dias de uma semana). A sazonalidade pode ser classificada como aditiva, se não depende do nível da série, ou multiplicativa, quando é proporcional ao mesmo (Chatfield, 2000);

Componente cíclica (C trata-se de um padrão de flutuação que não apresenta qualquer periodicidade definida (i.e., a sua duração não é fixa e, portanto, o seu comprimento varia frequentemente de ciclo para ciclo) nem causa atribuída a fatores "sazonais". Generalizando, os ciclos são qualquer componente não-sazonal que apresenta um padrão reconhecível;

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

Na generalidade, os métodos para a análise de séries temporais baseiam-se na decomposição da variação da série nas componentes verificadas anteriormente. Considere-se Y_t

como o valor da série temporal no tempo t, T_t a tendência no tempo t, S_t a componente sazonal no tempo t e E_t a componente irregular no tempo t. Acrescente-se que a presença de uma componente cíclica (C) encontra-se incorporada na tendência .

Considere-se o modelo de decomposição aditivo que descreve cada valor da série temporal como sendo a soma das suas componentes, ou seja,

$$Y_t = T_t + S_t + E_t, \tag{4.1}$$

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

$$Y_t = T_t \times S_t \times E_t. \tag{4.2}$$

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

A decomposição multiplicativa é predominante, na medida que a maioria das séries em estudo apresentam uma variação sazonal que obedece ao nível da série. Nesses casos é aplicada uma transformação dos dados, nomeadamente a transformação logarítmica, de forma a converter um modelo multiplicativo num modelo aditivo. Desta forma, aplicando a transformação logarítimica em 4.2 obtém-se

$$\log Y_t = \log T_t + \log S_t + \log E_t. \tag{4.3}$$

Desta forma obtém-se um modelo multiplicativo por ajustamento de um modelo aditivo ao logaritmo dos dados. Note que um modelo de decomposição multiplicativo não deve ser implementado para séries temporais de valores negativos ou nulos (Caiado, 2011).

Os modelos de decomposição aditivo e multiplicativo não são as únicas formas de decompor uma série temporal e, com a sua combinação, podem origem a outros modelos que incluem relações tanto aditivas como multiplicativas. A título de exemplo, o resultado dessa combinação pode originar um modelo multiplicativo com erros aditivos expresso por

$$Y_t = T_t \times S_t + E_t. \tag{4.4}$$

Cleveland & Terpenning (1982) introduziram uma ferramenta proveitosa com vista à decomposição de séries temporais: os gráficos de decomposição. Estes gráficos permitem visualizar graficamente as várias componentes (Wheelwright, 1998). No ambiente `R`, segundo Hyndman (2019), aplicando a função `decompose`, a série temporal é dividida em três componentes: sazonalidade (S), tendência (T) e aleatoriedade (E). Na prática esta função torna-se vantajosa na medida que possibilita a avaliação das diversas componentes de forma distinta e, desta forma, auxiliar na identificação do comportamento das mesmas.

4.2 Processos Estocásticos

Considerando um processo estocástico $\{Y(t),\ t \in \mathcal{T}\}$, uma série temporal é um conjunto de observações do processo estocástico em instantes $t_1, t_2, \ldots, t_n$. Generalizando, considera-se t inteiro (i.e., $t = 0, \pm1, \pm2, \ldots$) e as observações são feitas em intervalos de tempo regulares, isto é, com a mesma amplitude (Alpuim, 1998).

Definição 2. *Um processo estocástico é qualquer família ou coleção de variáveis aleatórias* $Y(t), t \in \mathcal{T}$*, em que* $\mathcal{T}$ *é um conjunto de índices representando o tempo.*

O conjunto de índices $\mathcal{T}$ designa-se de *espaço de parâmetros* e o contradomínio das variáveis aleatórias $Y(t)$ é definido de *espaço de estados*, representado por S. Quanto à natureza de $\mathcal{T}$, se $\mathcal{T} = \mathbb{Z}$ ou $\mathcal{T} = \mathbb{N}$ diz-se que o processo é de tempo discreto e se $\mathcal{T} = \mathbb{R}$ ou, mais comummente, $\mathcal{T} = \mathbb{R}^+$ diz-se que o processo é de tempo contínuo.

Para caracterizar um processo estocástico deve-se especificar a distribuição de probabilidade conjunta de n variáveis aleatórias $(Y(t_1), \ldots, Y(t_n))$ para todos os inteiros n e quaisquer pontos $t_1, \ldots, t_n$. Contudo, esta forma de definir um processo estocástico é complexa e, na prática, é inexequível. Com isto, uma alternativa mais acessível para descrever um processo estocástico é através dos momentos do processo, em particular os primeiro e segundo momentos, designados por valor médio $\mu(t) = E[Y(t)]$ e função de autocovariância $\gamma(t_1, t_2) = E[(Y(t_1) - \mu(t_1))(Y(t_2) - \mu(t_2))]$, respetivamente. A variância $\sigma^2(t) = Var[Y(t)]$ é um caso particular da função de autocovariância (quando $t_1 = t_2$) mas, por si só não é suficiente para definir os segundos momentos de uma sequência de variáveis aleatórias (Chatfield, 2003).

A série de valores observados, que compõem a série temporal, é considerada apenas uma única realização (ou trajetória) de um processo estocástico, de entre todas as possíveis.

Geralmente, na análise de séries temporais é pretendida a inferência sobre um processo estocástico desconhecido tendo como informação disponível uma única realização observada (Cordeiro, 2011).

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

4.2.1 Processos Estocásticos Estacionários

Em termos gerais, os processos estacionários refletem a situação em que o sistema se apresenta num estado de equilíbrio estatístico em torno de um nível médio fixo, ou seja, tem propriedades probabilísticas que são estáveis ou invariantes ao longo do tempo (Murteira, 2000). As Definições 3 e 4, segundo Menezes (2019), apresentam os conceitos

de processo estocástico estritamente estacionário (ou fortemente estacionário) e processo estocástico de 2.ª ordem (ou fracamente estacionário).

Definição 3. *Um processo estocástico* $\{Y(t),\ t \in \mathcal{T}\}$ *diz-se estritamente estacionário (ou fortemente estacionário) se e só se a distribuição conjunta de* $(Y(t_1), \ldots, Y(t_n))$ *é igual à distribuição conjunta de* $(Y(t_1+\delta), \ldots, Y(t_n+\delta))$ *qualquer que seja o n-úplo* $(t_1, \ldots, t_n)$ *e para qualquer* δ*, ou seja,*

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

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

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

A estacionariedade no sentido estrito é uma propriedade demasiado exigente e, na maioria dos casos, de difícil verificação. Os processos estacionários de 2.ª ordem (ou fracamente estacionários) obedecem a uma propriedade mais fraca mas que, grosso modo, descreve o mesmo tipo de comportamento (Murteira, 2000).

Definição 4. *Um processo* $\{Y(t), t \in \mathcal{T}\}$ *diz-se estacionário de 2.ª ordem (ou fracamente estacionário) se e só se todos os momentos até à 2.ª ordem de* $(Y(t_1), \ldots, Y(t_n))$ *existem e são iguais aos momentos correspondentes até à 2.ª ordem de* $(Y(t_1+\delta), \ldots, Y(t_n+\delta))$*. Logo, num processo fracamente estacionário:*

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

Considere-se que os momentos até à 2.ª ordem existem e são finitos, logo se $Y(t)$ é estritamente estacionário e os seus momentos até à 2.ª ordem são finitos, então $Y(t)$ também é estacionário de 2.ª ordem. Acrescente-se que a recíproca pode não se verificar .

Na apresentação dos próximos conceitos, considerem-se apenas processos estacionários de 2.ª ordem, designados por processos estacionários, e Y_t a representar um processo estocástico, independentemente do tempo t.

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

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

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

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

1. $\gamma_0 = Cov[Y_t, Y_t] = Var[Y_t] = \sigma^2$;

2. $\gamma_k = \gamma_{-k}$, isto é, a função é par e dispensa a representação gráfica para $k < 0$;

3. $|\gamma_k| \leq \gamma_0$, como consequência da desigualdade de Cauchy-Schwarz dada por $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$;

4. A função γ_k é semidefinida positiva, isto é, para qualquer conjunto de números reais $\alpha_1, \ldots, \alpha_n$ e instantes de tempo $t_1, \ldots, t_n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \gamma(|t_i - t_j|) \geq 0. \tag{4.5}$$

Definição 6. *Para um processo estacionário, define-se a função de autocorrelação*

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

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

A representação gráfica de ρ_k em função de k, designada por correlograma, permite obter indicações essenciais sobre as características da série e auxilia na identificação do modelo que lhe é mais adequado. Geralmente, o aumento de k traduz-se no decrescimento de ρ_k e de γ_k.

À medida que a amplitude do intervalo (k) aumenta é de esperar que a capacidade de memória do processo seja limitada, e, portanto, que no momento $t + k$ se encontre pouco refletido o que se passou no momento t (Murteira, 2000). Com isto, espera-se que a correlação temporal diminua ($\rho_k \to 0$) com o aumento do desfasamento entre duas observações ($k \to +\infty$).

Pode interpretar-se ρ_k, de forma intuitiva, como uma medida da semelhança entre cada realização e a mesma realização deslocada k unidades de tempo (Murteira, 2000).

Analogamente à função de autocovariância, a função de autocorrelação ρ_k pode estar definida para $k \in \mathbb{R}$ ou para $k \in \mathbb{Z}$, consoante o processo for de tempo contínuo ou discreto, respetivamente. Esta função satisfaz as seguintes propriedades:

1. $\rho_0 = Corr[Y_t, Y_t] = 1$;

2. $\rho_k = \rho_{-k}$, isto é, a função é par e dispensa a representação gráfica para $k < 0$;

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

4. A função ρ_k é semidefinida positiva, isto é, para qualquer conjunto de números reais $\alpha_1, \ldots, \alpha_n$ e instantes de tempo $t_1, \ldots, t_n$,

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

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

– para a média μ usar $\bar{Y} = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} Y_t$;

– para a autocovariância γ_k usar $\hat{\gamma}_k = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$;

– para a autocorrelação ρ_k usar $\hat{\rho}_k = \dfrac{\hat{\gamma}_k}{\hat{\gamma}_0} = \dfrac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$.

Além do estudo da correlação de uma forma global, interessa também investigar a correlação parcial que existe entre Y_t e Y_{t+k} quando se fixam as variáveis intermédias $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}$, isto é, a correlação simples entre Y_t e Y_{t+k} depois de eliminar o efeito que as variáveis intermédias exercem sobre elas (Caiado, 2011). Sejam $E[Y_t] = 0$ e $Var[Y_t] = 1$ e considere-se a regressão linear múltipla de Y_{t+k} sobre $Y_{t+k-1}, \ldots, Y_t$,

$$Y_{t+k} = \phi_{k1}Y_{t+k-1} + \cdots + \phi_{kk}Y_t + \epsilon_{t+k}, \tag{4.6}$$

onde $\phi_{kj},\ j = 1, \ldots, k$, são os coeficientes do modelo de regressão linear que se considera ter erros gaussianos. O valor ϕ_{kk} é o coeficiente de correlação do modelo de regressão linear (4.6) onde $\{\epsilon_t,\ t \in \mathbb{Z}\}$ tem distribuição Normal de parâmetros $(0, \sigma^2)$ e ϵ_{t+k} é independente de $\{Y_{t+k-j},\ j \geq 1\}$. O coeficiente ϕ_{kk} exprime a variação em Y_{t+k} que acompanha em média uma variação unitária em Y_t quando $Y_{t+1}, \ldots, Y_{t+k-1}$ são constantes; tal variação pode interpretar-se como a correlação parcial entre Y_t e Y_{t+k}. Multiplicando, então, ambos os membros de 4.6 por $Y_{t+k-j},\ j = 1, \ldots, k$, calculando os valores esperados e dividindo por ρ_0 obtém-se um sistema constituído pelas equações

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{k-j}\ ,\ j = 1, 2, \ldots, k; \tag{4.7}$$

resolve-se o sistema em ordem a $\phi_{kj}, j = 1, 2, \ldots, k$, utilizando a regra de Cramer & Gabriel (1750), e obtém-se, assim, a função de autocorrelação parcial, ϕ_{kk}.

Uma definição alternativa para esta função é:

Definição 7. *O conjunto de autocorrelações parciais de desfasamento (lag) k é dado por* $\{\phi_{kk} : k = 1, 2, \ldots\}$ *onde*

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

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

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

Resolvendo o sistema constituído pelas equações em 4.7 ou seguindo a Definição 7, obtêm-se as seguintes propriedades:

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

Definição 8. *Um processo estocástico* $\{\epsilon_t, t \in \mathbb{Z}\}$ *diz-se um processo puramente aleatório ou processo de ruído branco quando é formado por uma sucessão de variáveis aleatórias não correlacionadas e identicamente distribuídas, de média e variância constantes, ou seja, um processo estocástico diz-se um ruído branco se e só se satisfaz as seguintes condições:*

1. $E[\epsilon_t] = \mu_\epsilon$ *(usualmente* $\mu_\epsilon = 0$*);*

2. $Var[\epsilon_t] = \sigma_\epsilon^2$*;*

3. $Cov(\epsilon_t, \epsilon_{t+k}) = \gamma_k = 0, \quad k = \pm 1, \pm 2, \ldots.$

Se, além disso, as variáveis aleatórias seguem uma distribuição Normal ($\epsilon_t \sim N(\mu_\epsilon, \sigma_\epsilon^2)$), então o processo é designado de ruído branco gaussiano. Um ruído branco é, então, um processo estacionário cujas funções de autocorrelação (FAC) e autocorrelação parcial (FACP) são nulas para todo o $k \neq 0$. A Figura 4.1 representa uma trajetória de um processo de ruído branco e as respetivas FAC e FACP empíricas, a título de exemplo.

Segundo Caiado (2011), o ruído branco, apesar de ser difícil de observar em séries reais, executa um papel fundamental na construção de modelos probabilísticos ou estocásticos. Com isto, acrescenta-se que um bom modelo de previsão deve ser aquele que produz erros de previsão com comportamento análogo a um ruído branco, isto porque um ruído branco é imprevisível.

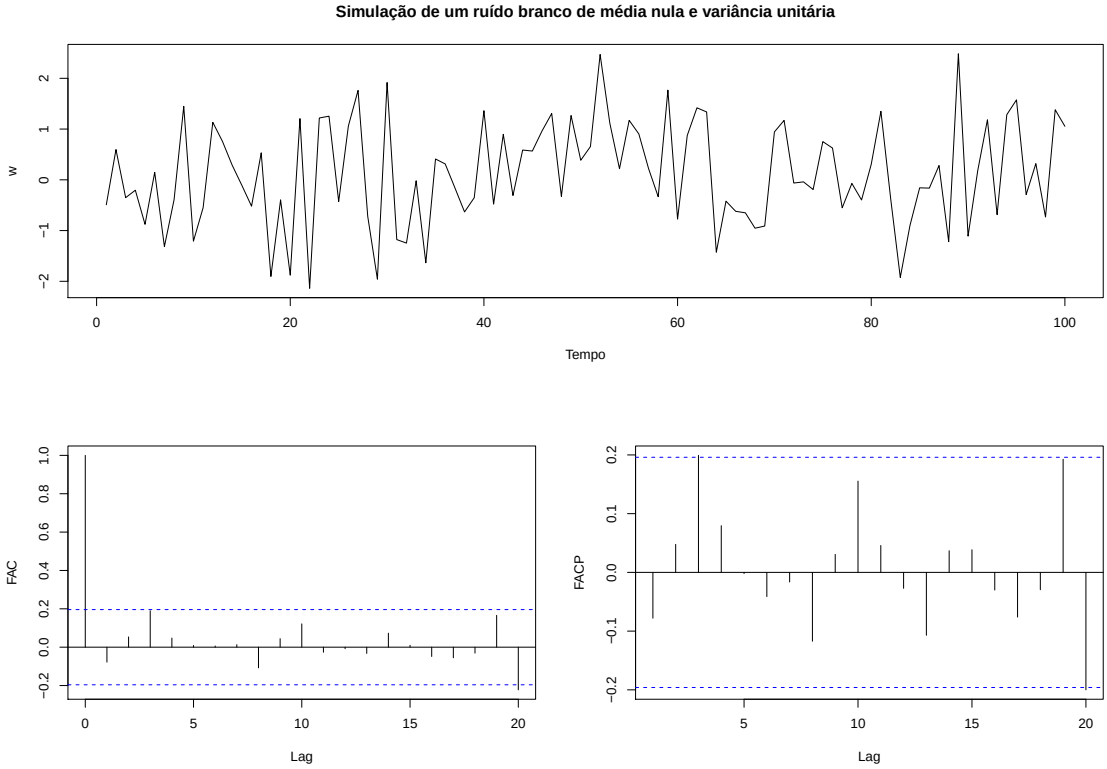


Figura 4.1: Representação da simulação de um ruído branco e respetivas FAC e FACP empíricas.

4.2.2 Processos Estocásticos não Estacionários

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

Muitas séries temporais, nomeadamente as associadas a fenómenos ambientais ou económicos, são não estacionárias. Um processo pode ser não estacionário por a média e/ou a variância serem funções do tempo e não constantes. Uma série estacionária em média não é necessariamente estacionária em variância. De forma a ultrapassar este problema pode recorrer-se a transformações que estabilizam a média e/ou a variância convertendo, assim, uma série temporal não estacionária numa série estacionária. No caso de se estar perante uma série não estacionária em média nem em variância deve proceder-se, em primeiro lugar, à estabilização da variância e só depois da média (Murteira, 2000; Caiado, 2011).

Em diversas situações, existem procedimentos que têm como objetivo a remoção da tendência e da sazonalidade a uma série temporal, permitindo, desta forma, que se atinja a estacionariedade. Estes métodos seguem os modelos de decomposição anteriormente descritos e consistem, naturalmente, na estimação das componentes tendência (T_t) e sazonalidade (S_t) através de funções determinísticas (ou outras abordagens), de modo que

a série após "remoção" dessas componentes passe a ser bem modelada por um processo estacionário (Alpuim, 1998). No entanto, em muitas séries temporais é possível realizar outro tipo de transformações, que permitem transformar séries não estacionárias em estacionárias. Uma das formas que permite a estabilização da média consiste no uso de processos de diferenciação, que resultam da aplicação do operador diferença ∇, definido como $\nabla Y_t = Y_t - Y_{t-1}$, à série temporal não estacionária. Assim, se uma série, Y_t, for não estacionária, pode pensar-se em transformá-la numa série estacionária, aplicando uma diferenciação de primeira ordem,

$$\nabla Y_t = Y_t - Y_{t-1} \ , \quad t = 2, 3, \ldots, n. \tag{4.8}$$

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

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = \nabla(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2} \ , \quad t = 3, 4, \ldots, n. \tag{4.9}$$

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

$$\nabla^d Y_t = \nabla(\nabla^{d-1} Y_t) \ , \quad t = d + 1, \ldots, n. \tag{4.10}$$

A diferenciação inapropriada de uma série já estacionária é indesejável e deve evitar-se. De facto, o objetivo é determinar a série estacionária obtida pela menor diferenciação, uma vez que a variância aumenta conforme a diferenciação efetuada. De uma forma geral, se a série transformada $\nabla^{d_0} Y_t$ é estacionária, então, para qualquer $d > d_0$, a série $\nabla^d Y_t$ é também estacionária, mas tem maior variância. Conclui-se, portanto, que se deve evitar a sobrediferenciação para não introduzir variação indesejada na série transformada.

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

Com o objetivo de estabilização da variância de uma série não estacionária, pode utilizar-se um método de transformação paramétrica, conhecido como transformação de Box-Cox, baseado na seguinte expressão

$$Z_t = T(Y_t) = \begin{cases} \frac{Y_t^{\lambda} - 1}{\lambda} \ , & \lambda \neq 0 \\ \log Y_t \ , & \lambda = 0 \end{cases} \ , \tag{4.11}$$

onde os valores de λ estão no intervalo $[-1; 1]$. As transformações mais utilizadas encontram-se sumariadas na Tabela 4.1.

Tabela 4.1: Transformações usuais de Box-Cox.

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

Note-se que algumas destas transformações (nomeadamente a do logaritmo) apenas estão definidas para séries de valores positivos.

Como foi verificado na Secção 4.1, a transformação logarítmica permite converter o efeito sazonal multiplicativo em aditivo sendo, por isso, uma das transformações mais usadas dentro da família Box-Cox. Esta transformação permite, também, estabilizar as diferenças entre os valores da série temporal, ela também pode ser utilizada para atenuar o efeito de possíveis *outliers*. É importante realçar que, sempre que se aplica alguma transformação aos dados, todas as previsões geradas pelo modelo selecionado estarão nas unidades transformadas. Uma vez ajustado o modelo e estimados os parâmetros, devem, então, ser revertidas as transformações de modo a obter previsões nas unidades originais (Jebb & Tay, 2015).

Análise de estacionariedade

Com o intuito de analisar a estacionariedade de uma série, de uma forma elementar, pode-se representar graficamente os dados ao longo do tempo. Esta análise, entretanto, é subjetiva e, apesar de útil, deve ser confirmada através de testes estatísticos formais. Existem vários testes para realizar este estudo, baseados, na sua maioria, em encontrar uma raiz unitária. Conforme a literatura, deve utilizar-se mais do que um teste de raiz unitária de modo a avaliar a estacionariedade da série. Alguns dos testes mais utilizados são os testes de Dickey-Fuller Aumentado (*Argumented Dickey Fuller*, ADF) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS).

Conforme Dickey & Fuller (1979) e Said & Dickey (1984), o teste ADF é realizado cuja hipótese a testar é a presença de uma raiz unitária, ou seja, a não estacionariedade, e, no caso de esta não ser rejeitada, são fornecidas informações sobre o número de diferenciações necessárias para atingir a estacionariedade. Já o teste KPSS, segundo Kwiatkowski & Phillips (1992), é realizado sob a hipótese nula de estacionariedade.

Teste de Dickey-Fuller Aumentado

Muitas séries apresentam uma estrutura mais complexa do que a captada pelo modelo 4.12. Para lidar com esses casos, surge uma variante do teste de Dickey-Fuller (DF) capaz de integrar modelos mais complexos. Esta inovação deve-se a Said & Dickey (1984) e,

desde então, este tem sido um dos testes mais usados no estudo da estacionariedade de séries temporais.

Desta forma, considere-se um processo definido por

$$Y_t = \phi Y_{t-1} + \epsilon_t, \quad -1 \leq \phi \leq 1, \tag{4.12}$$

onde ϵ_t é um ruído branco. Este processo é estacionário se $|\phi| < 1$ (trata-se de um processo autorregressivo de ordem 1, como se verá na Secção 5.1.1). No entanto, quando $\phi = 1$ este processo é um passeio aleatório o que, como foi visto anteriormente, equivale a fazer uma diferenciação de 1.ª ordem. Assim, se $\phi = 1$, pode dizer-se que a série é não estacionária. O processo descrito pela equação 4.12 pode ser escrito na forma das diferenças, ou seja,

$$\begin{aligned} Y_t = \phi Y_{t-1} + \epsilon_t &\Leftrightarrow Y_t - Y_{t-1} = \phi Y_{t-1} - Y_{t-1} + \epsilon_t \\ &\Leftrightarrow \nabla Y_t = (\phi - 1) Y_{t-1} + \epsilon_t \\ &\Leftrightarrow \nabla Y_t = \delta Y_{t-1} + \epsilon_t, \end{aligned} \tag{4.13}$$

onde $\delta = \phi - 1$ e ϵ_t é um processo estacionário.

Com isto e, neste caso, um processo mais complexo que o apresentado em 4.12 é definido por

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t, \tag{4.14}$$

Assim, repetindo o processo apresentado 4.13, tem-se

$$\nabla Y_t = \delta Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \nabla Y_{t-j} + \epsilon_t, \tag{4.15}$$

onde $\delta = \sum_{i=1}^{p} \phi_i - 1$, $\gamma_j = -\sum_{i=j+1}^{p} \phi_i$, $\nabla Y_{t-j} = Y_{t-j} - Y_{t-j-1}$ e ϵ_t é um ruído branco. Esta decomposição separa o modelo 4.14 em dois termos: Y_{t-1} e as $p-1$ primeiras diferenças. No caso em que Y_t é I(1) (passeio aleatório), esta separação envolve uma componente I(1) e $p-1$ componentes I(0) (estacionárias). Na terminologia comum, diz-se que o modelo original foi aumentado por $p-1$ componentes de primeiras diferenças, dando origem à designação ADF $(p-1)$.

Para o modelo 4.15, a existência de uma raiz unitária é garantida se $\sum_{i=1}^{p} \phi_i = 1$, isto é, se $\delta = 0$. Assim, as hipóteses a testar são, também neste caso,

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

e a não rejeição da hipótese nula implica a não estacionariedade de Y_t. Como este teste é unilateral à esquerda, a hipótese nula é rejeitada a um nível de significância α se a estatística de teste for inferior ou igual ao quantil $(1-\alpha)100\%$ da distribuição correspondente (valor crítico).

Além da equação 4.15, são propostas duas outras equações

$$\nabla Y_t = a_0 + \delta Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \nabla Y_{t-j} + \epsilon_t, \tag{4.16}$$

$$\nabla Y_t = a_0 + a_1 t + \delta Y_{t-1} + \sum_{j=1}^{p-1} \gamma_j \nabla Y_{t-j} + \epsilon_t, \tag{4.17}$$

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

Um dos principais problemas do teste ADF é decidir qual o número de termos a incluir na equação a ser testada, ou seja, o valor de p. Para este estudo é utilizada a estratégia baseada na regra proposta por Ng & Perron (1995) que consiste, numa primeira etapa, em definir um limite máximo para p, p_{max}. Numa segunda etapa calcula-se o teste ADF considerando $p = p_{max}$ e, se o valor absoluto da estatística t para testar a significância da diferença de ordem p for maior do que 1,6, define-se $p = p_{max}$ e prossegue-se com o teste, caso contrário reduz-se p em uma unidade e repete-se o processo. Segundo Schwert (2002) pode determinar-se p_{max} por

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

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

Teste de KPSS

O teste de KPSS permite avaliar a estacionariedade de um processo. No entanto, neste caso, as hipóteses a testar são diferentes do teste ADF, ou seja,

$$H_0 : \text{O processo é estacionário} \quad vs \quad H_1 : \text{O processo é não estacionário.}$$

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

$$Y_t = T_t + \mu_t + u_t, \tag{4.18}$$

$$\mu_t = \mu_{t-1} + \epsilon_t, \tag{4.19}$$

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

Capítulo 5

Métodos de Previsão em Séries Temporais

Segundo Cordeiro (2011), a distinção entre modelo, representação matemática da estrutura estocástica de uma série temporal através de uma equação ou sistema de equações, e método, procedimento para calcular previsões, nem sempre foi clara. Sabe-se que um modelo estatístico determina um processo gerador dos dados que, tem como finalidade, a obtenção de toda a distribuição de probabilidade para um momento futuro. Para além das previsões pontuais para um determinado horizonte temporal, um modelo também permite o cálculo de previsões intervalares (intervalos de previsão), para um nível de confiança associado.

Um método de previsão é um procedimento para calcular previsões a partir de valores presentes e passados. Como tal, pode ser simplesmente um algoritmo e não depender de um modelo de probabilidade subjacente ou, alternativamente, surgir da identificação de um modelo específico para os dados fornecidos e da localização de previsões condicionadas a esse modelo (Chatfield, 2000). A escolha do método depende de uma variedade de considerações, tais como o objetivo do cálculo das previsões, o tipo de série temporal em estudo e respetivas componentes, a dimensão da série temporal, o horizonte de previsão, o conhecimento e experiência do analista e, também, a disponibilidade dos programas informáticos. Tendo conhecimento sobre a diversidade de métodos de previsão que se podem aplicar a uma série temporal, cada um com as suas capacidades e limitações, deve escolher-se o método que pareça mais adequado. Na presente dissertação é estudada a abordagem clássica de Box-Jenkins.

5.1 A Metodologia Box-Jenkins

Box & Jenkins (1970) introduziram uma abordagem prática e sistemática para a construção de modelos SARIMA (*Seasonal Autoregressive Integrated Moving Average*), baseada nos trabalhos de Yule (1926) e Wold (1938), conhecida como metodologia Box-Jenkins. Esta metodologia trata-se de um processo de modelação iterativo dividido em três fases: identificação do modelo, estimação dos parâmetros e análise de diagnóstico (ou validação do modelo). Este processo é tipicamente repetido várias vezes até que um modelo satisfatório seja selecionado. A ideia que sustenta a identificação do modelo é que, se uma série temporal é gerada a partir de um processo SARIMA, então deve ter algumas propriedades teóricas de autocorrelação. Desta forma, ao comparar os padrões empíricos de autocorrelação com os teóricos, é frequentemente possível identificar um ou vários potenciais modelos para a série temporal a estudar. Box & Jenkins (1970) propuseram, então, usar a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) como ferramentas básicas para identificar as ordens do modelo SARIMA (Zhang, 2003).

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

Nos modelos SARIMA classificam-se em : modelo autorregressivo (AR), que considera que o comportamento da série pode ser explicado através do seu passado; o modelo de médias móveis (MA), que explica a série temporal através de uma sucessão de choques aleatórios; e ainda o modelo autorregressivo e de médias móveis (ARMA), que, tal como o nome indica, se trata de uma combinação dos dois modelos anteriores. Estes modelos são úteis para séries estacionárias, revelando-se, então, insuficientes para modelar casos de não estacionariedade. Nesses casos, deve optar-se pelos modelos integrados (ARIMA) ou, para séries que apresentam sazonalidade, pelos equivalentes sazonais (SARIMA).

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

Contudo, a fase correspondente à identificação do modelo é bastante subjetiva e complexa, resultando, várias vezes, em modelos distintos para uma mesma análise. De forma a contornar esta situação, os *softwares* já possuem algoritmos automatizados de seleção, que permitem não só que a metodologia possa ser utilizada por leigos, mas também que os resultados obtidos por diferentes analistas sejam os mesmos. Uma das, e maiores, vanta-

gens destes modelos é que estes têm em consideração uma das principais caraterísticas dos dados de séries temporais: a dependência temporal (autocorrelação). Com isto, afirma-se que os modelos SARIMA são apropriados quando se pode assumir que existe algum tipo de relação entre o passado e o futuro, sendo, no entanto, isso que os torna pouco recomendados para previsões a longo prazo e/ou previsões de séries com mudanças bruscas de comportamento.

5.1.1 Modelos de Processos Estacionários

Considera-se que uma série estacionária fica completamente definida pelas suas funções média, variância e de autocorrelação. Fazendo uso desta característica, pretende-se, com a metodologia Box-Jenkins, identificar um modelo com base no comportamento da função de autocorrelação empírica. Os processos ARMA são considerados como um grupo bastante diversificado e de grande fiabilidade na modelação de inúmeras séries temporais estacionárias. Contudo, para a modelação de séries que apresentem oscilações bruscas ao longo do tempo, este tipo de processos é insuficiente. No entanto, os processos ARMA têm especial importância na modelação de séries não estacionárias, uma vez que estas são facilmente convertidas em estacionárias através de transformações adequadas.

Processo Autorregressivo (AR)

Os processos autorregressivos, que pertencem à classe dos modelos mais utilizados no estudo de séries temporais estacionárias, baseiam-se no pressuposto de que a observação da variável no instante t se relaciona, de forma linear, com as observações nos instantes anteriores. Assim, o processo Y_t diz-se um processo autorregressivo de ordem p, AR(p), quando satisfaz a equação

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_{t-p} Y_{t-p} + \epsilon_t, \tag{5.1}$$

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

Alternativamente, a representação de um processo AR(p) pode ser feita através do operador atraso B^k, que se define como sendo $B^k Y_t = Y_{t-k}$. Com efeito, a equação 5.1 pode ser reescrita como

$$\Phi_p(B) Y_t = \epsilon_t, \tag{5.2}$$

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

modo

$$\Phi_p(B) = \prod_{i=1}^{p}(1 - G_i B). \tag{5.3}$$

Para que o processo seja estacionário é condição necessária e suficiente que as raízes da equação característica sejam todas de módulo maior do que a unidade, ou, de forma equivalente, que $|G_i| < 1$, para $i = 1, 2, \ldots, p$. Qualquer processo autorregressivo que seja estacionário é também invertível, o que, em termos práticos, significa que a dependência do passado se vai atenuando à medida que o passado se torna mais remoto.

Portanto, se o processo Y_t é um processo AR(p), então a sua função de autocorrelação parcial, ϕ_{kk}, é igual a zero para todo o $k > p$. Assim, a FACP de um processo AR(p) apresenta, graficamente, uma queda brusca para zero a partir do *lag* $p + 1$, enquanto que a respetiva FAC tem um decaimento exponencial ou sinusoidal amortecido para zero.

Na Figura 5.1 encontra-se representado um processo autorregressivo de ordem 1, AR(1), e as respetivas FAC e FACP empíricas.

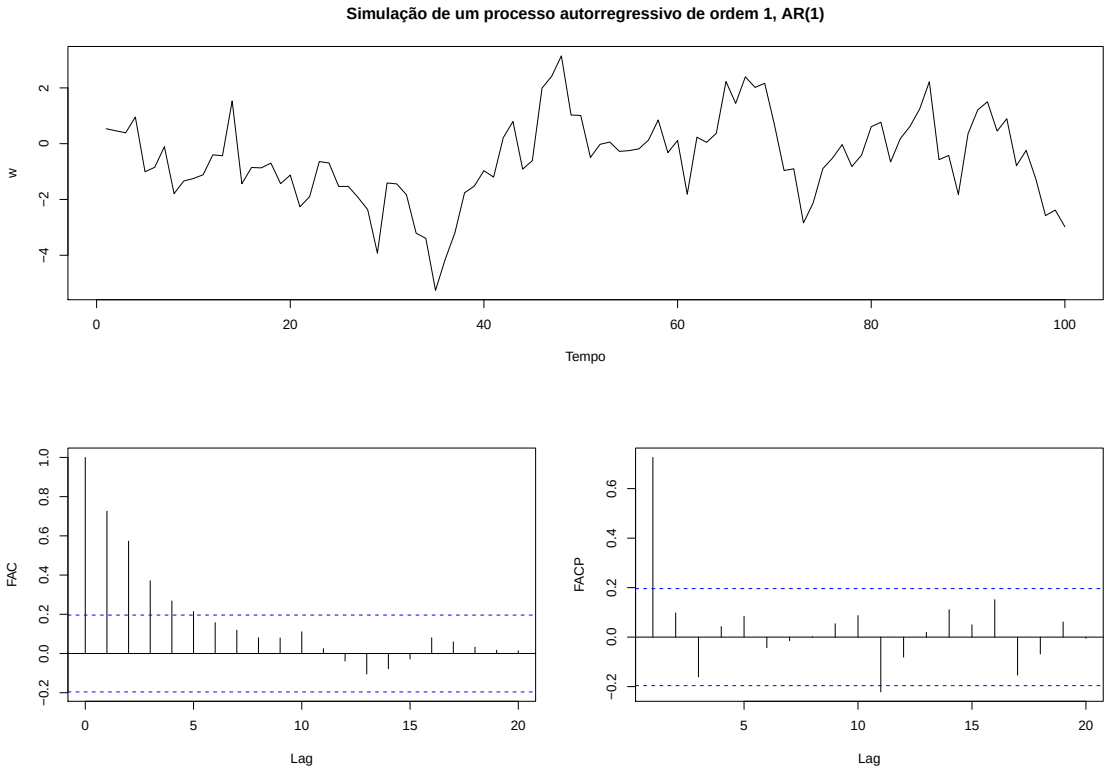


Figura 5.1: Simulação de um processo autorregressivo e respetivas FAC e FACP empíricas.

Processo de Médias Móveis (MA)

Diz-se que o processo Y_t é um processo de médias móveis de ordem q, MA(q), quando assume a expressão

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{5.4}$$

ou

$$Y_t = \Theta_q(B)\epsilon_t, \tag{5.5}$$

onde ϵ_t é um ruído branco de média nula e $\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ é o polinómio de médias móveis de ordem q. Pretende-se, através destes processos, exprimir Y_t em termos de um processo mais simples, como é o ruído branco. Assim, um processo de médias móveis de ordem q define-se, em cada instante t, como a média ponderada das $q + 1$ observações de um processo de ruído branco. Desta forma, graças à estacionariedade intrínseca ao ruído branco, os processos de médias móveis são sempre estacionários. Adicionalmente, um processo de médias móveis é invertível se puder ser escrito como um processo autorregressivo estacionário de ordem infinita. Para garantir a invertibilidade do processo, basta que, à semelhança do que acontece no caso da estacionariedade de processos autorregressivos, as raízes da equação característica $\Theta_q(B) = 0$ se encontrem todas fora do círculo unitário, isto é, sejam, em módulo, todas superiores a 1 (Metcalfe & Cowpertwait, 2009).

Na Figura 5.2 encontra-se representado um processo de médias móveis de ordem 1, MA(1), e as respetivas FAC e FACP empíricas.

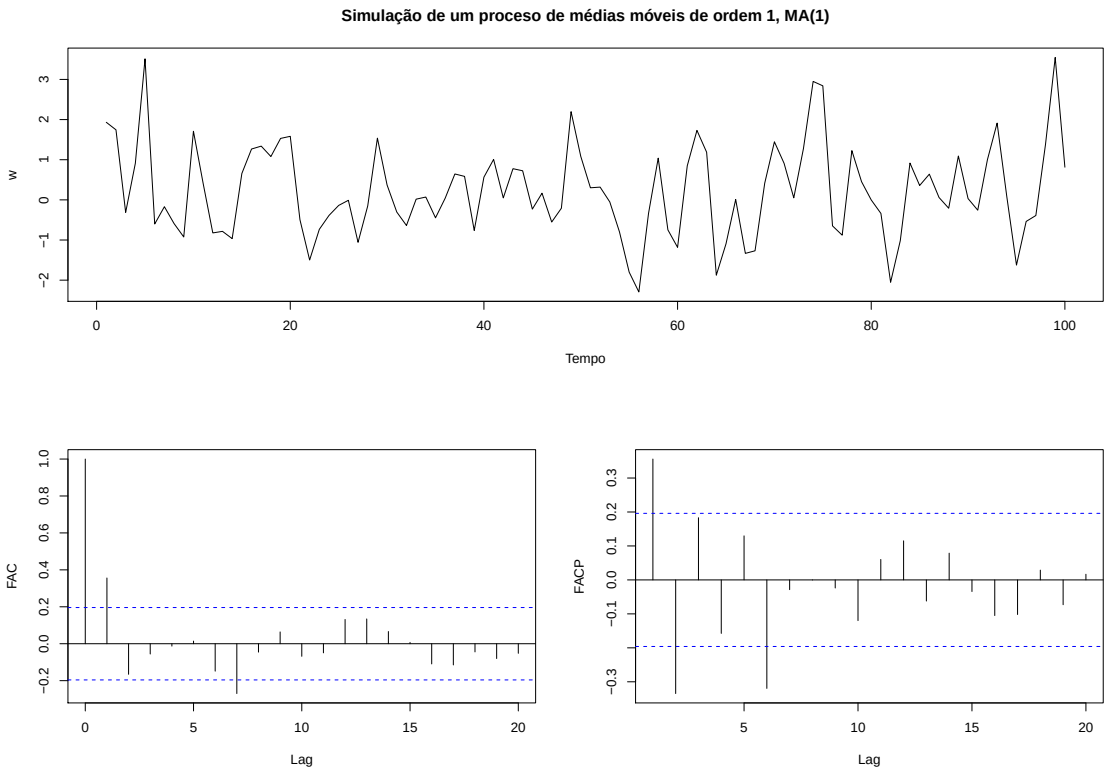


Figura 5.2: Simulação de um processo de médias móveis e respetivas FAC e FACP empíricas.

Se o processo Y_t é um processo MA(q), então a sua função de autocorrelação, ρ_k, é igual a zero para todo o $k > q$, e, então, a FAC de um processo MA(q) apresenta, graficamente,

uma queda brusca para zero a partir do *lag* $q+1$. No que respeita à FACP, esta expõe um decaimento exponencial ou sinusoidal amortecido para zero tendo, portanto, a mesma estrutura que a FAC de um processo AR(q).

Processo Autorregressivo e de Médias Móveis (ARMA)

Os processos estacionários e invertíveis podem ser representados tanto na forma autorregressiva quer na forma de médias móveis. No entanto, é possível que qualquer um destes processos tenha uma representação com um número excessivo de parâmetros, o que pode conduzir a uma perda de eficiência na sua estimação (Caiado, 2011).

Caso se afirme, pode construir-se um modelo mais parcimonioso que inclua tanto termos autorregressivos como de médias móveis. Este modelo designa-se de processo misto autorregressivo e de médias móveis de ordens p e q e representa-se por ARMA(p,q). Então o processo Y_t diz-se um processo autorregressivo e de médias móveis de ordens p e q, ARMA(p,q), se satisfaz a equação

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{5.6}$$

ou a equação

$$\Phi_p(B) Y_t = \Theta_q(B) \epsilon_t, \tag{5.7}$$

onde ϵ_t é um ruído branco de média nula, independente de Y_{t-k} para todo o $k \geq 1$, $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ e $\Theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ são os polinómios autorregressivo e de médias móveis de ordens p e q, respetivamente. A FAC e a FACP de um processo ARMA(p,q) resultam da combinação das respetivas funções dos processos AR(p) e MA(q).

Recorde-se que a FAC de um processo MA(q) é insignificante a partir do *lag* $q+1$, o mesmo acontecendo para a FACP de um processo AR(p) depois do *lag* p. Dado que o processo ARMA(p,q) é uma combinação dos processos AR(p) e MA(q), a estacionariedade e a invertibilidade do processo ficam garantidas se as raízes das equações características $\Phi_p(B) = 0$ e $\Theta_q(B) = 0$ são, em módulo, maiores do que a unidade. De facto, estes processos generalizam os processos anteriormente mencionados e, por exemplo, um processo ARMA$(p,0)$ é equivalente a um processo AR(p), o mesmo acontecendo com um ARMA$(0,q)$ relativamente a um MA(q).

Na Figura 5.3 encontra-se representado um processo autorregressivo e de médias móveis, ARMA$(1,2)$, e as respetivas FAC e FACP empíricas. Este processo é estacionário e invertível.

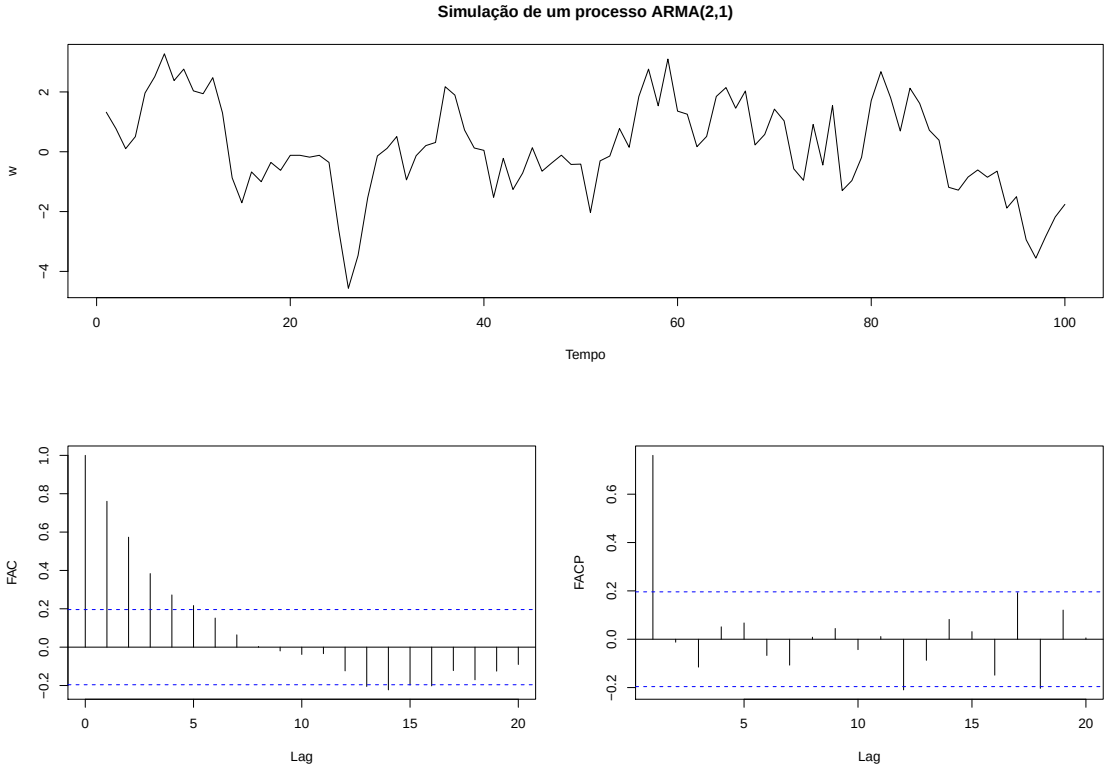


Figura 5.3: Simulação de um processo autorregressivo e de médias móveis, ARMA$(2, 2)$ e respetivas FAC e FACP empíricas.

5.1.2 Modelos de Processos Não Estacionários

Numa perspetiva prática, a maioria das séries temporais é não estacionária. Quando tal sucede, é necessária a remoção, nos dados, as fontes de variação não estacionárias (e.g., tendência, sazonalidade), de forma a possibilitar o ajustamento de um modelo estacionário. Como se verificou anteriormente, se a série temporal observada for não estacionária na média, pode aplicar-se uma (ou várias) diferenciação (regular) à mesma. Portanto, se se substituir Y_t por $\nabla^d Y_t$ na equação 5.7, é obtido um modelo capaz de descrever séries não estacionárias (modelo ARIMA). Este tipo de modelo é designado de modelo "integrado", uma vez que o modelo estacionário que é ajustado aos dados diferenciados deve ser somado ou "integrado" de forma a devolver um modelo para os dados não estacionários. Acrescente-se que estes modelos podem, à semelhança dos modelos ARMA, ser generalizados para incluir termos sazonais, dando origem aos modelos SARIMA.

Processo Autorregressivo Integrado de Médias Móveis (ARIMA)

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

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d Y_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\epsilon_t \tag{5.8}$$

ou

$$\Phi_p(B)\nabla^d Y_t = \Theta_q(B)\epsilon_t, \tag{5.9}$$

onde $\nabla^d Y_t = (1-B)^d Y_t$, com $d \geq 1$, é a série estacionária depois de diferenciada d vezes, $\phi_1, \phi_2, \ldots, \phi_p$ são os parâmetros autorregressivos, $\theta_1, \theta_2, \ldots, \theta_q$ são os parâmetros de médias móveis e $\Phi_p(B)$ e $\Theta_q(B)$ são os polinómios autorregressivo e de médias móveis regulares. Como se trata de um processo não estacionário, um processo deste tipo apresenta uma FAC com coeficientes positivos e decaimento muito lento para zero, pelo que a necessidade de uma diferenciação (regular) é facilmente identificável. Na Figura 5.4 encontra-se representado um processo autorregressivo e de médias móveis integrado, ARIMA$(2,1,1)$, e as respetivas FAC e FACP empíricas.

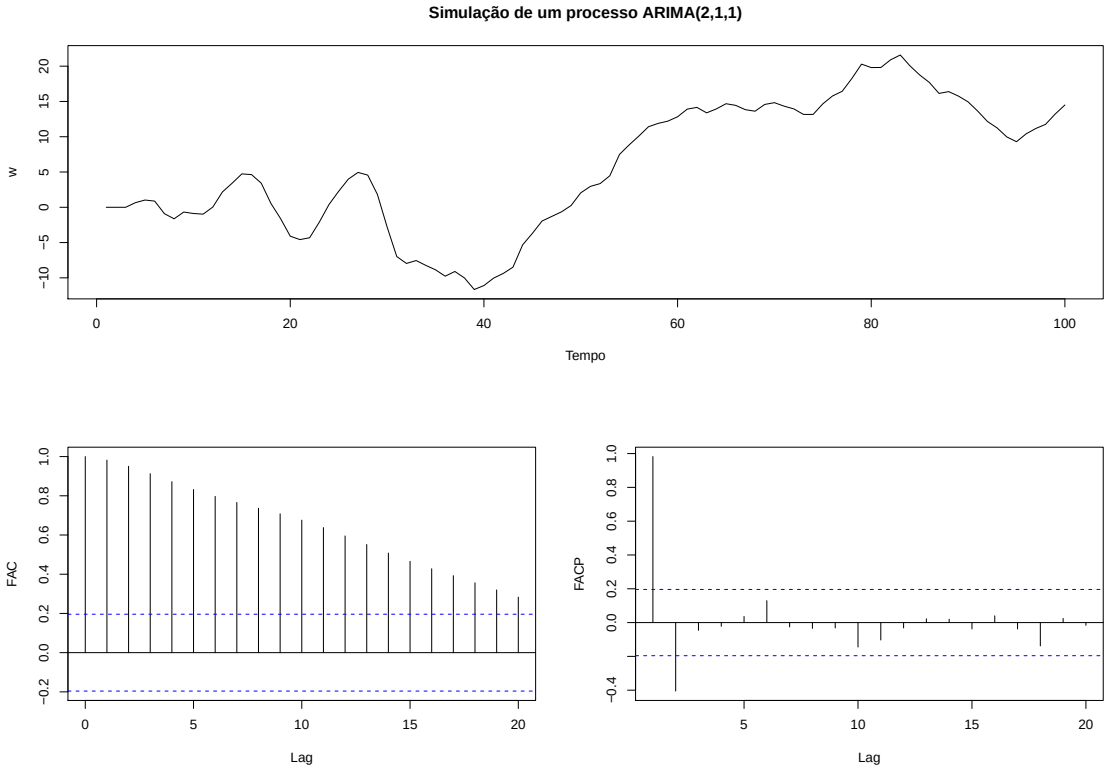


Figura 5.4: Simulação de um processo autorregressivo e de médias móveis integrado, ARIMA$(2,1,1)$ e respetivas FAC e FACP empíricas.

Processo Autorregressivo Integrado de Médias Móveis Sazonal (SARIMA)

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

e se representa por SARIMA$(p,d,q)(P,D,Q)_s$. Uma diferenciação sazonal consiste na diferença entre a observação no instante t e a observação que ocorre s momentos antes, ou seja, a observação no instante $t-s$. Desta forma, quando uma série apresenta um comportamento periódico, pode-se aplicar uma diferenciação sazonal da seguinte forma

$$\nabla_s Y_t = Y_t - Y_{t-s} = (1-B^s)Y_t. \tag{5.10}$$

Como consequência, a série resultante desta diferenciação corresponde, então, à mudança entre observações separadas por períodos de tempo s. Por exemplo, para uma série semanal, com $s \approx 52,18$, a série resultante de uma diferenciação sazonal representa a mudança que ocorre de ano para ano. De uma forma análoga ao que acontece para a diferenciação (regular) introduzida na Secção 4.2.2, a diferenciação sazonal pode ser aplicada a uma série D vezes, dando origem ao operador de diferenciação sazonal de ordem D, para qualquer inteiro $D \geq 1$, que se define por

$$\nabla_s^D Y_t = (1-B^s)^D Y_t. \tag{5.11}$$

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

$$\Phi_p(B)N_P(B^s)\nabla^d\nabla_s^D Y_t = \Theta_q(B)H_Q(B^s)\epsilon_t, \tag{5.12}$$

em que $\Phi_p(B), N_P(B^s), \Theta_q(B)$ e $H_Q(B^s)$ são os polinómios já referidos, d e D são as ordens de diferenciação das partes regular e sazonal, respetivamente. Geralmente, a necessidade de uma diferenciação sazonal pode ser indicada quando a FAC de um processo decai lentamente nos *lags* múltiplos de s e é insignificante nos restantes (Shumway & Stoffer, 2017). Na Figura 5.5 encontra-se representado um processo SARIMA$(2,1,1)(1,1,1)_{12}$ e as respetivas FAC e FACP empíricas. Segundo Caiado (2011), na maioria das aplicações práticas, os valores de p, q, P e Q são praticamente sempre inferiores ou iguais a 2, enquanto que os valores de d e D usualmente apresentam valores inteiros iguais a 0 ou 1.

Nos *lags* baixos o comportamento da FAC deve assemelhar-se ao da sua parte regular e nos *lags* sazonais deve observar-se apenas o efeito da parte sazonal. Em torno destes últimos pode observar-se a interação entre as partes regular e sazonal, que se manifesta na repetição em ambos os lados de cada *lag* sazonal da função FAC da parte regular. Também a FACP revela influência tanto da parte regular como da parte sazonal. Nos primeiros *lags*, o comportamento é idêntico ao da parte regular e nos *lags* sazonais a FACP reflete a parte sazonal. À direita de cada *lag* sazonal deve observar-se a FACP da parte regular, enquanto que à esquerda se deve notar o efeito da FAC da parte regular.

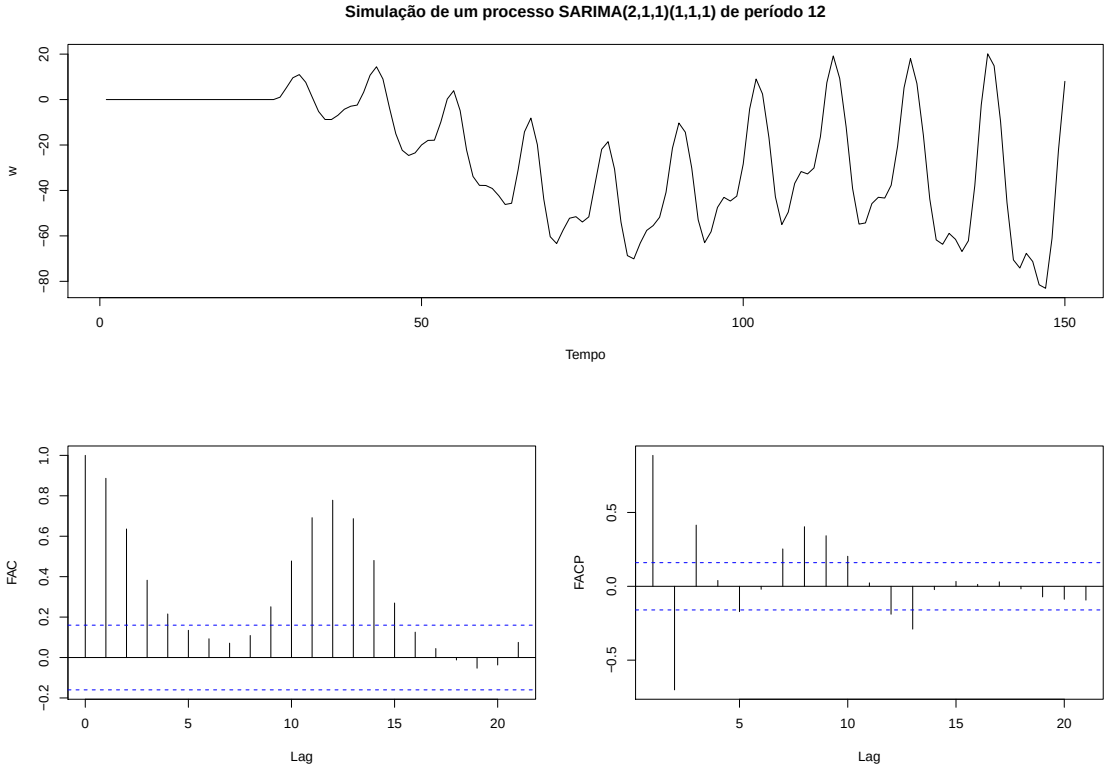


Figura 5.5: Simulação de um processo autorregressivo e de médias móveis integrado sazonal, SARIMA$(2,1,1)(1,1,1)_{12}$ e respetivas FAC e FACP empíricas.

5.1.3 Etapas da Metodologia Box-Jenkins

Nesta Secção apresenta-se com maior detalhe as etapas da metodologia Box-Jenkins: identificação, estimação e diagnóstico. A primeira etapa na modelação de uma série temporal consiste na identificação de um modelo SARIMA$(p,d,q)(P,D,Q)_s$ que descreva a relação existente entre as suas observações. Esta etapa compreende três importantes passos na tentativa de identificação dos valores dos inteiros d, s, D, p, q, P e Q:

1. Representação gráfica da série e respetiva estacionarização
   O estudo de uma série temporal deve iniciar-se pela análise detalhada da sua representação gráfica, com vista à identificação da existência ou não de fontes de não estacionariedade. Caso a série em estudo não seja estacionária, deve-se proceder à sua estacionarização através de uma transformação adequada: com vista à estabilização da variância recorrer a transformações Box-Cox; para a eliminação da tendência utiliza-se a diferenciação regular e para a eliminação de movimentos periódicos, a diferenciação sazonal. É importante salientar que, caso a estabilização da variância seja necessária, esta deve ser efetuada antes de qualquer outra transformação.

2. Estimação das FAC e FACP da série original
   Analisa-se o comportamento das FAC e FACP da série original, uma vez que as

conclusões retiradas podem ser úteis para complementar à informação obtida através da representação gráfica (passo 1), nomeadamente no que diz respeito à utilização de diferenciações. Por exemplo, um decaimento lento para zero na FAC de uma série temporal pode indicar a necessidade da aplicação de uma diferenciação não sazonal.

3. Estimação das FAC e FACP da série estacionária e identificação dos inteiros p, q, P e Q
   Identificadas as ordens de diferenciação, d e D, e o período, s, a escolha do modelo que descreve a série temporal só é considerada concluída quando são determinados os inteiros p, q, P e Q. Para identificar esses inteiros é efetuada a comparação do comportamento das FAC e FACP empíricas com o das FAC e FACP teórica.

Identificados os modelos candidatos a descrever a série em estudo, segue-se a etapa de estimação dos seus parâmetros. Nesta fase, é imprescindível o auxílio de um *software* estatístico adequado, uma vez que a estimação dos parâmetros requer a aplicação de um conjunto de métodos numéricos e de cálculos computacionais com alguma complexidade. Os dois principais métodos de estimação dos parâmetros do modelo SARIMA são o método da máxima verosimilhança e o método dos mínimos quadrados. O método da máxima verosimilhança fundamenta-se na ideia de determinar os valores dos parâmetros que tornam mais verosímil a ocorrência de um conjunto de observações idênticas aquelas de que efetivamente se dispõe. Segundo Box & Jenkins (2016), este método obtém estimativas dos parâmetros através de um processo iterativo em que se maximiza a função de verosimilhança dos estimadores. .

O método dos mínimos quadrados é, hipoteticamente, o método estatístico mais utilizado na estimação de modelos. A Tabela 5.1 apresenta os comportamentos das FAC e FACP dos modelos de previsão que, segundo Shumway & Stoffer (2017), permitem identificar os parâmetros p e q da ordem regular.

Tabela 5.1: Padrões teóricos das FAC e FACP dos modelos de previsão em séries temporais.

Modelo	**FAC**	**FACP**
AR(p)	Decaimento exponencial ou sinusoidal amortecido para zero	Queda brusca para zero a partir do *lag* $p+1$
MA(q)	Queda brusca para zero a partir do *lag* $q+1$	Decaimento exponencial ou sinusoidal amortecido para zero
ARMA(p, q)	Decaimento exponencial ou sinusoidal amortecido para zero	Decaimento exponencial ou sinusoidal amortecido para zero

Após a identificação do modelo SARIMA e a estimação dos respetivos parâmetros, é necessário verificar a adequação do modelo. A fase de diagnóstico engloba duas etapas imprescindíveis: a avaliação da qualidade das estimativas obtidas e a avaliação da qualidade do ajustamento do modelo às observações da série em estudo.

Na avaliação da qualidade das estimativas obtidas é crucial analisar a significância estatística dos parâmetros estimados. Para isso, a cada parâmetro, diga-se β_i (onde $i = 1, \ldots, m$ e m representa o número de parâmetros estimados), deve ser aplicado um teste de hipóteses apropriado que avalie a necessidade (ou não) de incluir esse parâmetro no modelo. Desta forma, interessa testar a hipótese de que β_i é estatisticamente nulo, isto é, $H_0 : \beta_i = 0$. A rejeição desta hipótese acontece, a um nível de significância α, quando a estatística T associada ao coeficiente estimado for, em valor absoluto, superior ou igual ao quantil $1 - \frac{\alpha}{2}$ de uma distribuição t de Student com $n - m$ graus de liberdade, (onde n representa o número de observações), isto é,

$$|T| = \left| \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \right| \geq t_{1-\alpha/2;n-m}. \tag{5.13}$$

De forma equivalente, a hipótese nula é rejeitada, a um nível de significância α, quando o valor de prova obtido é inferior ou igual a α. Para Caiado (2011), tendo sempre presente o princípio da parcimónia, devem incluir-se no modelo apenas os parâmetros que possam considerar-se significativamente diferentes de zero

Relativamente à avaliação da qualidade do ajustamento do modelo SARIMA, deve efetuar uma análise do comportamento dos respetivos resíduos. Caso os resíduos tenham um comportamento semelhante a um ruído branco, pode dizer-se que o modelo estimado descreve bem a série em estudo.

Um modelo que não satisfaça os critérios aplicados em alguma das duas etapas, deve ser rejeitado. Nesses casos, as informações recolhidas durante a avaliação podem sugerir indicações que ajudem na formulação de um novo modelo.

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

$$e_t = Y_t - \hat{Y}_t. \tag{5.14}$$

Um bom modelo deve gerar resíduos com o comportamento idêntico ao de um ruído branco, e, portanto, estes devem apresentar média nula e satisfazer o pressuposto da não correlação. Adicionalmente, para a construção de intervalos de previsão, torna-se pertinente verificar se os resíduos têm variância constante e apresentam uma distribuição aproximadamente Normal (Hyndman & Athanasopoulos, 2018).

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

uma reta correspondente à bissetriz dos quadrantes ímpares ($y = x$). Para uma verificação rigorosa, os testes estatísticos mais comuns são o teste de Shapiro-Wilk (para amostras de pequenas dimensões, com menos de 50 observações) e o teste de Kolmogorov-Smirnov. Em ambos é testada a hipótese nula "os erros seguem uma distribuição Normal".

Relativamente à hipótese de não correlação, esta deve ser verificada tanto individualmente como de forma conjunta no que respeita à verificação gráfica e à verificação analítica. De forma particular, as autocorrelações dos resíduos podem ser avaliadas através da observação da FAC que, se o modelo for apropriado, deve apresentar um comportamento semelhante ao da FAC de um ruído branco, ou seja, com autocorrelações não significativamente diferentes de zero. De forma a testar várias autocorrelações como um grupo pode recorrer-se a um teste de Portmanteau. Um dos mais utilizados dentro desta classe de testes é o de Ljung-Box, cuja estatística de teste Q é definida por

$$Q = n(n+2)\sum_{j=1}^{k}\frac{\hat{\rho}_j^2}{n-j} \tag{5.15}$$

e segue aproximadamente uma distribuição do Qui-Quadrado com $k-m$ graus de liberdade (com k que corresponde ao número de autocorrelações a serem testadas e m ao número de parâmetros estimados). No caso de se rejeitar a hipótese nula, $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0$, conclui-se que o modelo escolhido não é apropriado. Em relação ao valor de k, não existe um critério específico para a sua escolha, pelo que se apreende que a melhor abordagem passa por realizar o teste para vários valores distintos.

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

$$\left|\frac{\bar{e}\sqrt{n}}{s_e}\right| \geq t_{1-\frac{\alpha}{2};n-1}, \tag{5.16}$$

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

5.1.4 Seleção de modelos

Nas diversas análises estatísticas de dados, encontram-se diversos modelos adequados para descrever o fenómeno em estudo. Mais especificamente, na modelação de uma série

temporal pode existir mais do que um modelo que verifique os diferentes critérios de avaliação do diagnóstico, o que torna penosa a tarefa de escolher o melhor modelo. Assim sendo, devem procurar-se critérios de seleção de modelos que ponderem as estatísticas baseadas nos resíduos do modelo ajustado.

Um critério plausível para escolher o melhor modelo SARIMA passaria na escolha do modelo que fornece a menor soma dos quadrados dos erros (ou erro quadrático médio) ou o maior valor para a função de verosimilhança. Contudo, esta abordagem nem sempre funciona porque, frequentemente, o erro quadrático médio pode ser reduzido e a função de verosimilhança aumentada simplesmente pelo aumento do número de parâmetros no modelo. De forma a solucionar esta questão, a função de verosimilhança deve ser penalizada por cada parâmetro adicional no modelo, ou seja, se o parâmetro extra não melhorar o valor da função de verosimilhança mais do que o valor da penalização, esse parâmetro não deve ser acrescentado ao modelo (Wheelwright, 1998).

Os critérios, baseados na função de verosimilhança, existentes na literatura, são diversos, sendo os mais utilizados o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC, *Bayesian Information Criterion*). Estes critérios incorporam duas componentes, uma que consiste no logaritmo da função de verosimilhança, que decresce quando o número de parâmetros estimados aumenta, e outra mais "penalizadora", que aumenta à medida que o número de parâmetros também aumenta. O que estes critérios consideram é, então, uma situação de equilíbrio entre as duas componentes.

Critério de Informação de Akaike (AIC)

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

$$\text{AIC} = -2\log L + 2m, \tag{5.17}$$

onde L é a função de verosimilhança. De salientar que nem todos os *softwares* estatísticos possuem a capacidade de determinar o AIC ou a função de verosimilhança L e, por isso, nem sempre é possível encontrar o AIC exato para um determinado modelo. Contudo, uma aproximação útil para o AIC é obtida através da aproximação

$$-2\log L \approx n(1+\log 2\pi) + n\log\sigma^2, \tag{5.18}$$

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

$$\text{AIC} \approx n(1+\log 2\pi) + n\log\hat{\sigma}^2 + 2m. \tag{5.19}$$

Às vezes, o primeiro termo em 5.19 é omitido por ser igual para todos os modelos.

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

Critério de informação Bayesiano (BIC)

Schwarz et al. (1978), propõe o critério de informação Bayesiano, definido como

$$\text{BIC} = -2\log L + m\log(n), \tag{5.20}$$

onde L é a função de verosimilhança, m é o número de parâmetros do modelo e n é a dimensão da amostra. Contrariamente ao AIC, o BIC depende da dimensão da amostra (n) pelo que, para $\log(n) > 2$, isto é, para uma amostra de dimensão superior a 7, a penalização do BIC é superior à penalização do AIC. Como consequência, a minimização do BIC leva, em geral, à seleção de modelos com um menor número de parâmetros do que os obtidos pela minimização do critério AIC, evitando, de certa forma, a sobrestimação do número de componentes.

5.1.5 Previsão

Numa fase posterior, após a escolha do modelo que melhor descreve a série temporal, pode prosseguir-se para o cálculo de previsões, sejam estas pontuais ou intervalares. As previsões pontuais podem ser facilmente determinadas fazendo uso da própria expressão do modelo escolhido. De facto, para obter previsões a h-passos, isto é, para um instante $t+h$, basta calcular a esperança condicionada aos valores observados, ou seja, $E[Y_{t+h}|Y_1, Y_2, \ldots, Y_t]$. Iniciando o processo para uma previsão a 1-passo, isto é, para $h = 1$, e repetindo-o para $h = 2, 3, \ldots$, é, então, possível obter todas as previsões pretendidas.

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

$$\left(\hat{y}_{t+h|t} - z_{1-\frac{\alpha}{2}}\hat{\sigma}_h, \hat{y}_{t+h|t} + z_{1-\frac{\alpha}{2}}\hat{\sigma}_h\right), \tag{5.21}$$

onde z é o quantil da distribuição Normal padrão, $1-\alpha$ corresponde ao nível de confiança do intervalo e $\hat{\sigma}_h$ é a estimativa do desvio padrão da previsão para o passo h. Numa perspetiva generalizada, os intervalos de previsão aumentam conforme o horizonte de previsão, h, aumenta. Contudo, em modelos estacionários (isto é, com $d = 0$) as sucessões dos limites inferiores e superiores são convergentes e, portanto, para horizontes distantes, os intervalos de previsão terão amplitudes idênticas (Hyndman & Athanasopoulos, 2018).

A literatura apresenta mais detalhes sobre estes e outros processos para a obtenção de previsões pontuais e intervalares, respetivamente, sendo que se aconselha as conclusões apresentadas em Box & Jenkins (2016).

5.2 Avaliação de Modelos de Previsão

Na previsão de séries temporais existe uma diversidade de métodos para o efeito. Estes métodos podem envolver diferentes parâmetros de uma complexidade considerável como também se podem caracaterizar pela sua simplicidade e fácil assimilação. No entanto, a complexidade de um método não significa uma melhoria na qualidade das previsões então Yokuma & Armstrong (1995) descrevem uma série de fatores que devem ser considerados no momento de escolha do método de previsão que, segundo Chatfield (2000) abrangem a precisão das previsões, a facilidade de utilização, a interpretação e implementação, a poupança de custos, a flexibilidade e outros demais fatores consideráveis nesta opção.

As medidas de avaliação tratam-se, na maioria dos casos, do maior critério de seleção de um método de previsão onde é possível avaliar a eficácia de um determinado modelo ou método de previsão de forma que demonstre a capacidade de reprodução da série temporal em análise e obtenção de previsões que se revelem o mais precisas possível.

No exercício de previsão deve-se ter em consideração a escolha da amostra de treino com o objetivo de ajustar o modelo e a escolha da amostra de teste de forma a verificar a sua qualidade preditiva. Contudo, um bom modelo de previsão não significa que seja aquele que se ajuste bem aos dados da amostra de treino. A avaliação da precisão das suas previsões só pode ser efetuada com uma amostra de observações que não tenha sido utilizada na sua estimação, isto é, a avaliação terá de ser realizada com recurso a uma amostra de teste.

Como tal, uma série temporal é usualmente dividida em série de treino, usada para estimar o modelo, e série de teste, para avaliar as previsões. Desta forma, ao considerar uma uma série temporal de dimensão n, $\{Y_1, Y_2, \ldots, Y_n\}$, e um horizonte temporal de previsão, h, tem-se a seguinte partição da série $\underbrace{Y_1, Y_2, \ldots, Y_{n-h}}_{\text{série de treino}}, \underbrace{Y_{n-h+1}, \ldots, Y_n}_{\text{série de teste}}$.

A dimensão da série de teste depende da dimensão da amostra e do horizonte temporal pretendido para previsão, logo é aconselhado escolher uma série de teste de dimensão igual ou superior ao horizonte temporal pretendido.

Na prática são consideradas duas formas de avaliar a qualidade preditiva:

- **Previsão a 1-passo**: prevê uma unidade temporal à frente ($h = 1$) da última observação, ou seja, se Y_t é a observação no instante t e $\hat{Y}_t$ a sua estimativa obtida usando as observações $Y_1, Y_2, \ldots, Y_{t-1}$, então $\hat{Y}_t$ é a previsão a 1-passo de Y_t;

- **Previsão a multipassos**: utilizando todas as observações até ao tempo t (inclusive), é obtida a previsão h-passos à frente, $\hat{Y}_{t+h|t}$. Generalizando, a previsão a h passos à

frente (previsão multi-passos), $\hat{y}_{t+h}$, corresponde à previsão para y_{t+h} obtida usando as mesmas observações, $y_1, y_2, \ldots, y_{t-1}$.

Os erros de previsão indicam se a metodologia de previsão é apropriada, sendo, por isso, importante medir a sua magnitude (Cordeiro, 2011). Um erro de previsão trata-se da diferença entre o valor observado e a sua previsão, sendo estes indicadores da qualidade da metodologia de previsão aplicada. O erro de previsão (a h-passos) pode ser escrito como $e_{t+h} = Y_{t+h} - \hat{Y}_{t+h|t}$. Caso se considere uma previsão a 1-passo ($h = 1$) o erro de previsão a 1-passo é dado por $e_t = Y_t - \hat{Y}_t$.

5.2.1 Medidas de Avaliação

Neste estudo são utilizadas quatro medidas para se proceder à avaliação da qualidade preditiva das metodologias aplicadas: o Erro Quadrático Médio (EQM), o Erro Absoluto Médio (EAM), o Erro Percentual Absoluto Médio (EPAM) e o Erro Escalado Absoluto Médio (EEAM).

O Erro Quadrático Médio (EQM) consiste no quadrado do valor médio dos desvios entre os valores observados e as previsões para os instantes $1, 2, \ldots, n$, ou seja,

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

Esta medida, caracterizada pela dependência da escala dos dados, é calculada através do erro de previsão a 1-passo, $e_t = Y_t - \hat{Y}_t$. É frequente que a raiz do erro quadrático médio, $\text{REQM} = \sqrt{\text{EQM}}$, seja preferida em relação ao EQM uma vez que permite redução da grandeza dos valores para a mesma escala dos dados. Visto que se $Y_t = \hat{Y}_t$ se obtém EQM $= 0$, na comparação entre métodos de previsão, considera-se que o método mais preciso é o que apresenta menor EQM e, tratando-se da raiz do EQM, aquele que detenham menor valor de REQM.

Note-se que estas medidas descritas são mais sensíveis à presença de *outliers* em comparação, por exemplo, ao Erro Absoluto Médio (EAM)

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

O erro percentual absoluto médio, EPAM, consiste na percentagem média do erro de previsão em conformidade com a grandeza das observações. Esta medida define-se como:

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

Note-se que se $Y_t = \hat{Y}_t$ tem-se EPAM $= 0\%$ logo, quanto menor for o EPAM mais preciso é o método de previsão. Esta medida não pode ser calculada quando existem zeros na série e, quando as observações se aproximam de zero, o EPAM apresenta valores

extremos. O Erro Escalado Absoluto Médio, EEAM, permite a comparação relativa de qualquer método de previsão com o método *naïve* (referência). O EEAM define-se como

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

Nesta medida, o erro escalado, q_t, é resultado do quociente entre o erro de previsão, $e_t = Y_t - \hat{Y}_t$, e o erro absoluto médio da previsão *naïve*, $\text{EAMN} = \sum_{t=2}^{n}|Y_t - Y_{t-1}|/(n-1)$. Caso se verifique sazonalidade nas séries em estufo, o denominador do erro escalado deverá ser substituído pelo erro absoluto médio da previsão *naïve* sazonal, $\text{EAMNs} = \sum_{t=s+1}^{n}|Y_t - Y_{t-s}|/(n-s)$, onde $s = 12$.

Acrescente-se que o EEAM pode ser empregado de forma a comparar a qualidade preditiva de séries com diferentes escalas, visto que a grandeza q_t se caracteriza pela independência da escala dos dados.

Caso se verifiquem valores de EEAM superiores a 1, estes indicam que as previsões para o método adotado são menos precisas, em média, do que as *naïve* e, por isso, quanto mais próximo de zero maior será a precisão do método. Então, na comparação entre diferentes métodos de previsão, considera-se que o mais preciso é o que apresenta o menor EEAM.

Capítulo 6

Análise Exploratória de Dados

Neste Capítulo é apresentada uma análise preliminar dos dados fornecidos pelo SHU da VITRUS referentes aos circuitos de recolha de resíduos indiferenciados, distribuídos pelas 44 freguesias do município de Guimarães e pela recolha indiferenciada e seletiva efetuada no Centro Histórico de Guimarães intramuros (sistema PAYT).

Também é apresentada uma análise de dados referente aos tipos de recolha dos resíduos em contentores de profundidade, porta a porta ou recolha mista, nos circuitos de recolha indiferenciada em estudo (nas freguesias do município).

O objetivo principal destas análises é avaliar o comportamento e a distribuição das diferentes variáveis em estudo, de forma a permitir uma aplicação mais adequada das metodologias de modelos de Regressão Linear e de modelos de Séries Temporais.

6.1 Circuitos de recolha indiferenciada

A VITRUS é responsável pela recolha dos resíduos indiferenciados em 44 freguesias do concelho de Guimarães. Inicialmente, o SHU era responsável pela recolha em seis circuitos. Com o decorrer dos anos o número de circuitos tem vindo a aumentar e, neste momento, o SHU opera em doze circuitos de recolha indiferenciada (ver Apêndice A, Tabela A.1).

Nestes serviços de recolha indiferenciada há três tipos de recolha: recolha de contentores de profundidade, recolha porta a porta e a recolha mista.

Os dados relativos aos circuitos de recolha indiferenciada foram extraídos a 30 de agosto de 2019 e contêm a informação relativa às pesagens diárias obtidas de resíduos indiferenciados, nos respetivos circuitos, conforme a tipologia de recolha. Os dados, numa fase preliminar, foram transformados em observações semanais, por circuito de recolha, com o intuito de efetuar uma análise descritiva de forma a descrever e compreender o comportamento destas variáveis. Acrescente-se que, em particular, a variável respetiva à recolha de resíduos indiferenciados em contentores de profundidade será, posteriormente, analisada via modelos de Séries Temporais com o objetivo de se obterem previsões a longo prazo. O período observado foi entre a 16.ª semana de 2016 (semana de 17 de abril) e a 34.ª

semana de 2019 (semana de 18 de agosto). Todos os circuitos associados aos contentores de profundidade, os primeiros a serem adjudicados pela VITRUS, contêm informação desde 2016

Analisando os resultados descritos na Tabela 6.1, relativamente aos circuitos de contentores de profundidade (Circuitos 1, 2, 4, 5 e 6), comprova-se um aumento anual da quantidade de resíduos indiferenciados, à exceção do circuito 6 que apresenta um decréscimo entre 2017 e 2018 de 308,48 toneladas. Em relação aos circuitos porta a porta (Circuitos 9, 10, 11 e 12), cujos dados são observados desde 2017, verifica-se um decréscimo nos Circuitos 9, 11 e 12 entre 2017 e 2018. Apenas o Circuito 10 apresenta um aumento das quantidades recolhidas de resíduos indiferenciados. Por último, na recolha mista (Circuitos 3, 7 e 8), o Circuito 3 apresenta um aumento ao longo dos três anos apresentados, devendo-se ao facto de numa fase inicial, em 2016, a maioria dos resíduos eram recolhidos em contentores de profundidade sendo que, em 2017, após o incremento da recolha porta a porta neste circuito é verificado um aumento nas quantidades de resíduos recolhidos.

Já os Circuitos 7 e 8, observados desde 2017, também apresentam um aumento significativo nas quantidades recolhidas. As conclusões descritas também podem ser observadas na Figura 6.1. De uma forma geral, é notório o aumento da produção de resíduos indiferenciados nos circuitos de recolha indiferenciada, operados pelo SHU da VITRUS.

Tabela 6.1: Evolução anual das quantidades de resíduos indiferenciados (em toneladas), por tipo de recolha.

Contentores de profundidade				**Porta a porta**				**Mista**			
	2016	**2017**	**2018**		**2016**	**2017**	**2018**		**2016**	**2017**	**2018**
C1	3166,28	4416,16	4556,30	**C9**	-	1253,56	1158,80	**C3**	2586,70	3965,32	4082,98
C2	3089,86	4393,44	4651,96	**C10**	-	989,30	1109,74	**C7**	-	930,18	1135,90
C4	1776,84	2487,44	3332,22	**C11**	-	468,84	682,32	**C8**	-	588,16	1222,58
C5	2316,78	3609,88	4250,72	**C12**	-	889,58	622,88				
C6	588,60	1418,68	1110,20								
Total	10938,36	16325,60	17901,40		-	3601,28	3573,74		2586,70	5483,66	6441,46

Pela Figura 6.2 é notório que no período observado a recolha de contentores de profundidade (Circuitos 1, 2, 4 e 5) é responsável pelas elevadas quantidades de resíduos indiferenciados recolhidos. Este gráfico também indica que a recolha mista (Circuitos 3, 7 e 8) tem valores inferiores em relação aos outros tipos de recolha, uma vez que o número de circuitos afetos a esta recolha são menores em comparação aos restantes. Pode-se afirmar, ainda, que a recolha de contentores de profundidade evolui de forma crescente, contrariamente à recolha porta a porta (Circuitos 9, 10, 11 e 12) e à recolha mista que evoluem de uma forma mais ou menos constante ao longo do tempo, desde o ano de 2018. Note-se que as recolhas de contentores de profundidade e mista, respetivamente, apresentam uma evolução crescente ao longo dos anos. A recolha porta a porta apresenta um ligeiro decréscimo nas quantidades recolhidas entre os anos de 2017 e 2018.

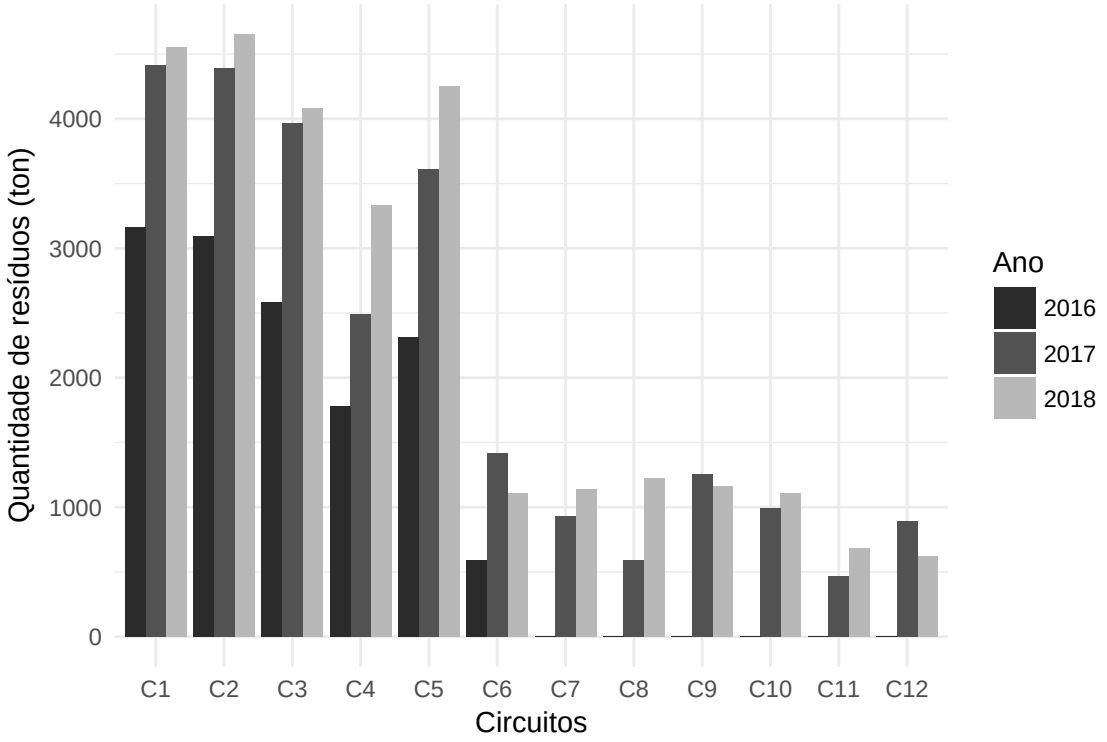


Figura 6.1: Evolução das quantidades de resíduos indiferenciados, recolhidos por ano.

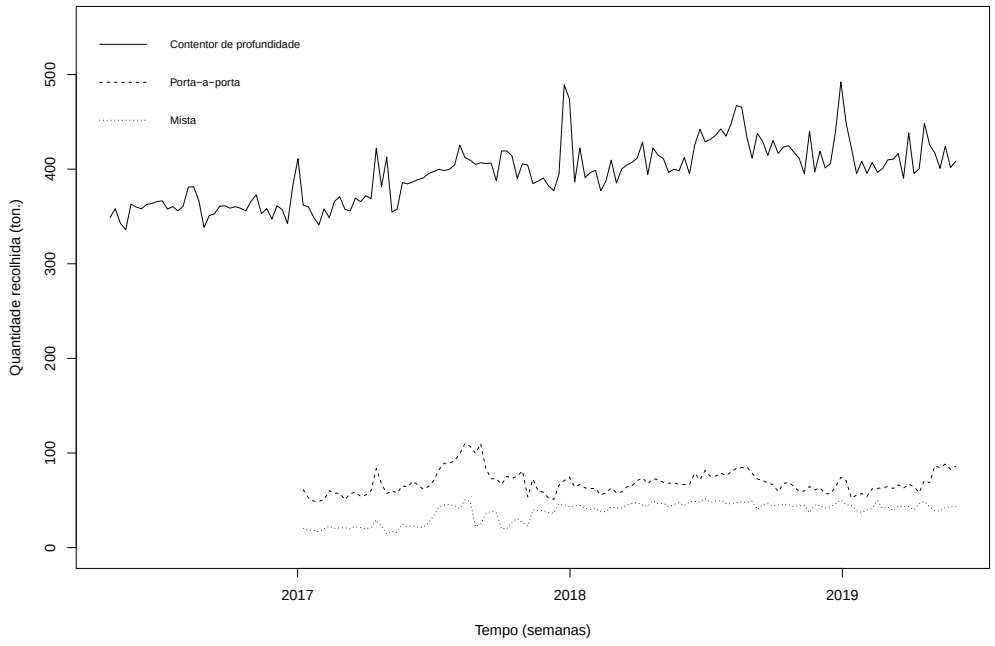


Figura 6.2: Evolução das quantidades de resíduos indiferenciados por tipo de recolha.

De acordo com a Tabela 6.2, os circuitos que detêm um valor médio de produção de

resíduos indiferenciados superior são o Circuito 1 e o Circuito 2. De realçar que o Circuito 3, 4 e 5, também apresentam valores elevados, podendo dever-se ao facto de terem sido os circuitos pioneiros na recolha de resíduos, operados pelo SHU. Em relação aos Circuitos 6, 7, 8, 9 e 10, o SHU apresenta uma produção média de 20 toneladas semanais. Já os Circuitos 11 e 12 detêm um valor médio semanal entre as 15 e as 16 toneladas, aproximadamente. São verificados valores omissos nos dados de alguns circuitos, uma vez que só são verificadas observações no ano de 2017.

Tabela 6.2: Estatísticas descritivas dos circuitos de recolha indiferenciada no período observado.

	Mínimo	**Máximo**	**Média**	**1.º Quartil**	**Mediana**	**3.º Quartil**	**Desvio Padrão**	**NA's**
Circuito 1	69,04	108,30	86,01	82,67	85,20	90,48	6,29	1
Circuito 2	28,38	106,50	85,02	81,90	84,93	88,86	7,21	0
Circuito 3	17,98	111,50	76,34	69,82	75,24	82,49	10,50	0
Circuito 4	15,32	82,24	55,29	46,43	53,09	65,36	10,69	0
Circuito 5	0,00	98,82	72,99	64,04	73,93	81,04	11,14	0
Circuito 6	0,00	52,20	21,13	16,32	18,38	27,76	8,86	3
Circuito 7	0,00	29,94	20,40	19,84	21,38	23,23	5,85	37
Circuito 8	0,00	31,22	22,27	20,70	23,10	24,39	4,35	62
Circuito 9	0,00	42,46	22,16	19,58	21,22	23,84	5,17	37
Circuito 10	0,00	35,28	19,67	17,44	19,58	21,82	4,37	37
Circuito 11	0,00	37,58	15,92	11,76	13,98	19,34	6,42	63
Circuito 12	0,00	42,66	15,48	10,56	12,82	22,07	8,55	37

De facto, de acordo com as Figuras 6.3 e 6.4, verifica-se que a recolha de contentores de profundidade é superior às recolha porta a porta e recolha mista, tendo esses circuitos uma elevada dispersão. De realçar que, tal como referido anteriormente, os Circuitos 1 e 2 são os que detêm uma maior produção de resíduos sendo notório o elevado valor mediano nestes circuitos. Em relação à presença de *outliers*, realça-se que apenas nos circuitos 4 e 5 não se verifica a presença de *outliers*. O gráfico da Figura 6.4 sugere uma tendência crescente na produção de resíduos indiferenciados nos Circuitos 3, 4 e 5 ao longo do tempo observado. Também, a partir dos diagramas de caixa com bigodes verifica-se uma baixa dispersão dos Circuitos 8 a 11.

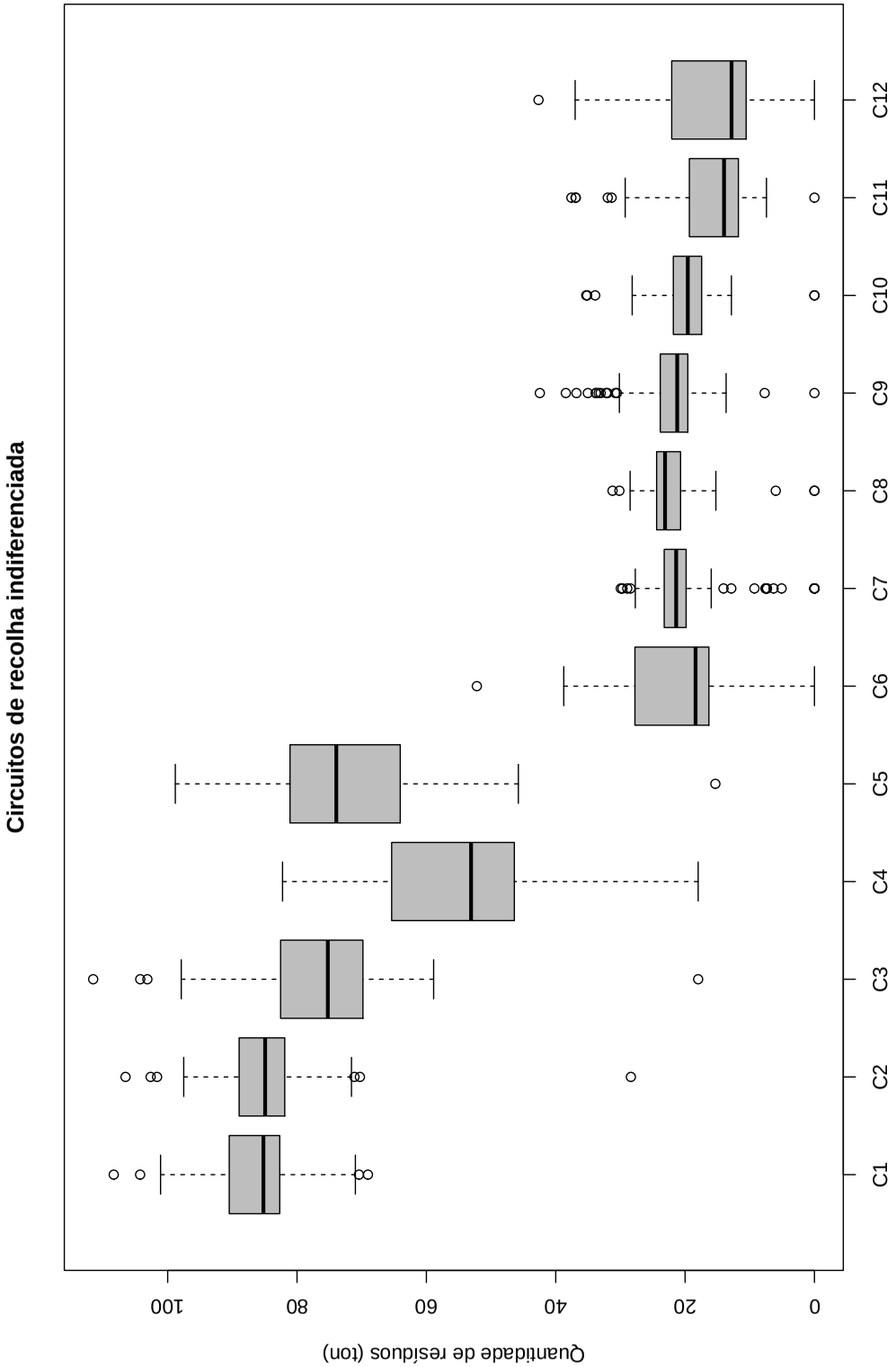


Figura 6.3: Diagramas em caixa de bigodes da produção de resíduos semanal dos circuitos (1 a 12) de recolha operados pela VITRUS.

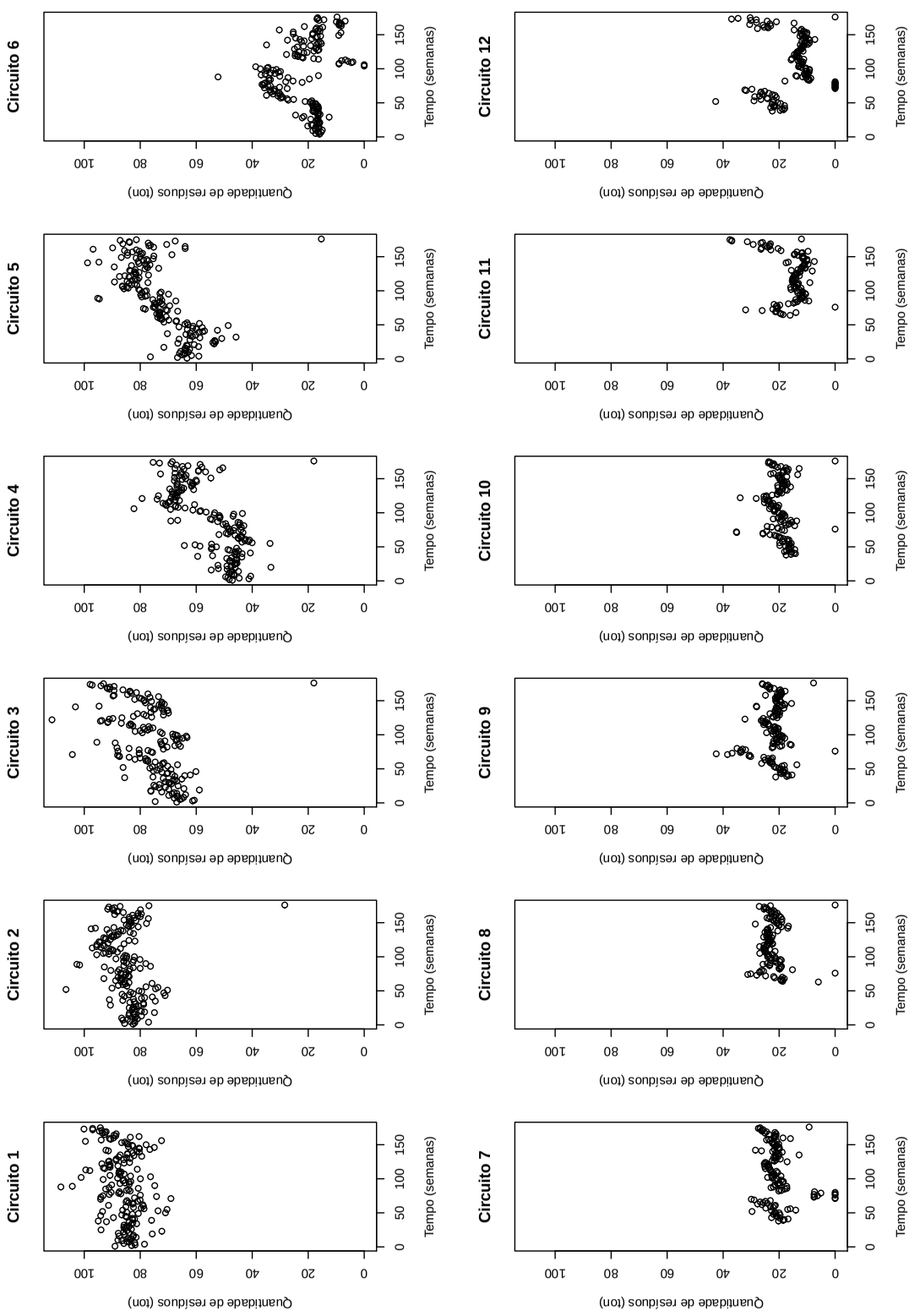


Figura 6.4: Diagramas de dispersão da produção de resíduos semanal dos circuitos (1 a 12) de recolha operados pela VITRUS.

6.2 *Pay-as-you-throw*: o caso de Guimarães

Na Tabela 6.3 são apresentadas as variáveis em estudo, a sua respetiva classificação e unidades de medida.

Tabela 6.3: Descrição das variáveis em estudo, relativamente ao sistema PAYT, implementado no Centro Histórico intramuros.

Variável	**Descrição**	**Classe**	**Unidade de medida/ Categorias**
MONTH	Mês	Quantitativa discreta	
IND	Quantidade de resíduos indiferenciados	Quantitativa contínua	toneladas
SEL	Quantidade de resíduos seletivos	Quantitativa contínua	toneladas
PAP	Quantidade de resíduos de papel/cartão	Quantitativa contínua	toneladas
PLA	Quantidade de resíduos de plástico/embalagens	Quantitativa contínua	toneladas
VID	Quantidade de resíduos de vidro	Quantitativa contínua	toneladas
SAC15UD	Número de sacos, de 15 litros, vendidos a utilizadores domésticos	Quantitativa discreta	
SAC15UND	Número de sacos, de 15 litros, vendidos a utilizadores não domésticos	Quantitativa discreta	
SAC30UD	Número de sacos, de 30 litros, vendidos a utilizadores domésticos	Quantitativa discreta	
SAC30UND	Número de sacos, de 30 litros, vendidos a utilizadores não domésticos	Quantitativa discreta	
SAC50UD	Número de sacos, de 50 litros, vendidos a utilizadores domésticos	Quantitativa discreta	
SAC50UND	Número de sacos, de 50 litros, vendidos a utilizadores não domésticos	Quantitativa discreta	
SAC100UD	Número de sacos, de 100 litros, vendidos a utilizadores domésticos	Quantitativa discreta	
SAC100UND	Número de sacos, de 100 litros, vendidos a utilizadores não domésticos	Quantitativa discreta	
UDC	Número de utilizadores domésticos que efetuaram compras	Quantitativa discreta	
TAC	Número de utilizadores não domésticos, da tipologia A, que efetuaram compras	Quantitativa discreta	
TBC	Número de utilizadores não domésticos, da tipologia B, que efetuaram compras	Quantitativa discreta	
TCC	Número de utilizadores não domésticos, da tipologia C, que efetuaram compras	Quantitativa discreta	
TDC	Número de utilizadores não domésticos, da tipologia D, que efetuaram compras	Quantitativa discreta	
TEC	Número de utilizadores não domésticos, da tipologia E, que efetuaram compras	Quantitativa discreta	
KMS	Distância percorrida na recolha de resíduos	Quantitativa contínua	quilómetros
FRT	Número de fretes efetuados	Quantitativa discreta	
DEP	Número de deposições ilegais	Quantitativa discreta	
OFERECE	Entrega de sacos para reciclagem de forma gratuita	Qualitativa nominal	1: Não efetuou 0: Efetuou
REC30	Número de sacos, de 30 litros, para a reciclagem, oferecidos	Quantitativa discreta	
REC50	Número de sacos, de 50 litros, para a reciclagem, oferecidos	Quantitativa discreta	
REC100	Número de sacos, de 100 litros, para a reciclagem, oferecidos	Quantitativa discreta	

Relativamente aos dados dos resíduos recolhidos no CHG, foram extraídos a 30 de agosto de 2019 e contêm a informação relativa às pesagens obtidas, de resíduos indiferenciados e seletivos, respetivamente, na zona de implementação do sistema PAYT.

Os dados diários da produção dos resíduos (`SEL`, `IND`, `PAP`, `PLA` e `VID`), após tratamento e manipulação, foram reduzidos a 33 observações mensais e, posteriormente, as variáveis `SEL` e `IND` foram transformadas em 149 observações semanais para posterior aplicação das metodologias para previsão de séries temporais. Resumidamente os dados estão compreendidos entre a 9.ª semana de 2016 e a 1.ª semana de 2019, ou seja, o período observado está compreendido entre 6 de março de 2016 e 6 de janeiro de 2019.

A partir da Tabela 6.4 são observadas as principais estatísticas descritivas das variáveis observadas no sistema PAYT. Em relação à quantidade total de resíduos produzidos, na zona piloto deste sistema a quantidade média de resíduos produzidos é igual a 76,55 toneladas por mês e com uma quantidade mínima e máxima igual a 45,42 toneladas e 103,84 toneladas, respetivamente.

Em relação às quantidades de resíduos indiferenciados e seletivos, verificam-se valores médios iguais a 51,63 toneladas e 24,92 toneladas, respetivamente. De notar que os resíduos indiferenciados detêm um valor máximo igual a 79,44 toneladas, enquanto que, os resíduos seletivos detêm um valor inferior, igual a 36,92 toneladas.

Continuando a análise dos resíduos seletivos, os resíduos de papel e cartão têm uma produção média de 6,58 toneladas por mês e, no período observado, uma produção entre as 3,02 toneladas e as 11,24 toneladas. Em relação aos resíduos de plástico e mistura de embalagens têm uma produção média de 4,53 toneladas em que, no período observado a produção se situou entre as 3,24 toneladas e as 6,72 toneladas. Por último, os resíduos de vidro têm uma produção média de 13,81 toneladas, sendo os resíduos que representam a grande parte da recolha seletiva. No período observado foram observados valores entre as 7,80 toneladas e as 21,70 toneladas mensais.

Em relação à quilometragem efetuada para a recolha de resíduos, as viaturas afetas ao serviço de recolha de resíduos na zona PAYT no Centro Histórico intramuros andam, em média 1785 quilómetros mensais. Deve realçar-se que o valor mínimo percorrido foi de 575 quilómetros e o valor máximo percorrido foi de 3764 quilómetros.

Os fretes realizados, isto é, o número de vezes que as viaturas se dirigem à estação de triagem para deposição de resíduos, situa-se entre 13 e 31 vezes. A variável correspondente aos fretes realizados detém um valor mediano igual a 18 e um valor médio igual a cerca de 19 fretes.

Um dos pontos fulcrais do sistema é a fiscalização e a posterior contagem do número de deposições ilegais por parte dos utilizadores. Em média, são verificadas, aproximadamente, 286 deposições ilegais mensais desde a implementação do projeto. Acrescente-se que o valor máximo observado foi igual a 526 deposições e o valor mínimo igual a 55 deposições ilegais.

Tabela 6.4: Estatísticas descritivas relativas às variáveis em estudo do circuito PAYT (mensais).

	Minimo	**Máximo**	**1.º Quartil**	**Mediana**	**Média**	**3.º Quartil**	**Desvio Padrão**	**NA's**
IND	29,34	79,44	44,88	50,06	51,63	58,64	10,21	0
SEL	16,08	36,92	21,08	24,40	24,92	28,14	5,31	0
PAP	3,02	11,24	5,00	6,44	6,58	8,00	2,15	0
PLA	3,24	6,72	3,72	4,26	4,53	5,30	0,93	0
VID	7,80	21,70	10,44	12,88	13,81	16,74	3,86	0
KMS	575,00	3764,00	1017,00	1888,00	1785,00	2127,00	847,61	0
FRT	13,00	31,00	17,00	18,00	19,45	21,00	4,13	0
DEP	55,00	526,00	200,50	239,00	285,80	395,50	137,97	0

Na Figura 6.5, os gráficos da evolução da produção de resíduos, indicam a presença de uma tendência crescente na produção de resíduos indiferenciados. Já na produção de resíduos seletivos indicam a presença de sazonalidade, na sua globalidade, e, de forma particular, na produção de resíduos de vidro e de resíduos de plástico.

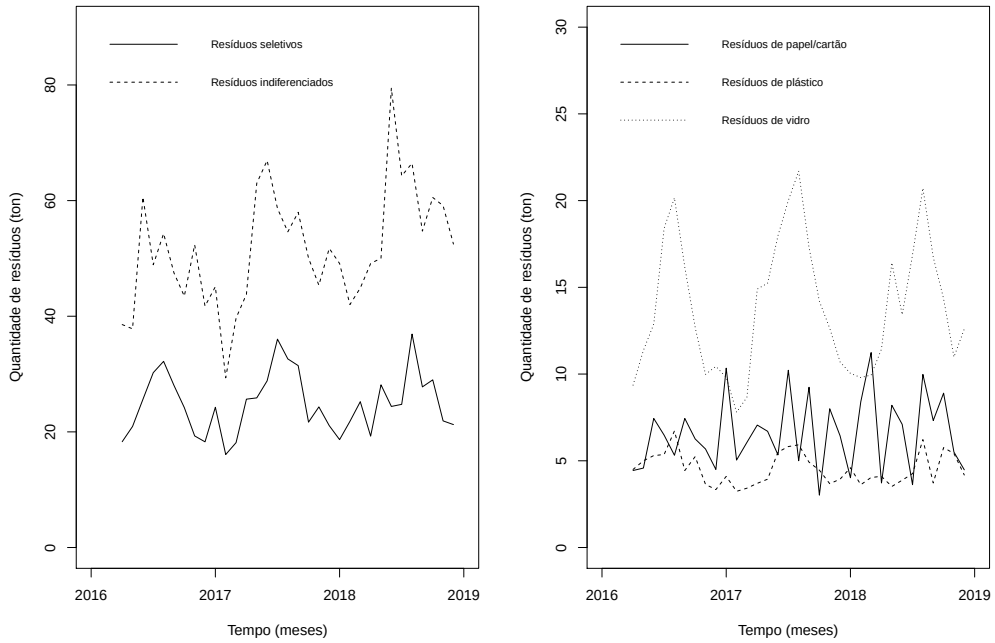


Figura 6.5: Representação gráfica da evolução da produção de resíduos.

Já na Tabela 6.5 é explorada de uma forma analítica, a produção dos diversos resíduos no Centro Histórico de Guimarães intramuros, a zona piloto do sistema PAYT. Assim, verifica-se que, com a implementação do sistema PAYT no CHG, a produção de resíduos seletivos, em relação a 2015, aumentou para as 276 toneladas correspondente a um acréscimo de 53,6%. Já os resíduos sofreram uma queda das 821 toneladas para as 538 toneladas, perfazendo um decréscimo igual a 52,6%. No ano de 2017 continuam a verificar-se aumen-

tos nos resíduos seletivos, das 276 para as 306,04 toneladas, com acréscimos percentuais iguais a 9,8% e, também, nos resíduos indiferenciados um acréscimo de 11,3% correspondente ao aumento das 538 para as 606,26 tonelada anuais. Relativamente ao último ano observado, 2018, os resíduos seletivos sofreram um decréscimo de 2,3%, enquanto os resíduos indiferenciados mantêm a tendência crescente, verificando-se um aumento de 9,8% em relação a 2017.

Tabela 6.5: Evolução da produção mensal de resíduos no Centro Histórico intramuros, zona piloto do sistema PAYT.

	Papel/Cartão	Plástico	Vidro	**Seletivo**	**Indiferenciado**	**Total**
2015	33,00	28,00	67,00	128,00	821,00	949,00
2016	74,00	54,00	148,00	276,00	538,00	814,00
2017	82,46	52,64	170,94	306,04	606,26	912,30
2018	82,48	53,32	163,32	299,12	672,22	971,34

Nos diagramas de caixa com bigodes apresentados na Figura 6.6 é possível verificar uma assimetria positiva nos dados relativos aos resíduos indiferenciados, vidro e plástico. Em relação aos resíduos seletivos e aos resíduos de papel/cartão apresentam uma distribuição quase simétrica. De realçar que se verifica um *outlier* nos dados dos resíduos urbanos indiferenciados, correspondente ao mês de junho de 2018.

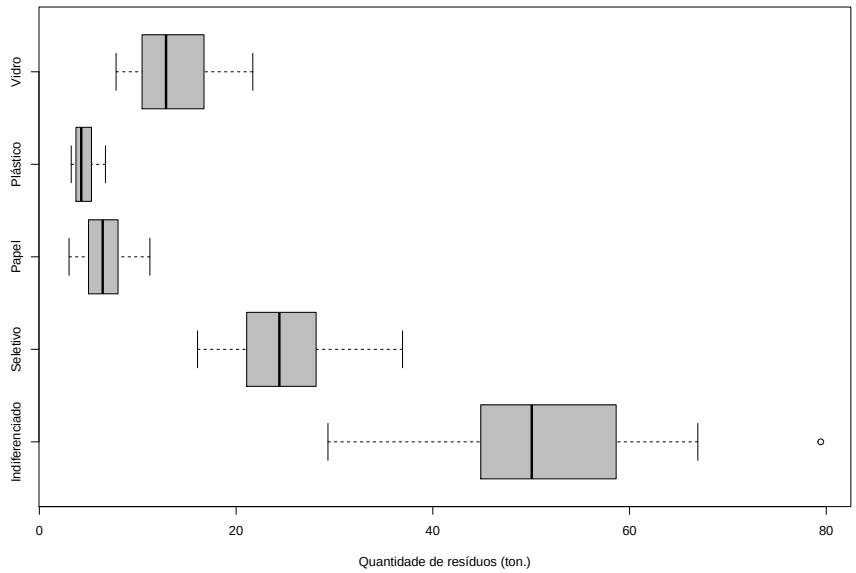


Figura 6.6: Diagramas de caixa com bigodes relativos às quantidades de resíduos, produzidas mensalmente.

A Tabela 6.6 apresenta as quantidades vendidas aos utilizadores, conforme a litragem de sacos, por parte da VITRUS. De uma forma geral verificam-se aumentos das vendas em todas as litragens de sacos, sendo a mais notória a dos sacos de 50 litros que apresentam

um aumento de 16% nas vendas entre 2017 e 2018. Nas restantes litragens verificam-se aumentos entre os 1,5% e os 12%.

Tabela 6.6: Quantidades de sacos vendidas, anualmente, na globalidade, conforme a litragem de sacos.

Venda de sacos				
	15 litros	**30 litros**	**50 litros**	**100 litros**
2016	2005	7352	11858	6217
2017	5785	8539	16545	8339
2018	6506	8674	19218	9157

A partir dos gráficos representados na Figura 6.7, para cada tipo de utilizador: utilizador doméstico (UD) e utilizador não doméstico (UND), verifica-se que são os utilizadores não domésticos que adquirem mais sacos, contrariamente aos utilizadores domésticos. De realçar que os sacos da litragem de 50 litros e 100 litros, respetivamente, são aqueles que têm maior procura dentro do grupo dos UND. Já os sacos de 15 litros e 30 litros, nos utilizadores domésticos são os mais adquiridos. O sacos de 50 e 100 litros, no grupo dos utilizadores domésticos, apresentam um comportamento homogéneo, não se verificando grandes alterações das vendas ao longo do tempo. Também de realça que os sacos de 15 litros apresentam uma tendência crescente de compra nos UD e, contrariamente, os sacos de 30 litros apresentam uma tendência decrescente, pouco significativa, nas vendas mensais.

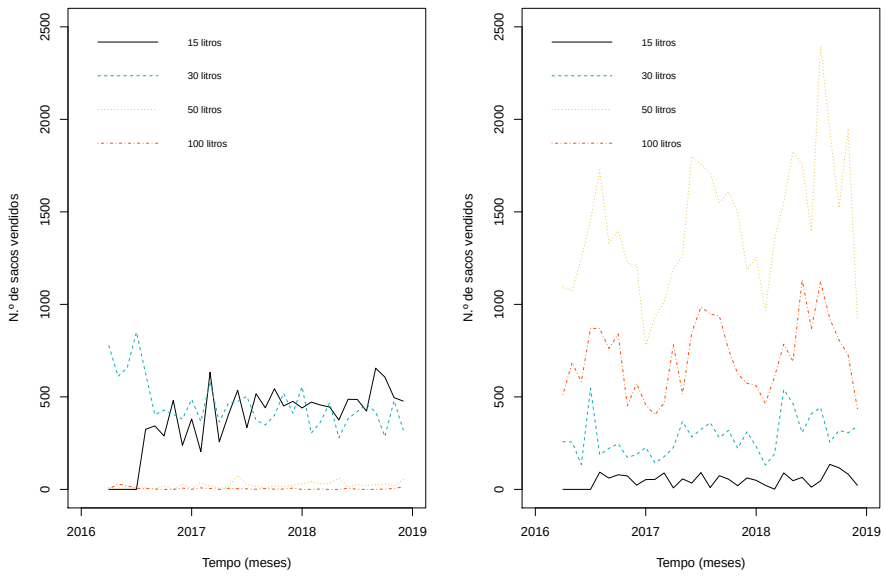


Figura 6.7: Evolução do número de sacos vendidos, por litragem, conforme o tipo de utilizador (esquerda: UD, direita: UND).

Relativamente aos utilizadores não domésticos, o gráfico sugere a presença de um

comportamento sazonal na venda mensal de sacos de 30, 50 e 100 litros, sendo estes também os sacos mais adquiridos por este tipo de utilizadores. O número de sacos de 15 litros apresentam uma evolução significativa, mas ténue em comparação aos restantes.

A Figura 6.8 apresenta os gráficos relativos à evolução anual, desde o início da implementação do sistema PAYT, da compra de sacos conforme a sua litragem. Daqui se retira que os utilizadores domésticos compram com regularidade sacos de 15 e 30 litros, respetivamente, uma vez que detêm uma maior número de compras em comparação com os utilizadores não domésticos. Já os sacos de 50 e 100 litros são adquiridos, aproximadamente na sua totalidade, pelos utilizadores não domésticos, sendo observável um valor residual nos utilizadores domésticos em relação à compra desta litragem de sacos. Refira-se também o aumento do número de vendas relativas aos sacos de 15 litros nos dois tipos de utilizadores e o aumento do número de sacos de 30 litros nos utilizadores domésticos. Nas restantes litragens verificam-se decréscimos no número de vendas, situação que se deverá ter em conta uma vez que pode ser influenciada por uma variedade de fatores, desde a mudança de habitação por parte dos utilizadores, quebras de *stock* ou até mesmo a prevaricação que terá como consequência as deposições ilegais.

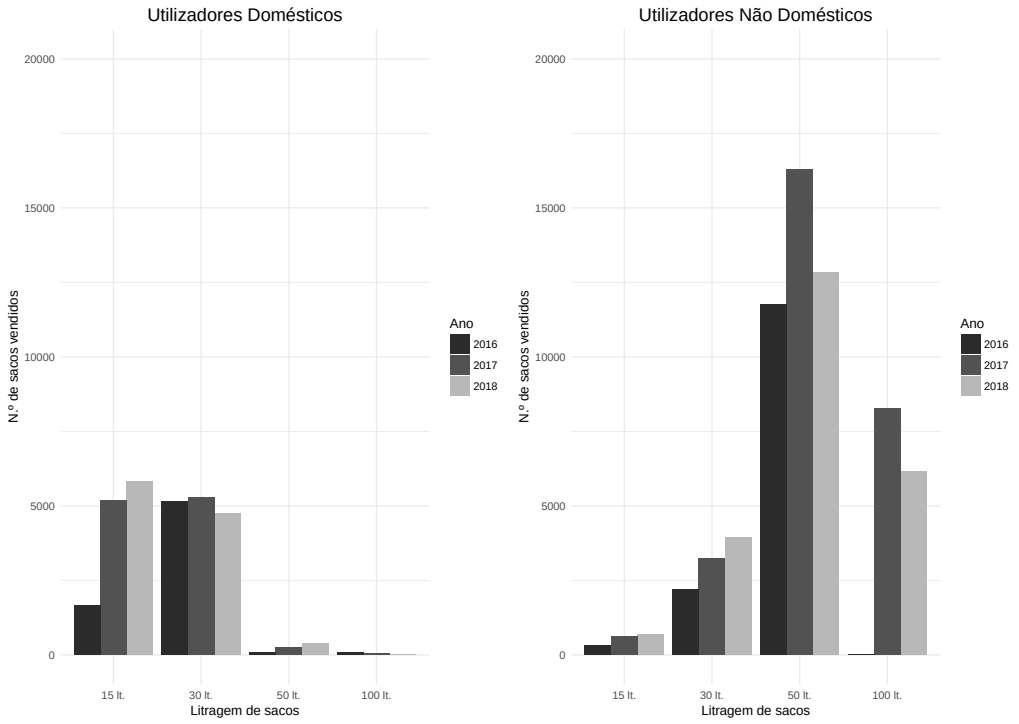


Figura 6.8: Gráficos relativos à evolução das compras efetuadas pelos utilizadores conforme a litragem do saco.

Relativamente às deposições ilegais é verificada, pela Figura 6.9, uma tendência crescente ao longo do tempo, situação que deverá servir de mote para uma agilização na fiscalização de forma a garantir a higienização e a disciplinação dos utilizadores.

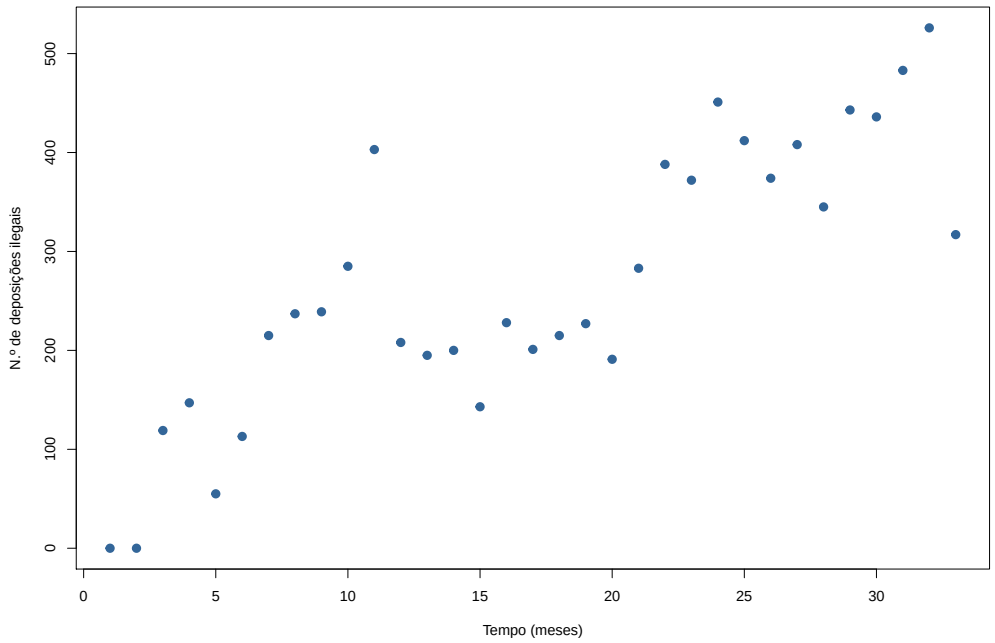


Figura 6.9: Evolução do número de deposições ilegais na zona piloto de implementação do sistema PAYT.

Para avaliar a associação entre as variáveis relativas ao número de sacos vendidos aos utilizadores e o número de deposições ilegais, foi aplicado o Teste de Correlação de Spearman (Spearman, 1987). Este teste foi utilizado uma vez que se tratam de variáveis quantitativas discretas. A Tabela 6.7 apresenta o coeficiente de correlação de Spearman, ρ_s, e o respetivo valor de prova da aplicação do Teste de correlação de Spearman para um nível de significância de 10% ($\alpha = 0,10$). Daqui se retira que apenas as variáveis `SAC30UD` e `SAC100UND` não apresentam correlações significativas com a variável respeitante às deposições ilegais. Das variáveis restantes, retiram-se as seguintes interpretações dos coeficientes de correlação de Spearman:

- A variável `SAC15UD` apresenta uma correlação positiva significativa com `DEP`, com um coeficiente de correlação igual a 0,580, o que significa que cada saco de 15 litros são adquirido pelos UD permite o acréscimo de 0,580 ao valor esperado do número de deposições ilegais;

- A variável `SAC15UND` apresenta uma correlação positiva significativa com `DEP`, o que significa que os sacos de 15 litros são adquiridos pelos UND contribuem para o aumento de 0,374 do valor esperado do número de deposições ilegais;

- A variável `SAC30UD` apresenta uma correlação negativa significativa com `DEP`, com um coeficiente de correlação igual a -0,569, o que significa que a venda de sacos de 15 litros aos UND contribui num decréscimo de -0,569 do valor esperado do número de deposições ilegais.

– A variável SAC50UD apresenta uma correlação positiva significativa com DEP, com um coeficiente de correlação igual a 0,362, o que significa que cada saco de 50 litros adquirido pelos UD contribui num acréscimo de 0,362 ao valor esperado do número de deposições ilegais;

– A variável SAC50UND apresenta uma correlação positiva significativa com DEP, que permite concluir que a venda de sacos de 50 litros aos UND permite um acréscimo de 0,288 ao valor esperado de deposições ilegais, por cada saco vendido;

– A variável SAC100UD apresenta uma correlação negativa significativa com DEP, com um coeficiente de correlação igual a -0,393, ou seja, cada saco de 100 litros adquirido pelos UD permite o decréscimo de -0,393 ao valor esperado das deposições ilegais.

Tabela 6.7: Teste de correlação de Spearman para avaliar a associação entre as variáveis (n.º de sacos vendidos) e o número de deposições ilegais.

	SAC15UD	SAC15UND	SAC30UD	SAC30UD	SAC50UD	SAC50UND	SAC100UD	SAC100UND
Valor de prova	<0,001	0,032	<0,001	0,267	0,039	0,105	0,024	0,605
r_s	0,580	0,374	-0,569	0,199	0,362	0,288	-0,393	0,093

Capítulo 7

Aplicação de Modelos de Regressão Linear

Os Modelos de Regressão Linear têm como principal objetivo avaliar o efeito de uma ou mais variáveis (covariáveis) sobre uma variável de interesse (variável resposta). Os resultados obtidos na aplicação da Regressão Linear Simples apenas nos indicam possíveis variáveis para integrar o modelo de Regressão Linear Múltipla pois, na prática, podem mudar as variáveis significativas do modelo devido a possíveis associações entre as covariáveis. Aprofundando a abordagem adotada no decorrer deste Capítulo, foram formulados modelos de regressão para a sazonalidade com recurso a indicadores sazonais, baseados nos modelos obtidos via regressão linear múltipla, com o objetivo de modelar os dados fornecidos conforme os períodos sazonais no período observado.

7.1 Regressão Linear Simples

Tendo em vista a influência que as covariáveis têm na produção de resíduos indiferenciados e seletivos, foram aplicados modelos de Regressão Linear Simples aos dados em estudo. Foram estimados 21 modelos de Regressão Linear Simples, para cada uma das variáveis resposta em estudo, respetivamente.

A Tabela 7.1 apresenta as estimativas dos parâmetros dos modelos de Regressão Linear Simples, os erros padrão das estimativas dos parâmetros, os valores de prova e respetivo coeficiente de determinação (R^2), tendo como variáveis resposta: `INDIFERENCIADO` e `SELETIVO`. De forma a verificar possíveis relações na produção de resíduos recicláveis/seletivos (papel/cartão, plásticos e vidro) apresentam-se no Apêndice B, nas Tabelas B.1, B.2 e B.3, os resultados obtidos após aplicação de Regressão Linear Simples.

Acrescenta-se que, durante o desenvolvimento da temática no presente Capítulo, será considerado um nível de significância de 10% ($\alpha = 0,10$).

Da análise dos resultados apresentados na Tabela 7.1 (onde se encontram as estimativas dos parâmetros do modelo, os erros padrão das estimativas dos parâmetros, o valor de

prova e o coeficiente de determinação (R^2)) é possível concluir que as covariáveis significativas na explicação da produção de resíduos resíduos indiferenciados, são significativas as covariáveis MTH, SAC15UD, SAC30UND, SAC50UND, SAC100UND, TAC, TBC, FRT, REC30 e REC50.

Tabela 7.1: Regressão Linear Simples, tendo como variável resposta SELETIVO e INDIFERENCIADO, respetivamente.

Regressão Linear Simples								
	SELETIVO				INDIFERENCIADO			
Variável	**Estimativas**	$\hat{\sigma}$	**p-valor**	R^2	**Estimativas**	$\hat{\sigma}$	**p-valor**	R^2
MTH	$\hat{\beta_0}$= 23,844	1,910	<0,001	0,013	$\hat{\beta_0}$= 43,610	3,308	<0,001	0,200
	$\hat{\beta_1}$= 0,06	0,100	0,524		$\hat{\beta_1}$= 0,472	0,17	**0,009**	
SAC15UD	$\hat{\beta_0}$=23,984	2,252	<0,001	0,007	$\hat{\beta_0}$=43,538	4,04	<0,001	0,135
	$\hat{\beta_1}$=0,002	0,005	0,652		$\hat{\beta_1}$=0,021	0,01	**0,035**	
SAC15UND	$\hat{\beta_0}$=23,994	1,580	<0,001	0,017	$\hat{\beta_0}$=48,504	2,985	<0,001	0,051
	$\hat{\beta_1}$=0,019	0,026	0,475		$\hat{\beta_1}$=0,064	0,049	0,205	
SAC30UD	$\hat{\beta_0}$=25,180	3,459	<0,001	0,000	$\hat{\beta_0}$=56,128	6,598	<0,001	0,016
	$\hat{\beta_1}$=-0,001	0,007	0,938		$\hat{\beta_1}$=-0,010	0,014	0,484	
SAC30UND	$\hat{\beta_0}$=19,963	2,532	<0,001	0,123	$\hat{\beta_0}$=41,470	4,822	<0,001	0,140
	$\hat{\beta_1}$=0,017	0,008	**0,045**		$\hat{\beta_1}$=0,036	0,016	**0,032**	
SAC50UD	$\hat{\beta_0}$=25,184	1,541	<0,001	0,002	$\hat{\beta_0}$=50,507	2,954	<0,001	0,007
	$\hat{\beta_1}$=-0,012	0,055	0,829		$\hat{\beta_1}$=0,050	0,105	0,636	
SAC50UND	$\hat{\beta_0}$=10,858	2,914	<0,001	0,444	$\hat{\beta_0}$=26,066	5,836	<0,001	0,397
	$\hat{\beta_1}$=0,010	0,002	**<0,001**		$\hat{\beta_1}$=0,018	0,004	**<0,001**	
SAC100UD	$\hat{\beta_0}$=25,626	1,138	<0,001	0,035	$\hat{\beta_0}$=52,187	2,221	<0,001	0,006
	$\hat{\beta_1}$=-0,162	0,153	0,296		$\hat{\beta_1}$=-0,128	0,298	0,671	
SAC100UND	$\hat{\beta_0}$=11,054	2,361	<0,001	0,545	$\hat{\beta_0}$=28,955	5,238	<0,001	0,394
	$\hat{\beta_1}$=0,019	0,003	**<0,001**		$\hat{\beta_1}$=0,032	0,007	<0,001	
UDC	$\hat{\beta_0}$=23,063	6,075	**<0,001**	0,003	$\hat{\beta_0}$=49,628	11,694	<0,001	0,001
	$\hat{\beta_1}$=0,027	0,087	0,759		$\hat{\beta_1}$=0,029	0,168	0,864	
TAC	$\hat{\beta_0}$=12,037	4,218	0,007	0,238	$\hat{\beta_0}$=31,484	8,528	<0,001	0,158
	$\hat{\beta_1}$=0,687	0,221	**0,004**		$\hat{\beta_1}$=1,074	0,446	**0,022**	
TBC	$\hat{\beta_0}$=5,872	5,217	0,269	0,306	$\hat{\beta_0}$=23,452	10,894	0,039	0,180
	$\hat{\beta_1}$=0,794	0,215	**0,001**		$\hat{\beta_1}$=1,174	0,449	**0,013**	
TCC	$\hat{\beta_0}$=22,888	2,076	<0,001	0,037	$\hat{\beta_0}$=51,087	4,067	<0,001	0,001
	$\hat{\beta_1}$=0,143	0,131	0,284		$\hat{\beta_1}$=0,038	0,256	0,883	
TDC	$\hat{\beta_0}$=18,017	2,945	<0,001	0,162	$\hat{\beta_0}$=45,767	6,089	<0,001	0,032
	$\hat{\beta_1}$=0,907	0,370	**0,020**		$\hat{\beta_1}$=0,771	0,766	0,322	
TEC	$\hat{\beta_0}$=25,792	3,235	<0,001	0,003	$\hat{\beta_0}$=54,228	6,208	<0,001	0,006
	$\hat{\beta_1}$=-0,127	0,448	0,780		$\hat{\beta_1}$=-0,376	0,86	0,665	
KMS	$\hat{\beta_0}$=23,986	2,210	<0,001	0,007	$\hat{\beta_0}$=46,333	4,133	<0,001	0,060
	$\hat{\beta_1}$=0,001	0,001	0,645		$\hat{\beta_1}$=0,003	0,002	0,167	
FRT	$\hat{\beta_0}$=8,700	3,494	0,018	0,420	$\hat{\beta_0}$=20,964	6,799	0,004	0,407
	$\hat{\beta_1}$=0,834	0,176	**<0,001**		$\hat{\beta_1}$=1,576	0,342	**<0,001**	
DEP	$\hat{\beta_0}$=25,449	2,078	<0,001	0,003	$\hat{\beta_0}$=46,788	3,88	<0,001	0,059
	$\hat{\beta_1}$=-0,002	0,007	0,776		$\hat{\beta_1}$=0,018	0,013	0,172	
REC30	$\hat{\beta_0}$=24,304	1,123	<0,001	0,030	$\hat{\beta_0}$=40,903	2,013	<0,001	0,156
	$\hat{\beta_1}$=0,003	0,003	0,341		$\hat{\beta_1}$=0,012	0,005	**0,023**	
REC50	$\hat{\beta_0}$=23,676	1,185	<0,001	0,078	$\hat{\beta_0}$=47,760	2,116	<0,001	0,204
	$\hat{\beta_1}$=0,002	0,001	0,116		$\hat{\beta_1}$=0,007	0,003	**0,008**	
REC100	$\hat{\beta_0}$=24,703	1,215	<0,001	0,078	$\hat{\beta_0}$=50,646	2,322	<0,001	0,014
	$\hat{\beta_1}$=0,001	0,002	0,782		$\hat{\beta_1}$=0,003	0,004	0,510	

Já na explicação da produção de resíduos seletivos na zona piloto de implementação do sistema PAYT são: `SAC30UND`, `SAC50UND`, `SAC100UND`, `TAC`, `TBC`, `TDC` e `FRT`.

Nesta análise não são ajustados um modelos de Regressão Linear Simples com a covariável `OFERECE` uma vez que, tratando-se de uma variável qualitativa nominal, não é apropriada para este tipo de modelação.

7.2 Regressão Linear Múltipla

O objetivo da análise de regressão múltipla é determinar se as covariáveis explicam o comportamento da variável dependente. No presente estudo pretende-se prever mudanças da variável respetiva à produção de resíduos indiferenciados e seletivos (variáveis resposta), respetivamente, associadas a mudanças das covariáveis já consideradas.

Verificadas as principais conclusões retiradas da análise exploratória efetuada aos dados do PAYT, o presente Capítulo incidirá na aplicação das metodologias aos mesmos, isto é, serão formulados modelos de regressão múltipla, para posterior auxílio na tomada de decisão. Definidas as variáveis resposta e respetivas covariáveis (covariáveis), é formulado o modelo completo, com todas as variáveis em estudo. Posto isto, é aplicado o método regressivo (*backward elimination*), onde são eliminadas, de forma iterativa, as variáveis detentoras de maior valor de prova (variável à qual corresponde a estatísticas de teste com valor absoluto mais baixo), até obter um modelo em que todas as covariáveis sejam significativas ao nível de significância considerado. Este método é utilizado de forma a obter um modelo de regressão que detenha as covariáveis relevantes no estudo do comportamento da variável resposta, com uma boa percentagem de explicação da variabilidade dos dados. Uma vez efetuado este passo, é necessária a análise do comportamento dos resíduos dos respetivos modelos. Desta forma é pretendido que se verifiquem, além da média nula e variância constante, a normalidade e a independência dos resíduos.

No presente estudo, os pressupostos da média nula e variância constante dos resíduos são avaliados de forma analítica e gráfica, respetivamente, mas para a verificação da condição exigida à média, caso os pressupostos de independência e normalidade dos resíduos não sejam rejeitados, recorre-se a um teste t para o valor médio. A condição da normalidade/gaussianidade dos resíduos é avaliada a partir de um histograma dos resíduos que deverá assemelhar-se ao comportamento da função densidade de uma distribuição Normal. De uma forma mais cuidada, complementando à análise gráfica, o teste de Shapiro-Wilk poderá ser realizado sob a hipótese nula da normalidade dos erros. De forma a avaliar a independência dos erros, uma vez que são estimados modelos via regressão linear, utiliza-se a estatística de Durbin-Watson, que não rejeita a hipótese de independência quando toma valores próximos de 2. Também a observação da FAC e da FACP dos resíduos deve servir de complemento à informação sobre os resíduos.

Nesta Secção serão apresentados os principais resultados após a modelação dos dados respetivos à produção de resíduos indiferenciados e seletivos. De uma forma detalhada,

também foram modelados os dados relativos à produção de resíduos de papel/cartão, plásticos e vidro, de forma a inferir sobre quais variáveis deverão ter uma atenção especial no que respeita à gestão do sistema implementado.

São apresentadas as estimativas dos parâmetros do modelo, os erros padrão das estimativas dos parâmetros, o valor de prova e o coeficiente de determinação ajustado (R_a^2), para cada tipo de resíduo. Após a formulação dos respetivos modelos será apresentada a devida análise de resíduos, com base nos pressupostos estudados, com vista à validação dos modelos. Note que no Apêndice C, nas Tabelas C.1, C.2 e C.3, são apresentados os valores obtidos após modelação dos dados relativos aos resíduos de papel/cartão, vidro e plástico, respetivamente.

Realça-se que, para todas as decisões, é considerado um nível de significância de 10%.

7.2.1 Resíduos indiferenciados

Ao analisar a Tabela 7.2 conclui-se que as variáveis `MTH`, `SAC50UD`, `FRT` e a não entrega, de forma gratuita, de sacos para a reciclagem, `OFERECE:1`, são significativas na explicação da produção de resíduos indiferenciados na zona em estudo.

A variável `MTH` contribui para um aumento de 1,118 toneladas na produção mensal de resíduos urbanos indiferenciados, que se traduz numa tendência crescente no tempo observado. Já a variável `SAC50UD` contribui para um decréscimo de 0,138 toneladas na produção de resíduos indiferenciados, a variável `FRT` contribui para um aumento de 1,653 toneladas e, caso a entidade responsável não entregue sacos para reciclagem, de forma gratuita, aos utilizadores, correspondente à variável `OFERECE:1`, verifica-se um acréscimo de 12,930 toneladas na produção mensal de resíduos indiferenciados.

Em relação ao valor observado do coeficiente de determinação ajustado, R_a^2, conclui-se que 66,8% da variabilidade da produção de resíduos indiferenciados é explicada pelo respetivo modelo de regressão múltipla.

Tabela 7.2: Modelo de regressão linear múltipla para a produção de resíduos indiferenciados.

INDIFERENCIADO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = -2,713$	8,850	0,042
`MTH`	$\hat{\beta_1} = 1,118$	0,262	<0,001
`SAC50UD`	$\hat{\beta_2} = -0,138$	0,075	0,076
`FRT`	$\hat{\beta_3} = 1,653$	0,277	<0,001
`OFERECE:1`	$\hat{\beta_4} = 12,930$	4,658	0,010
$R_a^2 = 0,668$			

Neste caso, o modelo de regressão linear múltipla, pode ser expresso por

$$IND_t = \beta_0 + \beta_1 MTH_t + \beta_2 SAC50UD_t + \beta_3 FRT_t + \beta_4 OFERECE(1)_t + \epsilon_t \qquad (7.1)$$

onde $t = 1, \ldots, 33$ representa os meses, IND_t a quantidade de resíduos indiferenciados produzidos e ϵ_t é um erro estocástico.

7.2.2 Resíduos seletivos

Pela Tabela 7.3, conclui-se que as variáveis `MTH`, `SAC15UD`, `SAC30UD`, `UDC`, `TAC`, `TBC` , `TDC`, `TEC`, `KMS`, `FRT`, `REC30` e ,`REC100`, são significativas na explicação da produção de resíduos seletivos na zona em estudo.

A variável `MTH` contribui para um decréscimo de -0,440 toneladas na produção mensal de resíduos urbanos indiferenciados, que se representa por uma pequena tendência decrescente ao longo do tempo no período observado. Em relação aos sacos de resíduos indiferenciados, a variável `SAC15UD` e a variável `SAC30UD` provocam um decréscimo de -0,017 e -0,038 toneladas, respetivamente. Em relação aos utilizadores que compram sacos, por cada utilizador doméstico(`UDC`) verifica-se um aumento de 0,230 toneladas na produção de resíduos seletivos. Em relação aos utilizadores não domésticos, realça-se os da tipologia A (`TAC`) que contribuem para um decréscimo de 0,478 toneladas por cada utilizador que compra sacos. Já os das tipologias B, D e E contribuem para um aumento de 0,426, 1,105 e 0,793 toneladas, por cada utilizador, respetivamente.

Em relação à quilometragem e ao número de vezes que a viatura se desloca à estação de triagem, verifica-se que por cada incremento unitário na variável `KMS` a produção de resíduos seletivos aumenta 0,003 toneladas e na variável `FRT` aumenta 1,222 toneladas.

Os sacos para reciclagem de 30 litros, `REC30`, contribuem para um aumento de 0,007 toneladas na produção mensal por cada unidade adquirida. Já os sacos de 100 litros, `REC100`, contribuem para um decréscimo de -0,003 toneladas por cada unidade adquirida pelos utilizadores.

Em relação ao valor observado do coeficiente de determinação ajustado, R_a^2, conclui-se que 79,9% da variabilidade da produção de resíduos seletivos é explicada pelo respetivo modelo de regressão múltipla.

Tabela 7.3: Modelo de regressão linear múltipla para a produção de resíduos seletivos.

	SELETIVO		
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = -3,851$	5,298	0,042
MTH	$\hat{\beta_1} = -0,440$	0,188	0,030
SAC15UD	$\hat{\beta_2} = -0,017$	0,008	0,046
SAC30UD	$\hat{\beta_3} = -0,038$	0,013	0,008
UDC	$\hat{\beta_4} = 0,230$	0,118	0,065
TAC	$\hat{\beta_5} = -0,478$	0,227	0,048
TBC	$\hat{\beta_6} = 0,426$	0,225	0,073
TDC	$\hat{\beta_7} = 1,105$	0,282	0,001
TEC	$\hat{\beta_8} = 0,793$	0,304	0,017
KMS	$\hat{\beta_9} = 0,003$	0,001	0,016
FRT	$\hat{\beta_{10}} = 1,222$	0,244	<0,001
REC30	$\hat{\beta_{11}} = 0,007$	0,003	0,039
REC100	$\hat{\beta_{11}} = -0,003$	0,002	0,087
	$R_a^2 = 0,799$		

Assim, o modelo de regressão linear múltipla para os resíduos seletivos é dado por

$$\begin{aligned} SEL_t &= \beta_0 + \beta_1 MTH_t + \beta_2 SAC15UD_t + \beta_3 SAC30UD_t + \beta_4 UDC_t + \beta_5 TAC_t \\ &+ \beta_6 TBC_t + \beta_7 TDC_t + \beta_8 TEC_t + \beta_9 KMS_t + \beta_{10} FRT_t + \beta_{11} REC30_t \\ &+ \beta_{12} REC100_t + \epsilon_t. \end{aligned} \tag{7.2}$$

onde $t = 1, \ldots, 33$ representa os meses, SEL_t a quantidade de resíduos seletivos produzidos e ϵ_t é um erro estocástico.

Ainda sobre os resíduos seletivos e de uma forma mais discriminada, segundo as conclusões retiradas das Tabelas C.1, C.2 e C.3, no Apêndice C, os modelos de regressão linear múltipla para os resíduos de papel/cartão, vidro e plástico, respetivamente, são dados por

$$\begin{aligned} PAP_t &= \beta_0 + \beta_1 SAC15UD_t + \beta_2 SAC30UD_t + \beta_3 SAC30UND_t + \beta_4 UDC_t + \beta_5 TAC_t \\ &+ \beta_6 TDC_t + \beta_7 TEC_t + \beta_8 KMS_t + \beta_9 FRT_t + \beta_{10} OFERECE(1)_t \\ &+ \beta_{11} REC50_t + \epsilon_t, \end{aligned} \tag{7.3}$$

$$\begin{aligned} VID_t &= \beta_0 + \beta_1 SAC15UD_t + \beta_2 SAC30UD_t + \beta_3 SAC30UND_t + \beta_4 UDC_t + \beta_5 TDC_t \\ &+ \beta_6 FRT_t + \beta_7 DEP_t + \beta_8 REC30_t + \beta_9 REC100_t + \epsilon_t, \end{aligned} \tag{7.4}$$

$$\begin{aligned} PLA_t &= \beta_0 + \beta_1 MTH + \beta_2 SAC15UD_t + \beta_3 SAC30UD_t + \beta_4 UDC_t + \beta_5 TAC_t \\ &+ \beta_6 TDC_t + \beta_7 TEC_t + \beta_8 FRT_t + \beta_9 OFERECE(1)_t + \beta_{10} REC50_t \\ &+ \beta_{11} REC100_t + \epsilon_t, \end{aligned} \tag{7.5}$$

onde $t = 1, \ldots, 33$ representa os meses, PAP_t, VID_t e PLA_t representam a quantidade de resíduos de papel, vidro e plástico produzidos e ϵ_t é um erro estocástico.

Em relação aos modelos apresentados nas equações 7.3, 7.4 e 7.5 relativos à produção de papel/cartão, plástico e vidro, é importante referir que apenas o modelo ajustado aos dados da produção de resíduos de plástico apresenta presença de componente de tendência, onde se verifica um decréscimo de -0,048 toneladas na produção mensal de resíduos de plástico.

Na análise da Tabela C.1 (Apêndice C) verifica-se que a covariável que mais contribui para o valor médio da produção de resíduos de papel/cartão é `OFERECE:1`, de onde se conclui que a não entrega gratuita de sacos para a reciclagem influencia a produção mensal de resíduos de papel/cartão em cerca de 2,643 toneladas. O coeficiente de determinação ajustado tem um valor igual a 0,526, isto é, cerca de 52,6% da variabilidade da produção de resíduos de papel/cartão é explicada pelo modelo definido em 7.3.

Em relação à produção de resíduos de vidro, a covariável `FRT` tem coeficiente em valor absoluto elevado, que permite concluir que é a que mais contribui para o valor esperado da produção de resíduos de vidro, ou seja, a produção de resíduos de vidro aumenta cerca de 0,705 toneladas com o acréscimo unitário ao número de fretes. O coeficiente de determinação ajustado tem um valor igual a 0,857, isto é, cerca de 85,7% da variabilidade da produção de resíduos de vidro é explicada pelo modelo definido em 7.4.

Por último, a covariável `FRT` tem coeficiente em valor absoluto elevado de valor igual a 0,273, de onde se constanta que é a que mais contribui para o valor esperado da produção de resíduos de plástico, ou seja, a produção de resíduos de plástico aumenta cerca de 0,273 toneladas com o acréscimo unitário ao número de fretes. O coeficiente de determinação ajustado tem um valor igual a 0,877, isto é, cerca de 87,7% da variabilidade da produção de resíduos de vidro é explicada pelo modelo definido em 7.5.

7.2.3 Análise de resíduos

Para validar os modelos é necessário efetuar uma análise dos resíduos. De forma a cumprir os pressupostos das "Condições de Gauss-Markov", estes devem apresentar um comportamento próximo de uma distribuição Normal, de média nula e variância constante, e não apresentarem correlação temporal.

Relativamente aos resíduos do modelo de regressão obtido para os dados relativos à produção de resíduos indiferenciados, pela Figura 7.1, numa análise visual poderemos considerar os resíduos como gaussianos devido ao seu comportamento simétrico no histograma e pela proximidade dos pontos, no *QQ plot*, à reta $y = x$ respeitante à bissetriz dos quadrantes ímpares (1.º quadrante). Complementando esta conclusão, o teste de Shapiro Wilk para a normalidade não rejeita a hipótese de normalidade (valor de prova de 0,426).

Analisando o gráfico de dispersão dos resíduos *versus* valores estimados (Figura 7.1), é notória a distribuição uniforme em torno do resíduo zero, o que leva a crer que os erros têm média nula e variância constante. De facto, o teste t para o valor médio confirma

a não rejeição da hipótese de média nula, resultando num valor de prova próximo de 1. Assim, conclui-se que não há evidência estatística para rejeitar a hipótese de que os erros seguem uma distribuição Normal de média nula e variância constante.

Quanto à independência, foi calculada a estatística de Durbin-Watson dos respetivos resíduos do modelo onde se obteve um valor igual a 1,778 (valor próximo de 2) que leva à não rejeição da independência dos resíduos. Este resultado pode ser também verificado pela representação gráfica da função de autocorrelação (FAC), onde não se observam correlações de valor significativo.

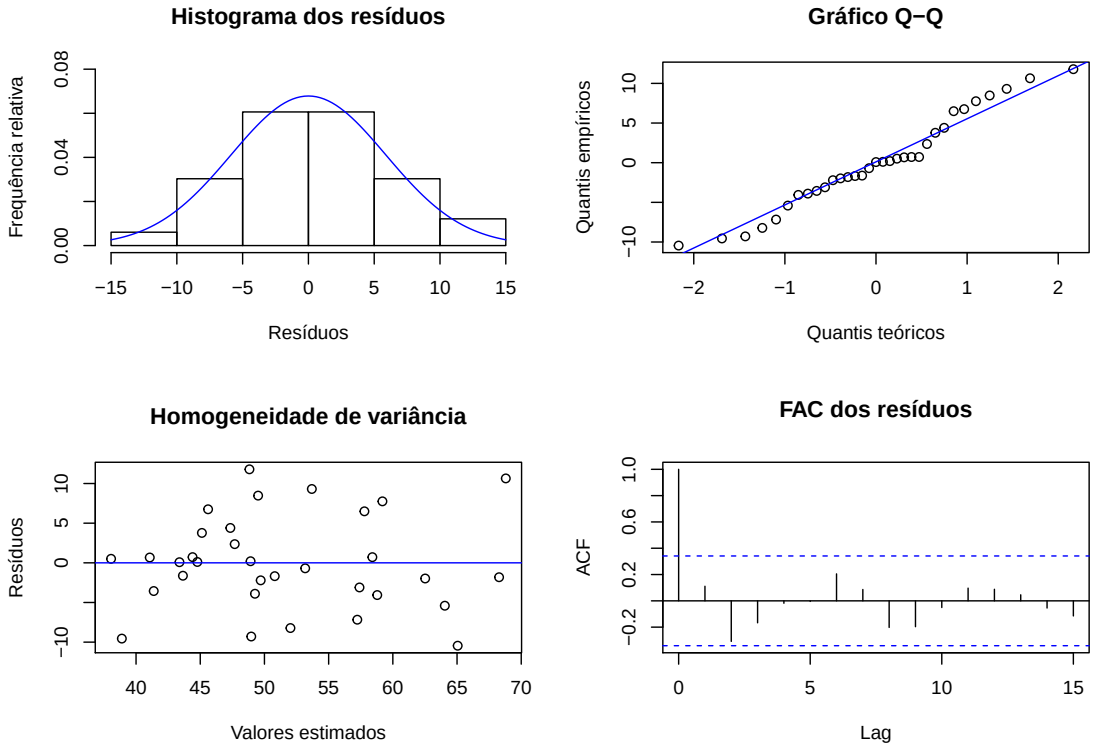


Figura 7.1: Histograma e *QQ plot* dos resíduos do modelo obtido para os resíduos indiferenciados.

Respeitante ao modelo obtido nos dados relativos à produção de resíduos seletivos, a partir da Figura 7.2, podemos afirmar que os resíduos, de uma perspetiva gráfica, não apresentam um comportamento semelhante ao de uma distribuição Normal devido à assimetria verificada no histograma apresentado e pelo afastamento das observações, no *QQ plot*, da reta $y = x$, correspondente à bissetriz dos quadrantes ímpares (1.º quadrante), à medida que o valor do quantil teórico aumenta. Numa perspetiva analítica, pelo teste de Shapiro-Wilk para a normalidade rejeita-se a hipótese de normalidade para um nível de significância de 10% (valor de prova de 0,076).

Para verificar o pressuposto de homocedasticidade da variância, analisando o gráfico de dispersão dos resíduos *versus* valores estimados, é notória a distribuição uniforme em torno do resíduo zero, o que leva a crer que os erros têm média nula e variância constante.

Novamente, numa perspetiva analítica de forma a reforçar as conclusões retiradas, no teste*t* para o valor médio confirma-se a não rejeição da hipótese de média nula, resultando num valor de prova próximo de 1. Com isto, conclui-se que não há evidência estatística para rejeitar a hipótese de que os erros seguem uma distribuição Normal de média nula e variância constante.

Para verificar a independência dos resíduos, pela estatística de Durbin-Watson, foi obtido um valor de 2,273 (valor próximo de 2), que leva à não rejeição da independência dos resíduos. Este resultado pode ser também verificado pela representação gráfica da função de autocorrelação (FAC) onde não se verificam correlações de valor significativo.

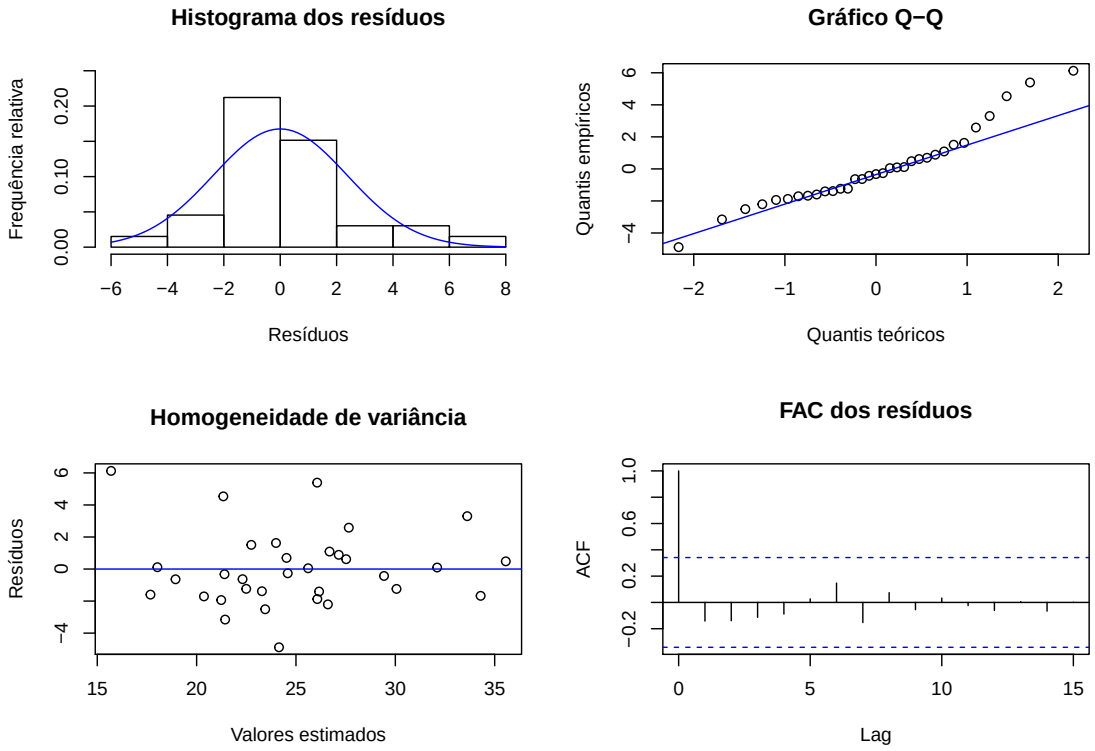


Figura 7.2: Histograma e *QQ plot* dos resíduos do modelo obtido para os resíduos seletivos.

7.3 Modelos de Regressão para a sazonalidade

Diz-se que os dados apresentam um comportamento sazonal quando os fenómenos que ocorrem ao longo do tempo se repetem em cada período idêntico de tempo, ou seja, fenómenos que ocorrem diariamente numa certa hora, todos os dias, ou num certo mês todos os anos. Como foi descrito anteriormente, a componente de sazonalidade (S_t) representa uma variabilidade periódica. Estas variações caraterizadas por aumentos ou decréscimos dos valores observados, ocorridos de forma regular que origina oscilações que se repetem, podem ser verificadas a partir da modelação dos dados conforme o procedimento descrito na Secção 3.2. Numa fase inicial foram formulados os modelos respetivos às quantidades de resíduos indiferenciados e seletivos, respetivamente, e considerando, exclusivamente, como covariáveis os indicadores sazonais. Obtendo noções preliminares sobre o comportamento sazonal dos resíduos indiferenciados e seletivos, numa perspetiva mensal, foram novamente formulados modelos de regressão linear múltipla, após aplicação do método regressivo, resultantes da combinação das variáveis dos indicadores sazonais com as covariáveis significativas obtidas nos modelos de regressão linear múltipla na Secção 7.2, para os dois tipos de resíduos.

Considera-se como mês de referência o mês de maio. Este mês também não foi estatisticamente significativo quando foi integrado no modelo e considerando outro mês de referência.

Por último, numa perspetiva exploratória foram elaborados modelos de regressão linear múltipla, após o método regressivo para seleção de variáveis, resultado da combinação das covariáveis respeitantes aos indicadores sazonais com as covariáveis significativas nos modelos de regressão linear simples, conforme se verifica na Tabela 7.1.

7.3.1 Resíduos indiferenciados

De forma a modelar a sazonalidade associada à produção mensal de resíduos indiferenciados, define-se o seguinte modelo

$$\begin{aligned} IND_t &= \beta_0 + \beta_1(ABR)_t + \beta_2(JUN)_t + \beta_3(JUL)_t + \beta_4(AGO)_t + \beta_5(SET)_t \\ &+ \beta_6(OUT)_t + \beta_7(NOV)_t + \beta_8(DEZ)_t + \beta_9(JAN)_t + \beta_{10}(FEV)_t \\ &+ \beta_{11}MAR_t + \epsilon_t, \end{aligned} \tag{7.6}$$

onde $t = 1, \ldots, 33$ representa os meses, IND_t a quantidade de resíduos indiferenciados produzidos e ϵ_t é um erro estocástico.

O modelo definido pela equação 7.6 foi ajustado à serie dos dados relativos à produção de resíduos indiferenciados no CHG.

Posto isto, a Tabela C.4, no Apêndide C, apresenta as estimativas dos coeficientes, valores de prova e respetivo coeficiente de determinação do modelo.

Uma vez que se deteta a existência de coeficientes sazonais estatisticamente não sig-

nificativos no modelo 7.6, é aplicado o método *backward* ao modelo inicial, de forma a encontrar o modelo que se ajuste melhor aos dados. Finalmente, é obtido o seguinte modelo:

$$IND_t = \beta_0 + \beta_1(ABR)_t + \beta_2(JUN)_t + \beta_3(AGO)_t + \beta_4(FEV)_t + \epsilon_t, \qquad (7.7)$$

onde $t = 1, \ldots, 33$ representa os meses, IND_t a quantidade de resíduos indiferenciados produzidos e ϵ_t é um erro estocástico.

A Tabela 7.4 apresenta as estimativas dos coeficientes, erros padrão, valores de prova e o coeficiente de determinação. De facto, é notório que existe um decréscimo significativo nos meses de abril e fevereiro iguais a 7,595 e 15,879 toneladas, respetivamente. Já nos meses de junho e agosto verificam-se aumentos de valor igual a 17,578 e 7,031 toneladas, respetivamente. O modelo final, via método regressivo, detém um coeficiente de determinação igual a 0,524 , ou seja, cerca de 52,4% da variabilidade dos dados é explicada pela componente sazonal destes meses.

Tabela 7.4: Valores obtidos após modelação do modelo sazonal final, dos resíduos indiferenciados.

INDIFERENCIADO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = 51,147$	1,319	<0,001
ABR	$\hat{\beta_1} = -7,595$	3,936	0,064
JUN	$\hat{\beta_2} = 17,578$	3,936	<0,001
AGO	$\hat{\beta_3} = 7,031$	3,936	0,085
FEV	$\hat{\beta_4} = -15,879$	4,569	0,002
$R_a^2 = 0,524$			

Aprofundando a análise relativa à sazonalidade, foi formulado um modelo de regressão linear múltipla combinando as variáveis obtidas no modelo de regressão linear múltipla, respeitante aos resíduos indiferenciados (Tabela 7.2), com a totalidade das variáveis respeitantes à sazonalidade, que foram estatisticamente significativas.

Após a aplicação do método *backward*, foi obtido o seguinte modelo

$$\begin{aligned} IND_t = \quad & \beta_0 \quad + \beta_1(MTH)_t + \beta_2(SAC50UD)_t + \beta_3(FRT)_t + \beta_4(OFERECE:1)_t + \\ & \beta_5 \quad (JUN)_t + \beta_6(FEV)_t + \beta_7(MAR)_t + \epsilon_t, \end{aligned} \qquad (7.8)$$

onde $t = 1, \ldots, 33$ representa os meses, IND_t a quantidade de resíduos indiferenciados produzidos e ϵ_t é um erro estocástico.

Na Tabela 7.5 são apresentados os valores das estimativas dos coeficientes, erros padrão e respetivos valores de prova. De facto, verifica-se que as variáveis constantes no modelo de equação 7.1 mantêm-se, sendo de realçar a alteração das variáveis sazonais, isto é, a saída dos meses de abril e agosto e a entrada do mês de março. Estas alterações devem-se

às relações entre as variáveis com a finalidade de explicar a variabilidade da produção mensal de resíduos indiferenciados.

Desta forma, com o modelo ajustado resultante da combinação das variáveis referidas, é verificada uma tendência crescente dos resíduos indiferenciados pela variável `MTH` onde, em cada mês acrescido, se verifica um aumento de 1,030 toneladas, na medida que é verificada uma tendência crescente na produção de resíduos indiferenciados. Já os sacos de 50 litros, adquiridos pelos UND (`SAC50UND`, contribuem para uma diminuição de 0,157 toneladas por cada saco vendido. Os fretes (`FRT`) e a não entrega de sacos de reciclagem de forma gratuita (`OFERECE:1`) contribuem para um acréscimo de 0,932 e 9,183 toneladas, respetivamente.

A produção de resíduos indiferenciados aumenta cerca de 14,418 toneladas aquando no mês de junho e descresce cerca de 9,183 e 7,291 toneladas nos meses de fevereiro e março, respetivamente.

De realçar o coeficiente de determinação que é igual a 0,814 que leva a afirmar que cerca de 81,4% da variabilidade da produção de resíduos indiferenciados é explicada pelo modelo.

Tabela 7.5: Modelo de regressão linear múltipla para a produção de resíduos indiferenciados com combinação das variáveis sazonais.

INDIFERENCIADO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = 14,556$	6,981	0,047
`MTH`	$\hat{\beta_1} = 1,030$	0,186	<0,001
`SAC50UD`	$\hat{\beta_2} = -0,157$	0,054	0,007
`FRT`	$\hat{\beta_3} = 0,932$	0,234	<0,001
`OFERECE:1`	$\hat{\beta_4} = 9,180$	3,346	0,011
`JUN`	$\hat{\beta_5} = 14,418$	2,747	<0,001
`FEV`	$\hat{\beta_6} = -9,183$	2,790	0,003
`MAR`	$\hat{\beta_7} = -7,291$	2,718	0,013
$R_a^2 = 0,814$			

Partindo para outra abordagem, considerando as covariáveis significativas após aplicação do modelo de regressão linear simples aos dados, como se pode verificar através da Tabela 7.1 é pretendida a formulação de um modelo de regressão linear múltipla que resulte da combinação das covariáveis supra mencionadas com as variáveis indicatrizes correspondentes à sazonalidade. O processo de modelação, tal como nos anteriores, consiste na utilização do método regressivo para a seleção das variáveis que melhor expliquem a variável resposta. Desta forma é obtido o seguinte modelo

$$\begin{aligned} IND_t &= \beta_0 + \beta_1(MTH)_t + \beta_2(SAC15UD)_t + \beta_3(JUN)_t + \beta_4(JUL)_t + \beta_5(AGO)_t \\ &+ \beta_6(DEZ)_t + \beta_7(FEV)_t + \beta_8(MAR)_t + \epsilon_t, \end{aligned} \tag{7.9}$$

onde $t = 1, \ldots, 33$ representa os meses, IND_t a quantidade de resíduos indiferenciados produzidos e ϵ_t é um erro estocástico.

Os principais resultados do modelo formulado estão apresentados na Tabela 7.6. Deste novo modelo retém-se um menor número de variáveis das que foram obtidas por regressão linear simples, sendo que apenas foram significativas as variáveis `MTH` e `SAC15UD`. Desta forma verifica-se uma tendência crescente na produção de resíduos indiferenciados na ordem das 0,354 toneladas por cada mês. Em relação à variável correspondente ao número de sacos de 15 litros, adquiridos por UD, verifica-se um aumento de 0,015 toneladas por cada incremento unitário na venda de sacos.

Importa realçar que as variáveis indicatrizes correspondentes à incorporação da sazonalidade no modelo têm um grande poder explicativo sobre a quantidade de resíduos indiferenciados, produzidos mensalmente, devido aos coeficientes em valor absoluto elevado.

Neste modelo, os meses de junho, julho e agosto contribuem de forma positiva para o aumento da produção de resíduos indiferenciados mensal, com valores iguais a 18,537, 7,441 e 6,107 toneladas, respetivamente. Já os meses de fevereiro, março e dezembro contribuem de forma negativa para a produção de resíduos indiferenciados, com valores iguais a 4,750, 15,968 e 12,728 toneladas, respetivamente, sendo expectável que haja uma maior taxa de reciclagem nestes meses.

Este modelo apresenta um coeficiente de determinação ajustado, R_a^2, igual a 0,843 , ou seja, cerca de 84,3% da variabilidade da produção de resíduos indiferenciados é explicada pelo modelo.

Tabela 7.6: Modelo de regressão linear múltipla para a produção de resíduos indiferenciados com combinação das variáveis sazonais e variáveis obtidas por regressão linear simples.

`INDIFERENCIADO`			
	Estimativas	$\hat{\sigma}$	**valor de prova**
	$\hat{\beta}_0 = 39,147$	2,114	<0,001
`MTH`	$\hat{\beta}_1 = 0,354$	0,131	0,012
`SAC15UD`	$\hat{\beta}_2 = 0,015$	0,008	0,066
`JUN`	$\hat{\beta}_3 = 18,537$	2,534	<0,001
`JUL`	$\hat{\beta}_4 = 7,441$	2,611	**0,009**
`AGO`	$\hat{\beta}_5 = 6,107$	2,556	0,025
`DEZ`	$\hat{\beta}_6 = -4,750$	2,562	0,076
`FEV`	$\hat{\beta}_7 = -15,968$	2,999	<0,001
`MAR`	$\hat{\beta}_8 = -12,728$	3,251	0,001
$R_a^2 = 0,651$			

7.3.2 Resíduos seletivos

Analisando agora os dados relativos às quantidades de resíduos seletivos produzidos, define-se o seguinte modelo

$$\begin{aligned} SEL_t &= \beta_0 + \beta_1(ABR)_t + \beta_2(JUN)_t + \beta_3(JUL)_t + \beta_4(AGO)_t + \beta_5(SET)_t \\ &+ \beta_6(OUT)_t + \beta_7(NOV)_t + \beta_8(DEZ)_t + \beta_9(JAN)_t + \beta_{10}(FEV)_t \\ &+ \beta_{11}(MAR)_t + \epsilon_t, \end{aligned} \tag{7.10}$$

onde $t = 1, \ldots, 33$ representa os meses, SEL_t a quantidade de resíduos seletivos produzidos e ϵ_t é um erro estocástico.

De forma análoga aos resíduos indiferenciados, o modelo definido pela equação 7.10 foi ajustado à série dos dados relativos à produção de resíduos seletivo no CHG. Assim, a Tabela C.4 (Apêndice C) apresenta as estimativas dos coeficientes, valores de prova e respetivo coeficiente de determinação do modelo.

A existência de coeficientes sazonais estatisticamente não significativos no modelo 7.10, leva a aplicar o método regressivo ao modelo inicial, de forma a encontrar o modelo que se ajuste melhor aos dados. Após a aplicação do referido método, foi obtido o seguinte modelo

$$SEL_t = \beta_0+\beta_1(ABR)_t+\beta_2(JUL)_t+\beta_3(AGO)_t+\beta_4(SET)_t++\beta_5(DEZ)_t+\beta_6(FEV)_t+\epsilon_t, \tag{7.11}$$

onde $t = 1, \ldots, 33$ representa os meses, SEL_t a quantidade de resíduos seletivos produzidos e ϵ_t é um erro estocástico.

A Tabela 7.7 apresenta as estimativas dos coeficientes, erros padrão, valores de prova e o coeficiente de determinação. Verifica-se que existe um decréscimo significativo nos meses de abril, dezembro e fevereiro iguais a 4,383, 5,250 e 6,880 toneladas, respetivamente. Já nos meses de julho, agosto e setembro verificam-se aumentos de valor igual a 4,890, 8,450 e 3,670 toneladas, respetivamente.

O modelo final, via método regressivo, detém um coeficiente de determinação ajustado, R^2_a, igual a 0,621 , ou seja, cerca de 62,1% da variabilidade da produção de resíduos seletivos é explicada pelas variáveis sazonais.

De forma análoga aos resíduos indiferenciados, também foi formulado um modelo resultante da combinação das variáveis constantes da equação 7.3 com as variáveis sazonais.

Tabela 7.7: Valores obtidos após modelação do modelo sazonal final, dos resíduos seletivos.

SELETIVO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta}_0 = 24,709$	0,636	<0,001
ABR	$\hat{\beta}_1 = -4,383$	1,952	0,033
JUL	$\hat{\beta}_2 = 4,890$	1,952	0,012
AGO	$\hat{\beta}_3 = 8,457$	1,952	<0,001
SET	$\hat{\beta}_4 = 3,637$	1,952	0,074
DEZ	$\hat{\beta}_5 = -5,250$	1,952	0,012
FEV	$\hat{\beta}_6 = -6,880$	2,310	0,006
$R_a^2 = 0,621$			

Desta forma, atendendo à Tabela 7.8 verifica-se que, após aplicação do método regressivo no modelo completo, as variáveis correspondente aos meses (MTH) e ao número de sacos de 30 litros oferecidos (REC30) não foram incluídas e, também, a remoção das variáveis indicatrizes correspondentes aos meses de julho JUL, dezembro DEZ e fevereiro FEV, pelo que no modelo obtido são consideradas as variáveis respeitantes aos meses de abril ABR, agosto AGO e setembro SET.

Interpretando os valores das estimativas dos coeficientes tem-se que os sacos de 15 litros adquirido por UD (SAC15UD) provocam um decréscimo de -0,016 toneladas por cada saco adquirido. Também os resíduos seletivos decrescem cerca de -0,010 e -0,004 toneladas por cada saco de 30 litros adquirido por UD e por cada saco de reciclagem de 100 litros oferecido aos utilizadores. Já a diversidade de tipologias que compõem os UND influenciam a produção de resíduos seletivos. Neste caso, é verificado um aumento de 0,474, 0,539 e 0,740 toneladas por cada aumento unitário dos utilizadores das tipologias B, D e E que compram sacos. A quilometragem também influencia as pesagens, concluindo-se que por cada quilómetro percorrido são recolhidas 0,003 toneladas de resíduos seletivos. Em relação à sazonalidade é verificado um decréscimo de 6,220 toneladas nas pesagens de resíduos seletivos, quando se trata do mês de abril e aumentos de 4,004 e 2,810 toneladas, aquando os meses de agosto e setembro, respetivamente.

O modelo, agora formulado, apresenta um coeficiente de determinação ajustado, R_a^2 igual a 0,714 , o que determina que 71,41% da variabilidade da produção mensal de resíduos seletivos é explicada pelo modelo. Desta forma, a equação do modelo é dada por

$$\begin{aligned} SEL_t &= \beta_0 + \beta_1(SAC15UD)_t + \beta_2(SAC30UD)_t + \beta_3(TBC)_t + \beta_4(TDC)_t \\ &+ \beta_5(TEC)_t + \beta_6(KMS)_t + \beta_7(FRT)_t + \beta_8(REC100)_t + \beta_9(ABR)_t \\ &+ \beta_{10}(AGO)_t + \beta_{11}(SET)_t + \epsilon_t, \end{aligned} \tag{7.12}$$

onde $t = 1, \ldots, 33$ representa os meses, SEL_t a quantidade de resíduos seletivos produzidos e ϵ_t é um erro estocástico.

Tabela 7.8: Modelo de regressão linear múltipla para a produção de resíduos seletivos com combinação das variáveis sazonais.

	SELETIVO		
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = -1,419$	5,260	0,095
SAC15UD	$\hat{\beta_1} = -0,016$	0,005	0,003
SAC30UD	$\hat{\beta_2} = -0,010$	0,006	0,094
TBC	$\hat{\beta_3} = 0,474$	0,174	0,013
TDC	$\hat{\beta_4} = 0,539$	0,234	0,032
TEC	$\hat{\beta_5} = 0,740$	0,281	0,016
KMS	$\hat{\beta_6} = 0,003$	0,001	0,006
FRT	$\hat{\beta_7} = 0,651$	0,180	0,002
REC100	$\hat{\beta_8} = -0,004$	0,001	0,026
ABR	$\hat{\beta_9} = -6,220$	1,740	0,002
AGO	$\hat{\beta_{10}} = 4,004$	1,980	0,056
SET	$\hat{\beta_{11}} = 2,810$	1,511	0,077
	$R_a^2 = 0,714$		

Partindo para outra metodologia, considerando as variáveis significativas após aplicação do modelo de regressão linear simples às duas variáveis resposta, como se pode verificar através da Tabela 7.1 é pretendida a formulação de um modelo de regressão linear múltipla que resulte da combinação das covariáveis estatisticamente significativas com as covariáveis correspondentes aos indicadores sazonais. Tal como nos anteriores é utilizado o método regressivo para seleção das variáveis que melhor expliquem a variável resposta. Desta forma é obtido o seguinte modelo

$$\begin{aligned} SEL_t &= \beta_0 + \beta_1(TBC)_t + \beta_2(ABR)_t + \beta_3(JUL)_t + \beta_4(AGO)_t \\ &+ \beta_5(SET)_t + \beta_6(NOV)_t + \beta_7(DEZ)t + \beta_8(FEV)_t + \epsilon_t, \end{aligned} \tag{7.13}$$

onde $t = 1, \ldots, 33$ representa os meses, SEL_t a quantidade de resíduos indiferenciados produzidos e ϵ_t é um erro estocástico.

Os principais resultados do modelo formulado estão apresentados na Tabela 7.9. Deste novo modelo, verifica-se uma maior retenção do número de variáveis sazonais. Em relação às variáveis restantes existe, apenas, uma variável significativa, correspondente ao número de UND da tipologia B que por cada acréscimo ao seu número aumenta, consequentemente, cerca de 0,401 toneladas na produção de resíduos seletivos.

Destaca-se então o poder explicativo das variáveis indicatrizes dos indicadores sazonais na produção dos resíduos seletivos. Neste caso as variáveis correspondentes aos meses de julho (JUL), agosto (AGO) e setembro (SET) aumentam as pesagens de resíduos seletivos na ordem dos 4,171, 7,337 e 3,854 toneladas, respetivamente. Já os meses de abril ABR, novembro (NOV), dezembro (DEZ) e fevereiro (FEV) provocam uma diminuição da produção de

resíduos seletivos de valores iguais a 3,899, 3,934, 3,697 e 4,515 toneladas, respetivamente. De realçar que este modelo tem um coeficiente de determinação ajustado igual a 0,614, ou seja, cerca de 61,4% da variabilidade da produção de resíduos seletivos é explicada pelo modelo.

Tabela 7.9: Modelo de regressão linear múltipla para a produção de resíduos seletivos com combinação das variáveis sazonais e variáveis obtidas por regressão linear simples.

	SELETIVO		
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = 15,158$	4,680	0,003
TBC	$\hat{\beta_1} = 0,401$	0,195	0,050
ABR	$\hat{\beta_2} = -3,899$	1,792	0,040
JUL	$\hat{\beta_3} = 4,171$	1,875	0,036
AGO	$\hat{\beta_4} = 7,337$	1,942	0,001
SET	$\hat{\beta_5} = 3,854$	1,794	0,042
NOV	$\hat{\beta_6} = -3,934$	1,827	0,042
DEZ	$\hat{\beta_7} = -3,697$	1,873	0,060
FEV	$\hat{\beta_8} = -4,515$	2,290	0,060
	$R_a^2 = 0,614$		

Capítulo 8

Aplicação de Métodos de Previsão

Depois de efetuada a análise exploratória aos dados facultados pela VITRUS, o passo seguinte é a aplicação das metodologias aos dados, que, neste caso, correspondem a modelos de previsão para séries temporais. Muitos estudos fundamentam-se, essencialmente pela metodologia de Box-Jenkins que será aplicada aos dados referentes à recolha de contentores de profundidade e aos dados relacionados com a recolha seletiva e indiferenciada, respetivamente, de resíduos na área de implementação do sistema PAYT.

A formulação do modelo SARIMA compreende as três fases constantes na metodologia Box-Jenkins: identificação, estimação e diagnóstico. Para o primeiro passo é necessária a estacionarização da série, por meio de transformações apropriadas, tanto em relação à média como à variância. Desta forma, e como a estabilização da variância deve ser efetuada em primeiro lugar, procede-se à transformação logarítmica (mais usual) dos dados. Relativamente à ordem de diferenciação, com vista à estabilização da média, esta é escolhida e fundamentada com base na análise gráfica (da série e das FAC e FACP empíricas) e nos testes de estacionariedade (ADF e KPSS).

Efetuadas as devidas transformações à série em estudo numa série estacionária, identificando a ordem de diferenciação regular, d, é necessário estimar a componente sazonal. Para tal, começa-se por determinar o período sazonal, s, através da análise das FAC e FACP da série estacionária e, de seguida, estimam-se vários modelos, fazendo variar os valor de P, D, e Q (ordens da parte sazonal), usualmente entre 0 e 1. A escolha das ordens, P, D e Q tem sempre em consideração a significância dos parâmetros associados e no critério AIC. Saliente-se que, cado o intervalo de confiança associado a $P = 1$ incluir o valor 1, se deve optar por uma diferenciação sazonal, ou seja, pelo modelo que considera $P = 0$ e $D = +1$.

Por último, identificam-se as ordens p e q, comparando o comportamento das FAC e FACP empíricas com o das FAC e FACP teóricas. Na medida que se deve realizar uma escolha mais cuidada, devem sempre ser explorados modelos "vizinhos", sendo estes analisado tanto em relação à significância dos parâmetros como ao comportamento dos resíduos. Quando em dúvida, a escolha entre dois ou mais modelos SARIMA fundamenta-

se no critério AIC, tendo sempre em mente que, se os AIC diferem em apenas duas unidades, se escolhe o modelo mais parcimonioso, ou seja, aquele com menor número de parâmetros.

A análise de resíduos, quando aplicável, tem como finalidade a verificação do comportamento dos resíduos e se estes se aproximam ao de um ruído branco. Desta forma é pretendido que se verifiquem, além da média nula e variância constante, a gaussianidade e a independência dos erros.

De forma a avaliar os pressupostos de média nula e variância constante, a representação gráfica dos resíduos é útil mas, para além disso, para a verificação da condição exigida à média, caso os pressupostos de independência e normalidade dos erros não sejam rejeitados, pode recorrer-se ao teste t para o valor médio. A condição de normalidade dos resíduos é valiada a partir de um histograma dos resíduos que deverá aproximar-se de um comportamento da função densidade de uma distribuição Normal. No entanto, de uma forma mais rigorosa, complementando à análise gráfica, podem utilizar-se testes estatísticos, sendo o mais comum, para amostras de grandes dimensões, o teste de Kolmogorov-Smirnov sob a hipótese nula da normalidade dos erros.

Para avaliar a independência dos erros, são utilizadas diferentes metodologias dependendo do método de previsão aplicado. De facto, quando se estima um modelo SARIMA, recomenda-se a utilização de um teste de Portmanteau, sendo um dos mais utilizados o teste de Ljung-Box, que testa se as primeiras k autocorrelações são simultaneamente nulas. Como tal, e caso de rejeição da hipótese nula conclui-se que o modelo escolhido não é apropriado.

Realça-se que, para todas as decisões, é considerado um nível de significância de 10%.

8.1 Caso I: Recolha de contentores de profundidade

Inicia-se o estudo com a identificação de um modelo SARIMA onde, o primeiro passo, consiste na estacionarização da série em estudo. Desta forma, após a estabilização da variância, através da transformação logarítimica é, então, necessário definir a ordem de diferenciação regular para a estabilização da média.

É aplicada uma diferenciação de 1.ª ordem ($d = 1$) onde, consequentemente, a série passa a ser estacionária em média (Figura 8.1). De forma a sustentar a afirmação, fundamentada com a análise gráfica, são aplicados dois testes de estacionariedade – o teste ADF e o teste KPSS – à série após diferenciação. O teste ADF é realizado, sob a hipótese nula de que a série não é estacionária, e conclui-se que a hipótese nula é rejeitada de onde se confirma que a série dos resíduos é estacionária enquanto no teste KPSS, a rejeição da hipótese nula implica a não estacionariedade da série. A escolha do número de *lags* para o teste ADF, ou seja, do valor de p, tem por base a regra proposta por Ng & Perron (1995). O mesmo número de *lags* é utilizado para o teste KPSS.

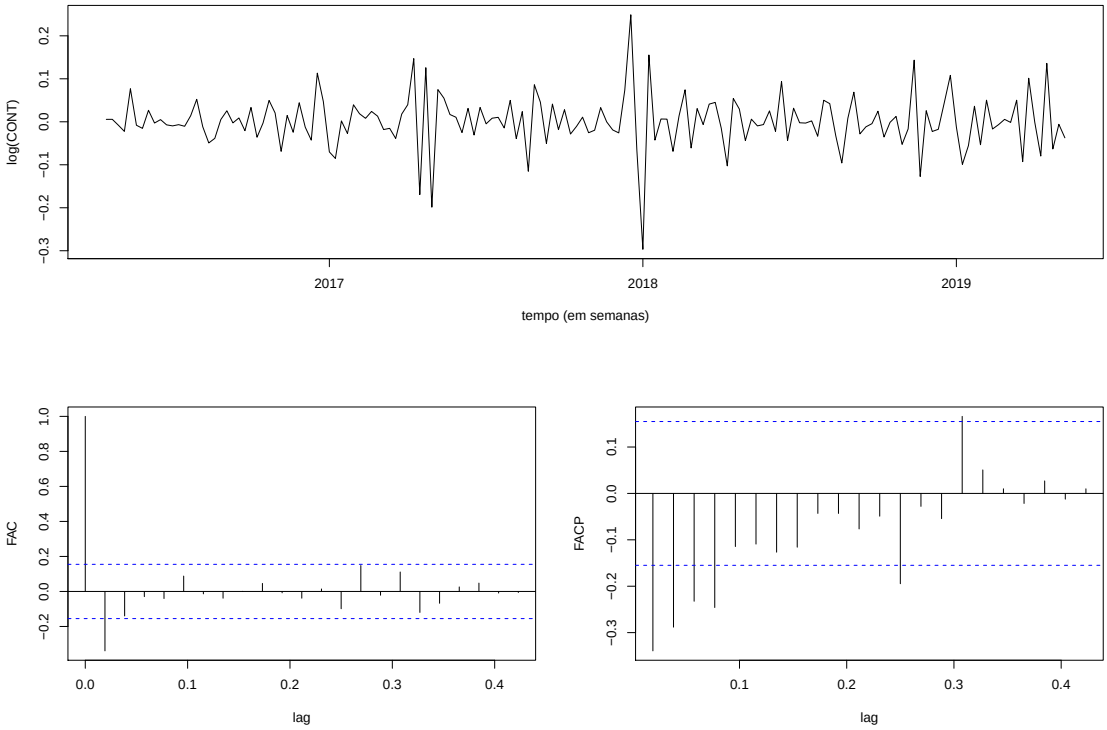


Figura 8.1: Série dos logaritmos das quantidades recolhidas em contentores de profundidade, após diferenciação de 1.ª ordem ($d = 1$), e respetivas FAC e FACP estimadas.

Conforme a distribuição associada a cada um dos testes, a estacionariedade é confirmada, em ambos os casos, se a estatística de teste for inferior ao valor crítico que, para um nível de significância de 10% é de -5,773 e 0,347 para os testes ADF e KPSS, respetivamente. Desta forma, conclui-se que, a um nível de significância de 10%, após uma diferenciação de 1.ª ordem, a série é estacionária em média.

Após a transformação da série original numa série estacionária em média e em variância, o passo seguinte consiste no ajuste da parte sazonal do modelo, que é verificada pela Figura 8.1 pela representação gráfica da série como na FAC correspondente. A periodicidade é, aparentemente, semanal, e, portanto, considera-se $s = \frac{365{,}25}{7} \approx 52,18 \approx 52$.

Estes valores relativos à sazonalidade são derivados do facto de existirem anos bissextos (366 dias) e com 53 semanas.Desta forma, este ajuste da sazonalidade irá permitir uma melhor formulação do modelo SARIMA. Neste estudo os anos estudados são considerados anos comuns, ou seja, anos com 365 dias.

A identificação das restantes ordens (P,D e Q) é explorada através da combinação de várias possibilidades, fazendo variar P,D e Q entre os valores de 0 e 1 (Tabela 8.1). De acordo com os resultados obtidos, o modelo que resulta no menor AIC é o que considera $P = 0$, $D = 0$ e $Q = 1$ e, desta forma, estas são as ordens escolhidas para a parte sazonal do modelo. De realçar que, estas ordens não são imutáveis e, de acordo com as necessidades futuras, estas poderão ser alteradas (aumentadas ou reduzidas).

Tabela 8.1: Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos das quantidades recolhidas em contentores de profundidade.

Modelo	$\hat{\nu}_1$	$\hat{\eta}_1$	**AIC**
SARIMA$(0,1,0)(1,0,0)_{52}$	0,123	-	-432,63
SARIMA$(0,1,0)(0,1,1)_{52}$	-	-1,000	-240,62
SARIMA$(0,1,0)(1,0,1)_{52}$	-0,592	0,739	-430,90
SARIMA$(0,1,0)(0,1,0)_{52}$	-	-	-217,81,
SARIMA$(0,1,0)(0,0,1)_{52}$	-	0,129	-432,71
SARIMA$(0,1,0)(1,1,0)_{52}$	-0,572	-	-238,28
SARIMA$(0,1,0)(1,1,1)_{52}$	-0,158	-0,998	-239,41

Verifique-se pela Figura 8.2 o comportamento dos resíduos após o ajustamento da parte sazonal.

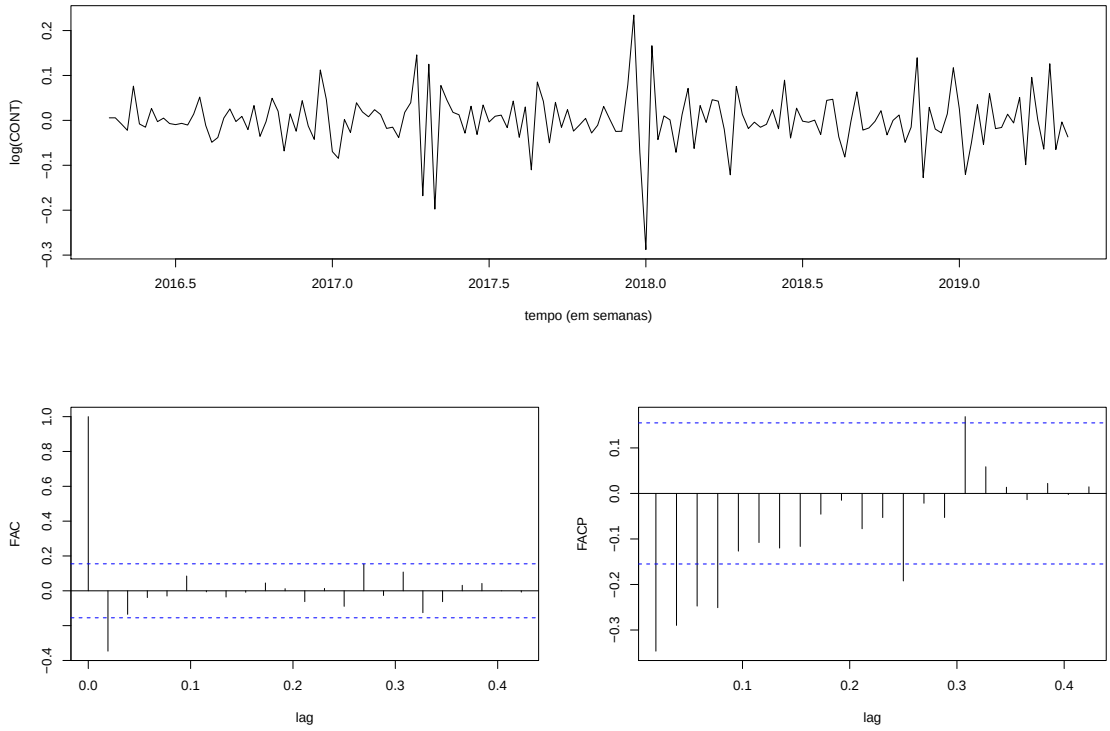


Figura 8.2: Série dos logaritmos das quantidades recolhidas em contentores de profundidade, após diferenciação de 1.ª ordem e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas.

Por último, na identificação de um modelo SARIMA, é necessário escolher as ordens p e q, através da comparação das FAC e das FACP empíricas com as FAC e FACP teóricas dos vários modelos conhecidos. De facto, de acordo com a Tabela 5.1, as FAC e FACP da Figura 8.2 sugerem o ajustamento de um modelo AR(3), ou, alternativamente, de um MA(1), ao resíduos obtidos após estimação da parte sazonal. Contudo, além destes modelos, são também ajustados modelos "vizinhos", de forma a realizar uma escolha mais

cuidada (Tabela 8.2). Ao analisar a Tabela 8.2, verifica-se que, para os modelos que consideram $P = 0$, $D = 0$ e $Q = 1$, os AIC são próximos em relação ao valor apresentado. De notar que quando $D = 1$ verifica-se um aumento significativo do AIC, levando à exclusão deste modelo. Relativamente ao modelo com um maior número de parâmetros, este inclui um coeficiente não significativo para o nível de significância considerado e, desta forma, não deverá ser considerado. Então, analisando os valores obtidos relativos aos modelos com igual número de parâmetros, seleciona-se o que detém menor AIC. Assim, o modelo escolhido é SARIMA(0,1,2)$(0,0,1)_{52}$.

Tabela 8.2: Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos das quantidades de resíduos indiferenciados em contentores de profundidade.

Modelo	$\hat{\phi_1}$	$\hat{\phi_2}$	$\hat{\phi_3}$	$\hat{\theta_1}$	$\hat{\theta_2}$	$\hat{\eta_1}$	**AIC**
SARIMA(0,1,1)$(0,0,1)_{52}$	–	–	–	-0,820	–	0,211	-488,30
SARIMA(1,1,1)$(0,0,1)_{52}$	0,177	–	–	-0,860	–	0,192	-490,18
SARIMA(1,1,0)$(0,0,1)_{52}$	-0.345	–	–	–	–	0,145	-450,96
SARIMA(0,1,0)$(0,0,1)_{52}$	–	–	–	–	–	0,129	-432,71
SARIMA(0,1,2)$(0,0,1)_{52}$	–	–	–	-0,659	-0,171	0,192	-490,96
SARIMA(1,1,2)$(0,0,1)_{52}$	-0,166	–	–	-0,500	-0,300	0,195	-489,15
SARIMA(2,1,0)$(0,0,1)_{52}$	-0,447	-0,288	–	–	–	0,155	-462,85
SARIMA(0,1,2)$(0,1,1)_{52}$	–	–	–	-0,855	0,009	-1,000	-294,64
SARIMA(3,1,1)$(0,0,1)_{52}$	0,127	-0,124	-0,093	-0,805	–	0,218	-489,04

* o coeficiente não é estatisticamente significativo, para um nível de significância $\alpha = 10\%$.

Os resultados da estimação do modelo escolhido podem ser consultados, em mais detalhe, na Tabela 8.3.

Tabela 8.3: Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos das quantidades de resíduos indiferenciados em contentores de profundidade.

Modelo final: SARIMA$(0,1,2)(0,0,1)_{52}$		AIC $= -490{,}96$	$\hat{\sigma}^2 = 0{,}002$
	θ_1	θ_2	η_1
estimativa	$-0{,}659$	$-0{,}171$	0,192
erro padrão	0,081	0,077	0,087

De forma a validar o modelo escolhido é necessário realizar uma análise dos resíduos. Estes devem apresentar, idealmente, um comportamento próximo de uma distribuição Normal, de média nula e variância constante, e não apresentar correlação temporal. O histograma da Figura 8.3 sugere, aparentemente, que os resíduos têm distribuição Normal, e de forma complelenta, o teste de Kolmogorov-Smirnov não rejeita a hipótese de normalidade (valor de prova de 0,215) dos resíduos. Também, de acordo com a representação gráfica da série dos resíduos (Figura 8.3) esta apresenta uma distribuição uniforme em

torno do resíduo zero, o que leva a concluir que os resíduos têm média nula e variância constante (homocedásticos). De facto, o teste t para o valor médio confirma a não rejeição da hipótese de média nula, resultando num valor de prova igual a 0,225. Desta forma, conclui-se que não existe evidência estatística para rejeitar a hipótese de que os erros seguem uma distribuição Normal de média nula e variância constante.

Quanto à independência, o teste de Ljung-Box é aplicado à série dos resíduos onde k varia entre 4 e 40 (k corresponde ao número de autocorrelações a serem testadas como grupo). Segundo os resultados do teste, a hipótese de independência não é rejeitada para nenhum dos valores de k, apresentando valores de prova entre 0,241 ($k = 4$) e 0,876 ($k = 40$). Realça-se que as FAC e FAC estimadas dos resíduos (Figura 8.3) assemelham-se às FAC e FACP de um ruído branco e, portanto, pode admitir-se a independência dos erros.

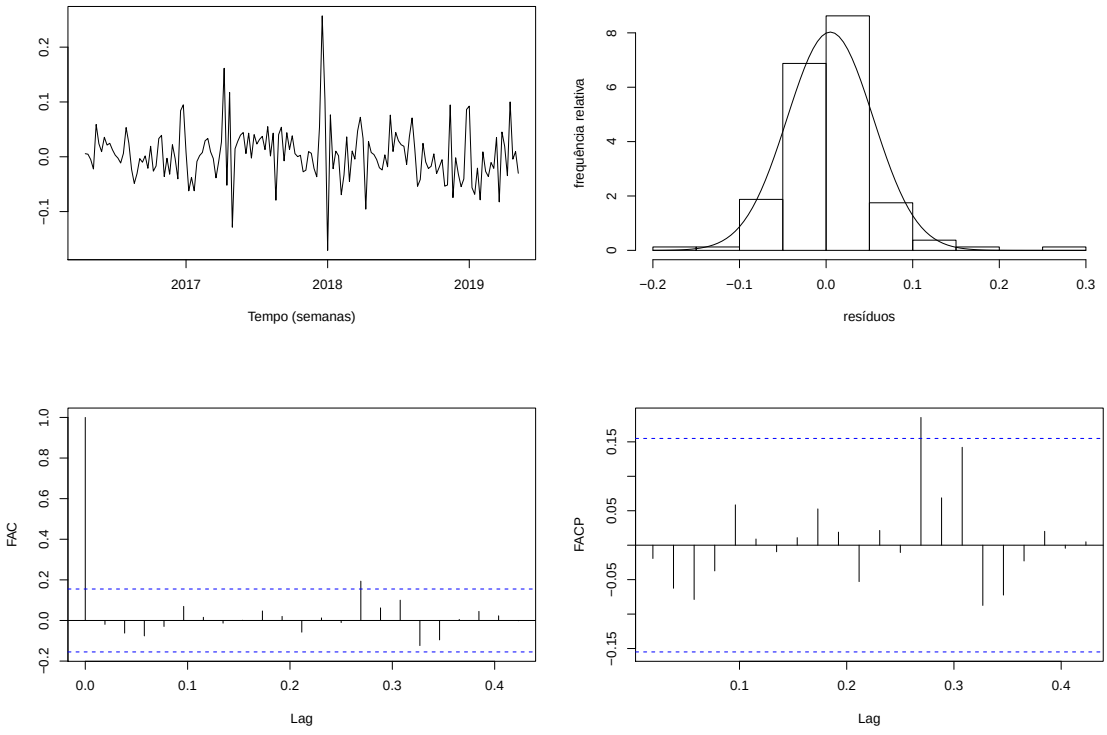


Figura 8.3: Série dos resíduos para a série dos logaritmos das quantidades de resíduos indiferenciados em contentores, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.

Na Figura 8.4 encontram-se representadas as previsões (no período de teste, isto é, da 19.ª semana de 2019 à 34.ª semana de 2019), pontuais e intervalares, e as estimativas pontuais (do período da 16.ª semana de 2016 à 18.ª semana de 2019) obtidas através do modelo final, nas unidades originais, sobrepostas à série em estudo.

A Figura 8.4 sugere que a qualidade preditiva do modelo é melhor na série de treino do que na série de teste, uma vez que se verifica uma descida atípica nas quantidades

de resíduos indiferenciados na série original que o modelo não seria capaz de explicar o fenómeno dadas as observações do passado.

Em relação aos intervalos e previsão, afirma-se que a sua taxa de cobertura é, neste caso, de 40%, uma vez que apenas 6 observações da série de teste pertencem ao interior dos mesmos.

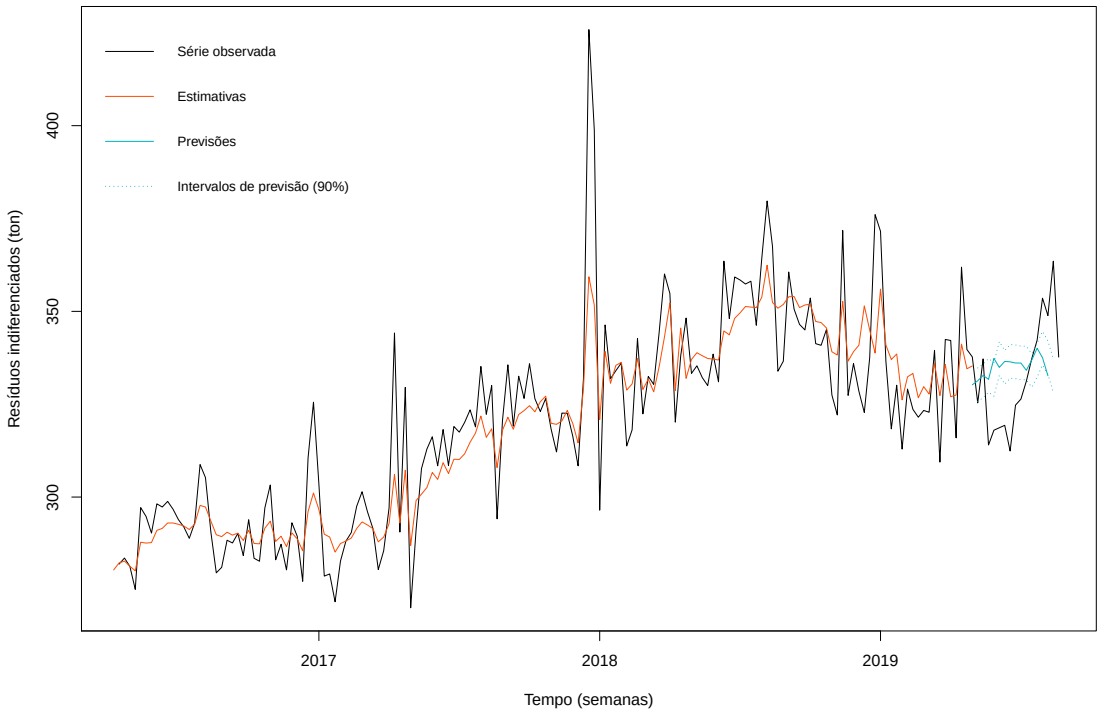


Figura 8.4: Previsões (no período de teste), pontuais e intervalares (90%), e estimativas pontuais (entre a 16.ª semana de 2016 e a 34.ª semana de 2019) obtidas através do modelo SARIMA, sobrepostas à série das quantidades de resíduos indiferenciados em contentores de profundidade.

8.2 Caso II: Recolha de resíduos em área PAYT

Resíduos seletivos

Em relação à serie temporal respeitante às quantidades de resíduos seletivos recolhidos, semanalmente, na zona piloto de implementação do sistema PAYT, inicia-se o estudo com a identificação de um modelo SARIMA onde, o primeiro passo, consiste na estacionarização da série em estudo. Desta forma, após a estabilização da variância, através da transformação logarítimica é, então, necessário definir a ordem de diferenciação regular para a estabilização da média. Numa perspetiva gráfica, através da análise da Figura 8.5 a série apresenta ser estacionária em média mas, de forma a sustentar esta afirmação, é necessário recorrer a uma análise com aplicação dos testes de estacionariedade - o teste ADF e o teste KPSS- à série dos logaritmos das quantidades de resíduos seletivos no CHG.

O teste ADF é realizado, sob a hipótese nula de que a série não é estacionária, e conclui-se que a hipótese nula é rejeitada de onde se confirma que a série dos resíduos é estacionária enquanto no teste KPSS, a rejeição da hipótese nula implica a não estacionariedade da série. A escolha do número de *lags* para o teste ADF, ou seja, do valor de p, tem como fundamento a regra proposta por Ng & Perron (1995). O número de *lags*, anteriormente calculado, é, também, utilizado para o teste KPSS.

De acordo com a distribuição associada a cada um dos testes, a estacionariedade é confirmada, em ambos os casos, se a estatística de teste for inferior ao valor crítico que, para um nível de significância de 10% é de -2,570 e 0,347 para os testes ADF e KPSS, respetivamente. Neste caso, para 3 *lags*, as estatísticas e teste são -3,647 e 0,143 Desta forma, conclui-se que, a um nível de significância de 10%, sem qualquer diferenciação aplicada, a série é estacionária em média.

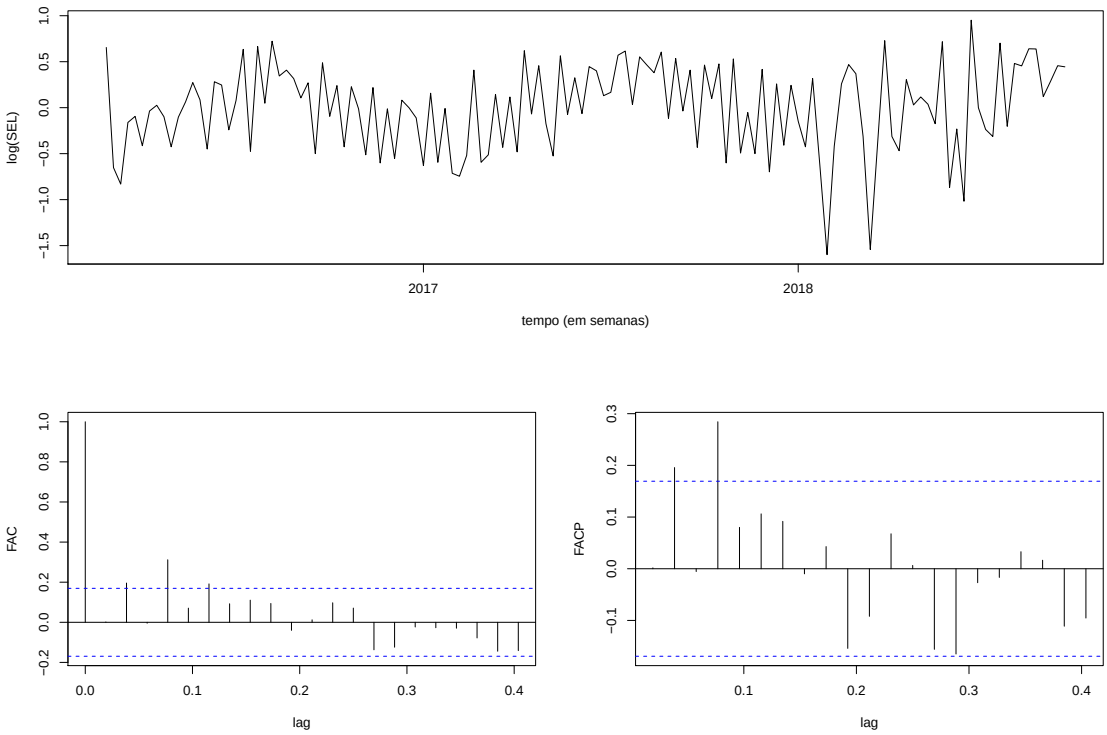


Figura 8.5: Série dos logaritmos das quantidades recolhidas de resíduos seletivos, no CHG, sem aplicação de diferenciação ($d = 0$), e respetivas FAC e FACP estimadas.

Com vista ao ajuste da parte sazonal do modelo, após análise da FAC correspondente à série logaritmizada (Figura 8.5), verifica-se que a sazonalidade é semanal e, por isso, considera-se $s = \frac{365{,}25}{7} \approx 52,18 \approx 52$. Como se verificou na modelação da série anterior, estes valores relativos à sazonalidade são derivados do facto de existirem anos bissextos (366 dias) e com 53 semanas. Desta forma, este ajustamento da sazonalidade irá permitir uma melhor formulação do modelo SARIMA. Neste estudo os anos estudados são considerados

anos comuns, isto é, anos com 365 dias.

A identificação das ordens da parte sazonal (P, D, Q) é analisada pela combinação das várias possibilidades, ou seja, trata-se de um processo iterativo, onde se varia P,D e Q entre 0 e 1 (Tabela 8.4). Posto isto, e de acordo com os resultados obtidos, o modelo que resulta no menor AIC é o que considera $P = 1$, $D = 1$ e $Q = 0$ sendo, portanto, estas as ordens selecionadas para a parte sazonal do modelo. Refira-se que estas não são imutáveis e que , de acordo com as necessidades futuras, podem ser alteradas.

Tabela 8.4: Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos das quantidades recolhidas em contentores de profundidade.

Modelo	$\hat{\nu}_1$	$\hat{\eta}_1$	**AIC**
SARIMA$(0,0,0)(1,0,0)_{52}$	0,224	–	179,65
SARIMA$(0,0,0)(0,1,1)_{52}$	–	-0,921	137,25
SARIMA$(0,0,0)(1,0,1)_{52}$	0,983	-0,898	179,42
SARIMA$(0,0,0)(0,1,0)_{52}$	–	–	144,14
SARIMA$(0,0,0)(0,0,1)_{52}$	–	0,183	180,45
SARIMA$(0,0,0)(1,1,0)_{52}$	-0,499	–	137,25
SARIMA$(0,0,0)(1,1,1)_{52}$	-0,595	0,129	139,25

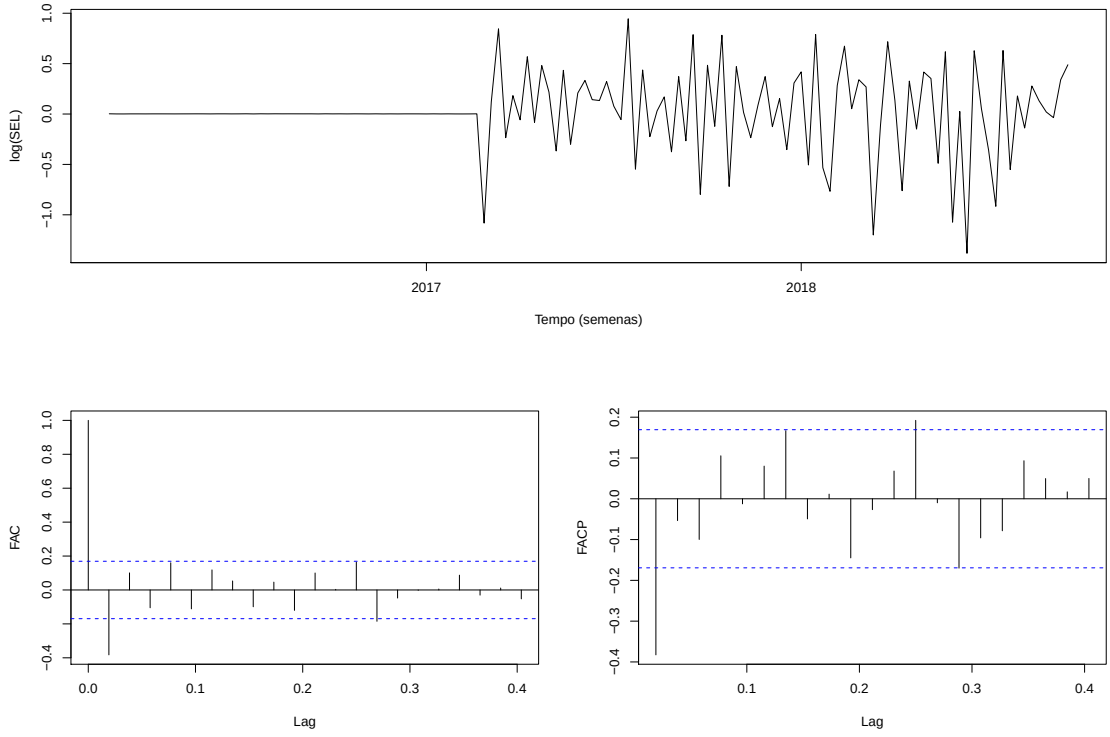


Figura 8.6: Série dos resíduos das quantidades recolhidas de resíduos seletivos, no CHG, sem diferenciação aplicada e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas.

Verifique-se pela Figura 8.6 o comportamento dos resíduos após o ajustamento da parte sazonal. De notar que numa fase inicial e numa perspetiva visual, os resíduos tendem a ter valor igual a 0, mas estes valores nunca atingem a sua nulidade de onde se conclui que os seus valores são muito reduzidos.

Por último, na identificação de um modelo SARIMA, é necessário escolher as ordens p e q, através da comparação das FAC e das FACP empíricas com as FAC e FACP teóricas dos vários modelos conhecidos. Conforme a Tabela 5.1, as FAC e FACP da Figura 8.6 sugerem o ajustamento de um modelo MA(1), ao resíduos obtidos após estimação da parte sazonal. Para além destes modelos, são também ajustados modelos "vizinhos", de forma a realizar uma escolha mais cuidada (Tabela 8.5).

Ao analisar a Tabela 8.5, verifica-se que, para os modelos que consideram $P = 0$, $D = 0$ e $Q = 1$, os AIC são próximos em relação ao valor apresentado. De notar que quando $D = 1$ verifica-se um aumento significativo do AIC, levando à exclusão deste modelo. Relativamente ao modelo com um maior número de parâmetros, este inclui um coeficiente não significativo para o nível de significância considerado e, desta forma, não deverá ser considerado. Então, analisando os valores obtidos relativos aos modelos com igual número de parâmetros, seleciona-se o que detém menor AIC. Assim, o modelo escolhido é SARIMA(0,0,1)$(1,1,1)_{52}$.

Tabela 8.5: Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos das quantidades de resíduos seletivos no CHG.

Modelo	$\hat{\phi_1}$	$\hat{\phi_2}$	$\hat{\theta_1}$	$\hat{\theta_2}$	$\hat{\nu_1}$	**AIC**
SARIMA(1,0,0)$(1,1,0)_{52}$	-0,393	–	–	–	-0,548	126,05
SARIMA(1,0,1)$(1,1,0)_{52}$	–	–	-0,377	–	-0,547	126,80
SARIMA(1,0,0)$(0,1,0)_{52}$	-0,354	–	–	–	–	135,66
SARIMA(2,0,1)$(1,1,0)_{52}$	-1,045	$-0,234^*$	$0,663^*$	–	-0,550	129,95
SARIMA(1,0,2)$(1,1,0)_{52}$	-0,734	–	0,328	-0,182	-0,557	129,31

* o coeficiente não é estatisticamente significativo, para um nível de significância $\alpha = 10\%$.

Desta forma, os resultados da estimação do modelo selecionado podem ser consultados, em mais detalhe, na Tabela 8.6.

Tabela 8.6: Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos das quantidades de resíduos seletivos no CHG.

Modelo final: SARIMA$(0,0,1)(1,1,0)_{52}$	AIC $= 126,05$	$\hat{\sigma}^2 = 0,204$
	θ_1	ν_1
estimativa	$-0,377$	-0,547
erro padrão	0,095	0,120

A análise dos resíduos é necessária, uma vez que é preciso validar o modelo escolhido.

Estes devem apresentar, idealmente, um comportamento próximo de uma distribuição Normal, de média nula e variância constante, e não apresentar correlação temporal. O histograma da Figura 8.7 sugere, aparentemente, que os resíduos têm distribuição Normal e, de forma complementar, o teste de Kolmogorov-Smirnov não rejeita a hipótese de normalidade (valor de prova de 0,461) dos resíduos. Também, de acordo com a representação gráfica da série dos resíduos (Figura 8.7) esta apresenta uma distribuição uniforme em torno do resíduo zero, o que leva a concluir que os resíduos têm média nula e variância constante (homocedásticos). De facto, o teste t para o valor médio confirma a não rejeição da hipótese de média nula, resultando num valor de prova igual a 0,323. Desta forma, conclui-se que não existe evidência estatística para rejeitar a hipótese de que os erros seguem uma distribuição Normal de média nula e variância constante.

Quanto à independência, o teste de Ljung-Box é aplicado à série dos resíduos onde k varia entre 3 e 35 (k corresponde ao número de autocorrelações a serem testadas como grupo). Segundo os resultados do teste, a hipótese de independência não é rejeitada para nenhum dos valores de k, apresentando valores de prova entre 0,181 ($k = 5$) e 0,904 ($k = 36$). Realça-se que as FAC e FAC estimadas dos resíduos (Figura 8.7) assemelham-se às FAC e FACP de um ruído branco e, portanto, pode admitir-se a independência dos erros.

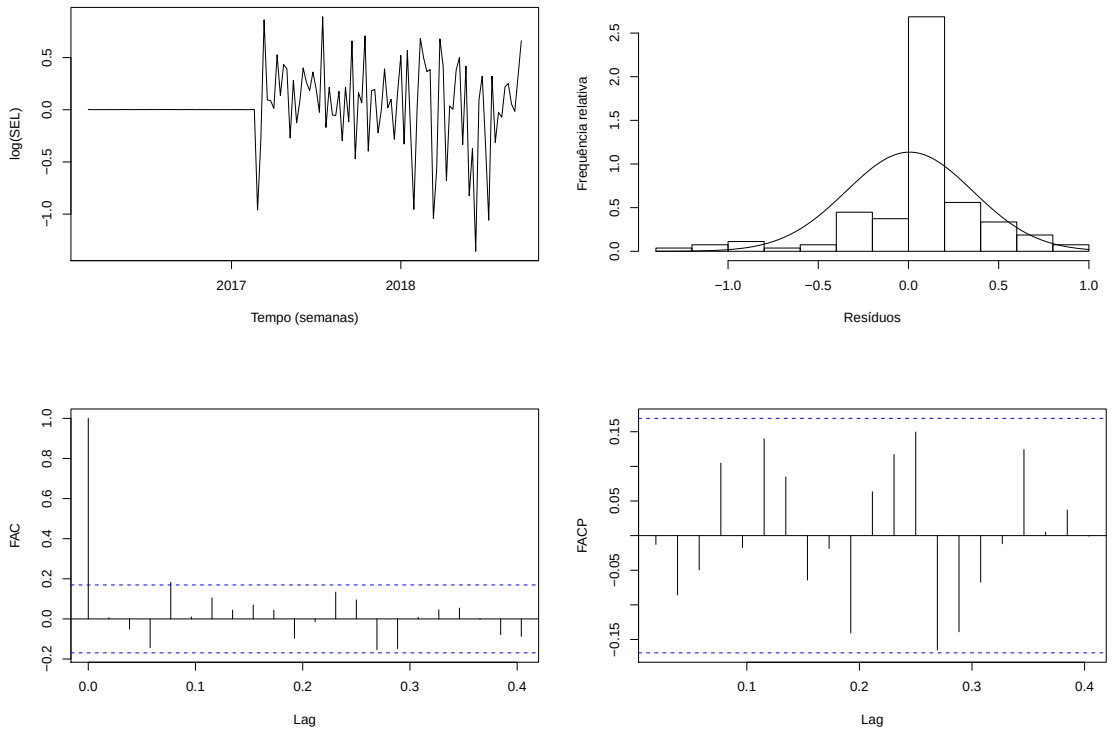


Figura 8.7: Série dos resíduos para a série dos logaritmos das quantidades de resíduos seletivos no CHG, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.

Na Figura 8.8 encontram-se representadas as previsões (no período de teste, isto é, da 39.ª semana de 2018 à 1.ª semana de 2019), pontuais e intervalares, e as estimativas pontuais (do período da 9.ª semana de 2016 à 38.ª semana de 2018) obtidas através do modelo final, nas unidades originais, sobrepostas à série em estudo.

A Figura 8.8 sugere que a qualidade preditiva do modelo é melhor na série de treino do que na série de teste, uma vez que se verifica uma descida atípica nas quantidades de resíduos indiferenciados na série original que o modelo não seria capaz de explicar o fenómeno dadas as observações do passado.

Em relação aos intervalos e previsão, afirma-se que a sua taxa de cobertura é, neste caso, de 40%, uma vez que 6 observações da série de teste pertencem ao interior dos mesmos.

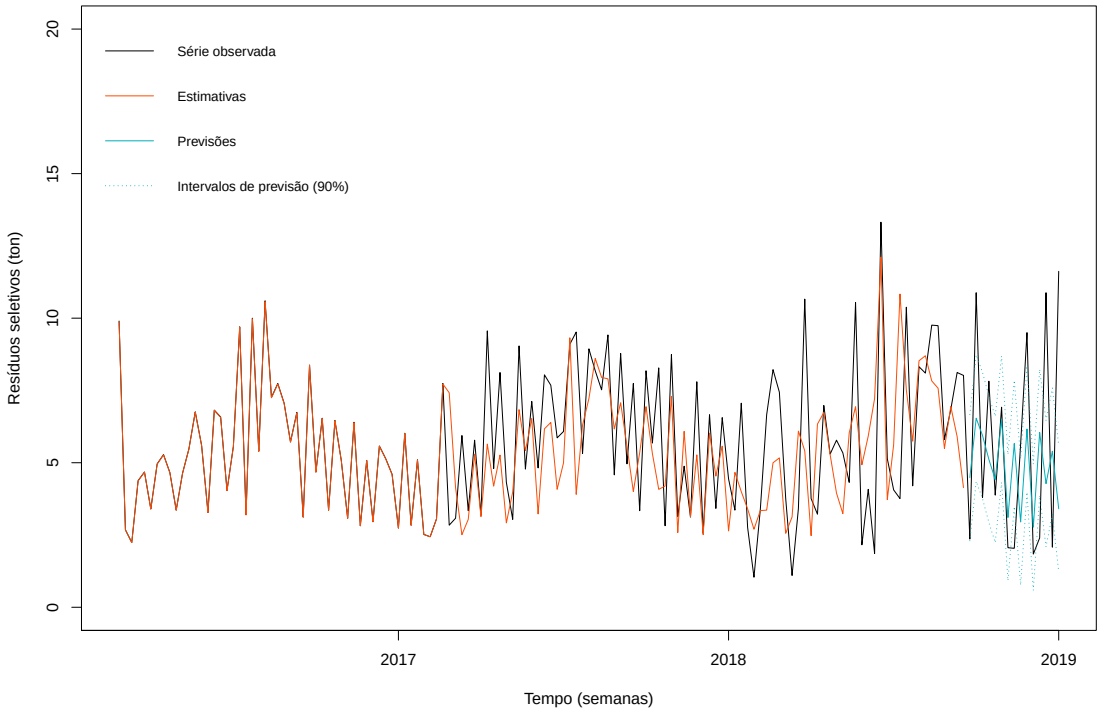


Figura 8.8: Previsões (no período de teste), pontuais e intervalares (90%), e estimativas pontuais (entre a 9.ª semana de 2016 e a 38.ª semana de 2018) obtidas através do modelo SARIMA, sobrepostas à série das quantidades de resíduos seletivos no CHG.

Resíduos indiferenciados

Por último, modelando a série temporal respeitante às quantidades de resíduos indiferenciados, recolhidos de forma semanal no CHG (zona piloto de implementação do sistema PAYT) inicia-se o processo com a identificação de um modelo SARIMA onde, o primeiro passo, consiste na estacionarização da série em estudo. Numa primeira instância, é efetuada a estabilização da variância, através da transformação logarítimica e, conse-

quentemente, é necessário definir a ordem de diferenciação regular para a estabilização da média.

É aplicada uma diferenciação de 1.ª ordem ($d = 1$) onde, consequentemente, a série passa a ser estacionária em média (Figura 8.9). De facto, esta conclusão é suportada, não só pela análise gráfica, mas também pelos dois testes de estacionariedade utilizados - o teste ADF e o teste KPSS. Neste caso, para 3 *lags*, as estatísticas de teste são -9,551 e 0,525 para os testes ADF e KPSS, respetivamente, o que leva a concluir, a um nível de significância de 10%, que, após uma diferenciação de 1.ª ordem, a série é estacionária em média. Assim que a série original é transformada numa série estacionária, o passo seguinte é ajustar a parte sazonal do modelo. Numa primeira instância, pela análise da FAC da Figura 8.9, considera-se um período $s = 52$, valor este derivado de $s \approx 52,18$ devido aos anos bissextos e/ou com 53 semanas. Desta forma, este ajuste da sazonalidade irá permitir uma melhor formulação do modelo SARIMA. Neste estudo os anos estudados são considerados anos comuns, isto é, anos com 365 dias.

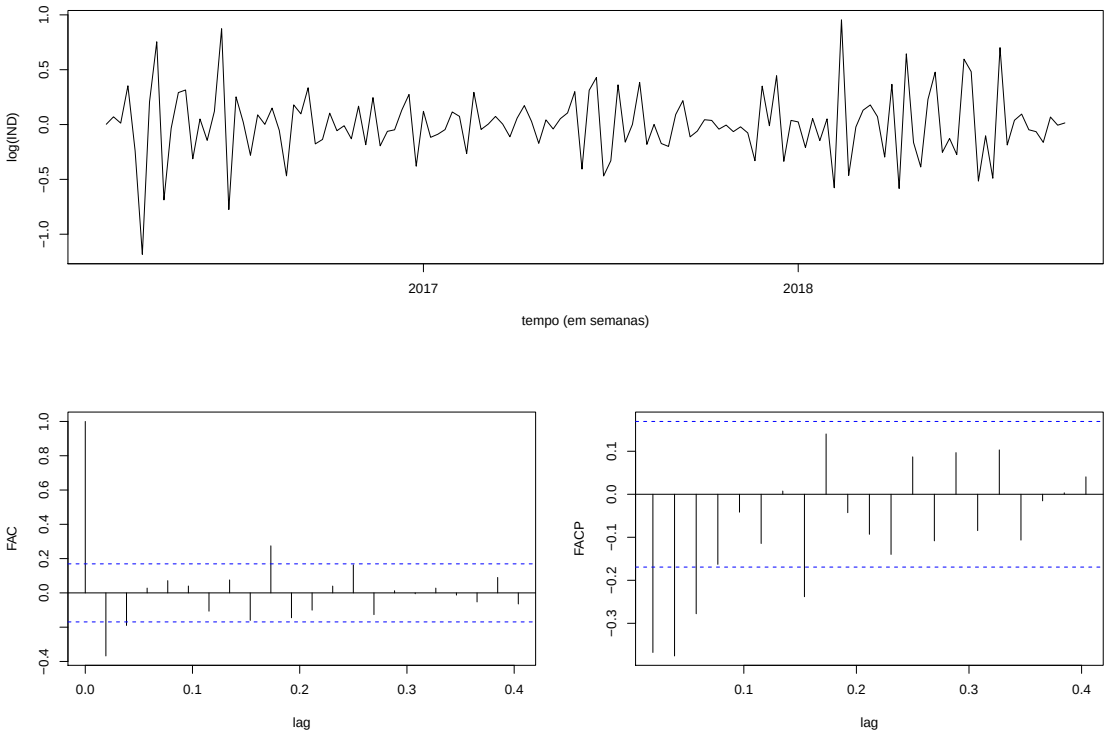


Figura 8.9: Série dos logaritmos das quantidades recolhidas de resíduos indiferenciados, no CHG, após diferenciação de 1.ª ordem ($d = 1$), e respetivas FAC e FACP estimadas.

A identificação das ordens da parte sazonal (P, D, Q) é analisada pela combinação das várias possibilidades, ou seja, trata-se de um processo iterativo, onde se varia P,D e Q entre 0 e 1 (Tabela 8.7). Posto isto, e de acordo com os resultados obtidos, o modelo que resulta no menor AIC é o que considera $P = 1$, $D = 0$ e $Q = 0$ sendo, portanto, estas as

ordens selecionadas para a parte sazonal do modelo. Refira-se que estas não são imutáveis e que, de acordo com as necessidades futuras, podem ser alteradas.

Tabela 8.7: Ajustamento de vários modelos para a parte sazonal, após escolha da ordem de diferenciação regular, à série dos logaritmos das quantidades recolhidas de resíduos indiferenciados no CHG.

Modelo	$\hat{\nu_1}$	$\hat{\eta_1}$	**AIC**
SARIMA$(0,1,0)(1,0,0)_{52}$	–	0,176	54,45
SARIMA$(0,1,0)(0,1,1)_{52}$	–	-0,297	74,87
SARIMA$(0,1,0)(1,0,1)_{52}$	0,949	-0,848	64,66
SARIMA$(0,1,0)(0,1,0)_{52}$	–	–	74,98
SARIMA$(0,1,0)(0,0,1)_{52}$	–	0,136	64,99
SARIMA$(0,1,0)(1,1,0)_{52}$	-0,273	–	74,87
SARIMA$(0,1,0)(1,1,1)_{52}$	-0,261	-0,013	76,87

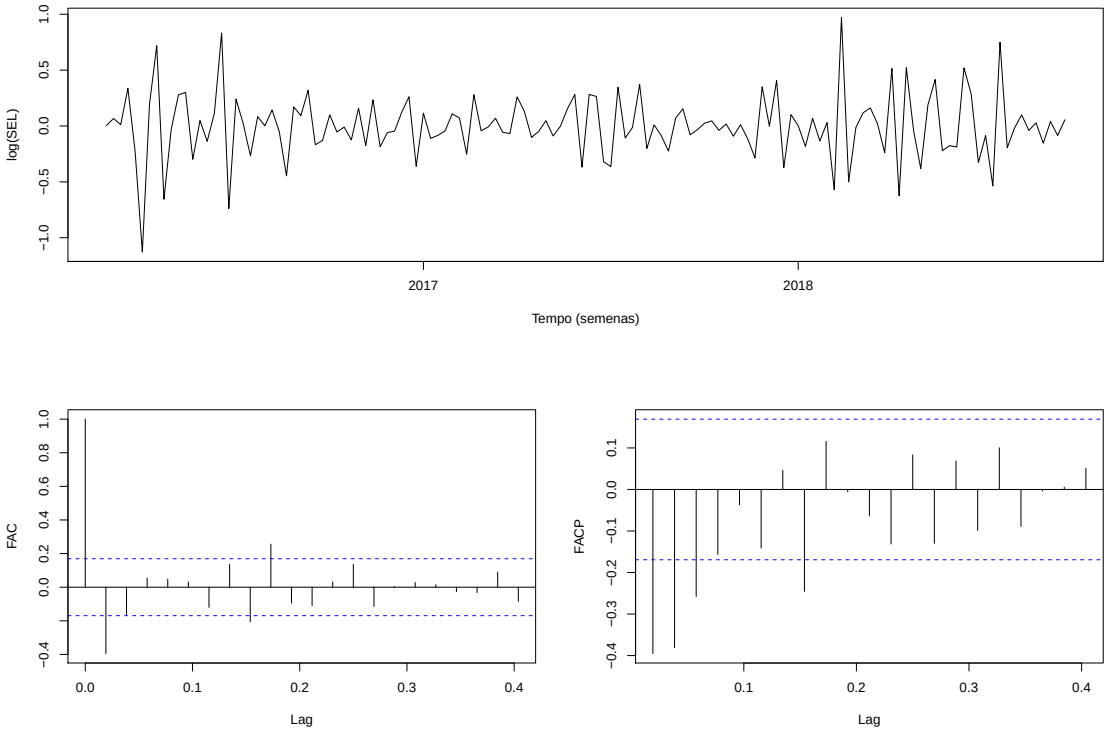


Figura 8.10: Série dos resíduos das quantidades recolhidas de resíduos seletivos, no CHG, após aplicação de uma diferenciação de 1.ª ordem ($d = 1$), e ajustamento da parte sazonal, e respetivas FAC e FACP estimadas.

Na Figura 8.10 pode ver-se o comportamento dos resíduos após o ajustamento da parte sazonal. O último passo para a identificação de um modelo SARIMA é a escolha das ordens p e q, utilizando como ferramenta as FAC e FACP empíricas. De facto, através da comparação das FAC e FACP da Figura 8.10 com as FAC e FACP teóricas (ver Tabela 5.1),

é possível reconhecer-se o comportamento de um modelo MA(1), ou, alternativamente, de um AR(3). No entanto, além destes modelos, são também ajustados modelos "vizinhos", de forma a realizar uma escolha mais pensada (ver Tabela 8.8).

Analisando a Tabela 8.8, dos nove modelos formulados, verifica-se que três deles não são adequados devido à não significância de alguns dos seus coeficientes. Com isto, A escolha entre os outros modelos foi realizada, com base no AIC e, de facto, de entre os modelos, opta-se, então, pelo que detém menor valor de AIC, ou seja, o modelo SARIMA(1, 1, 2)$(1,0,0)_{52}$.

Tabela 8.8: Ajustamento de vários modelos para a parte regular, após escolha da ordem de diferenciação regular e das ordens da parte sazonal, à série dos logaritmos das quantidades de resíduos indiferenciados no CHG.

Modelo	$\hat{\phi_1}$	$\hat{\phi_2}$	$\hat{\theta_1}$	$\hat{\theta_2}$	$\hat{\nu_1}$	$\hat{\eta_1}$	**AIC**
SARIMA(0,1,1)$(1,0,1)_{52}$	–	–	-0,800	–	0,908	$-0,741^*$	10,45
SARIMA(2,1,0)$(1,0,1)_{52}$	-0,563	-0,381	–	–	0,994	-0,931	24,41
SARIMA(0,1,0)$(1,0,1)_{52}$	–	–	–	–	0,949	-0,848	64,66
SARIMA(0,1,2)$(1,0,1)_{52}$	–	–	-0,721	$-0,116^*$	$0,875^*$	$-0,701^*$	11,25
SARIMA(1,1,2)$(1,0,1)_{52}$	0,911	–	-1,719	0,719	0,896	-0,721	11,61
SARIMA(0,1,1)$(0,0,1)_{52}$	–	–	-0,810	–	–	0,221	12,24
SARIMA(0,1,1)$(1,0,0)_{52}$	–	–	-0,812	–	0,291	–	10,54
SARIMA(1,1,1)$(0,0,1)_{52}$	$0,147^*$	–	-0,876	–	–	$0,217^*$	12,62
SARIMA(1,1,2)$(1,0,0)_{52}$	-0,815	–	0,145	-0,822	0,280	–	4,61

* o coeficiente não é estatisticamente significativo, para um nível de significância $\alpha = 10\%$.

Os resultados da estimação deste modelo podem ser consultados, em mais detalhe, na Tabela 8.9.

Tabela 8.9: Resultados da estimação do modelo SARIMA aplicado à série dos logaritmos das quantidades de resíduos indiferenciados no CHG.

Modelo final: SARIMA$(1,1,2)(1,0,0)_{52}$		AIC $= 4,61$	$\hat{\sigma}^2 = 0,054$
	$\hat{\phi_1}$	$\hat{\theta_1}$	$\hat{\theta_2}$	$\hat{\nu_1}$
estimativa	$-0,815$	0,145	-0,822	0,280
erro padrão	0,066	0,070	0,064	0,114

De forma a validar o modelo escolhido deve realizar-se uma análise dos resíduos (ver Figura 8.11). Idealmente, estes devem apresentar um comportamento próximo de uma distribuição Normal, de média nula e variância constante, e não ser correlacionados no tempo. De acordo com a representação gráfica da série dos resíduos (Figura 8.11), considera-se que esta apresenta uma distribuição (relativamente) uniforme em torno do resíduo zero, o que sugere que os erros têm média nula e variância constante. A média nula, numa perspetiva analítica, pode ser verificada pelo teste t para a média de onde se obtém um

valor de prova igual a 0,972 comprovando, desta forma, o pressuposto em análise.

Além disso, o histograma da Figura 8.11 sugere que, apesar da existência de algumas observações discrepantes, os resíduos têm uma distribuição Normal, o que é comprovado pela aplicação do teste de Kolmogorov-Smirnov, com um valor de prova de 0,111. Admite-se, assim, que os erros seguem uma distribuição Normal de média nula e variância constante.

Quanto à independência, o teste de Ljung-Box é aplicado à série dos resíduos onde k varia entre 5 e 35 (k corresponde ao número de autocorrelações a serem testadas como grupo). Segundo os resultados do teste, a hipótese de independência não é rejeitada para nenhum dos valores de k, apresentando valores de prova entre 0,215 ($k = 6$) e 0,843 ($k = 40$). Note-se que as FAC e FAC estimadas dos resíduos (Figura 8.11) assemelham-se às FAC e FACP de um ruído branco e, portanto, pode admitir-se a independência dos erros.

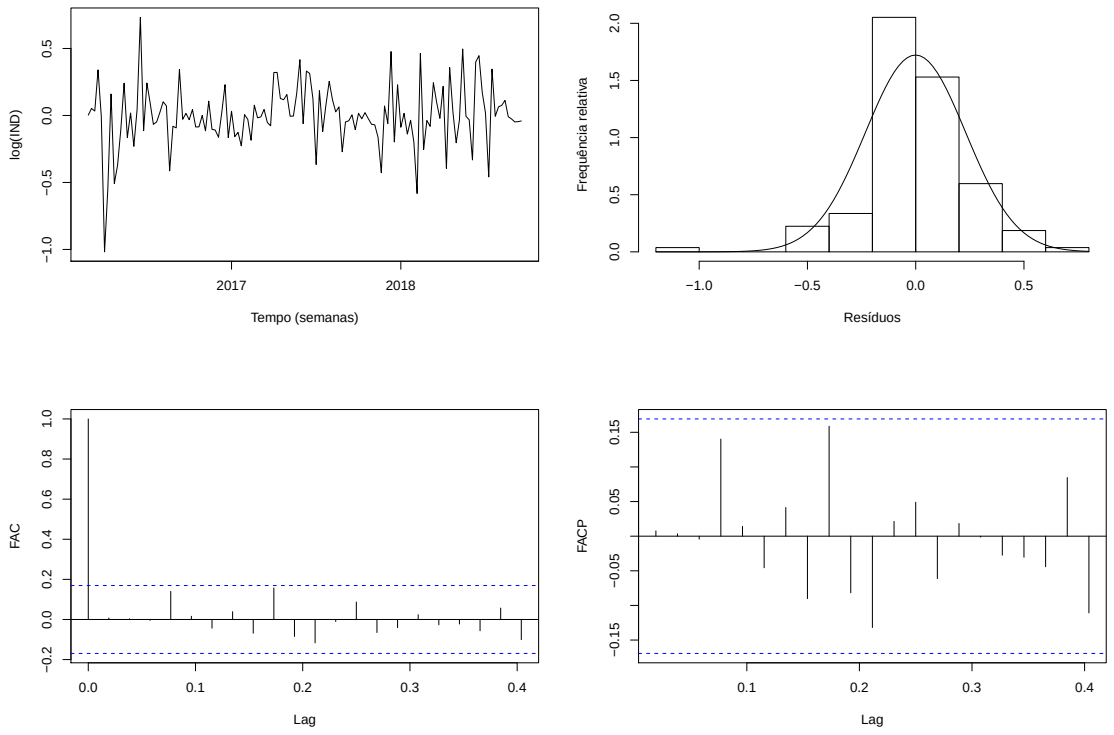


Figura 8.11: Série dos resíduos para a série dos logaritmos das quantidades de resíduos indiferenciados no CHG, após ajustamento do modelo SARIMA, e respetivo histograma, FAC e FACP estimadas.

Na Figura 8.12 encontram-se representadas as previsões (no período de teste: 39.ª semana de 2018 à 1ª semana de 2019), pontuais e intervalares, e as estimativas pontuais (no período compreendido entre a 9.ª semana de 2016 e a 38ª semana de 2018) obtidas através do modelo final, nas unidades originais, sobrepostas à série em estudo.

Em relação aos intervalos e previsão, afirma-se que a sua taxa de cobertura é, neste caso, de 60%, uma vez que apenas 9 observações da série de teste pertencem ao interior

dos mesmos.

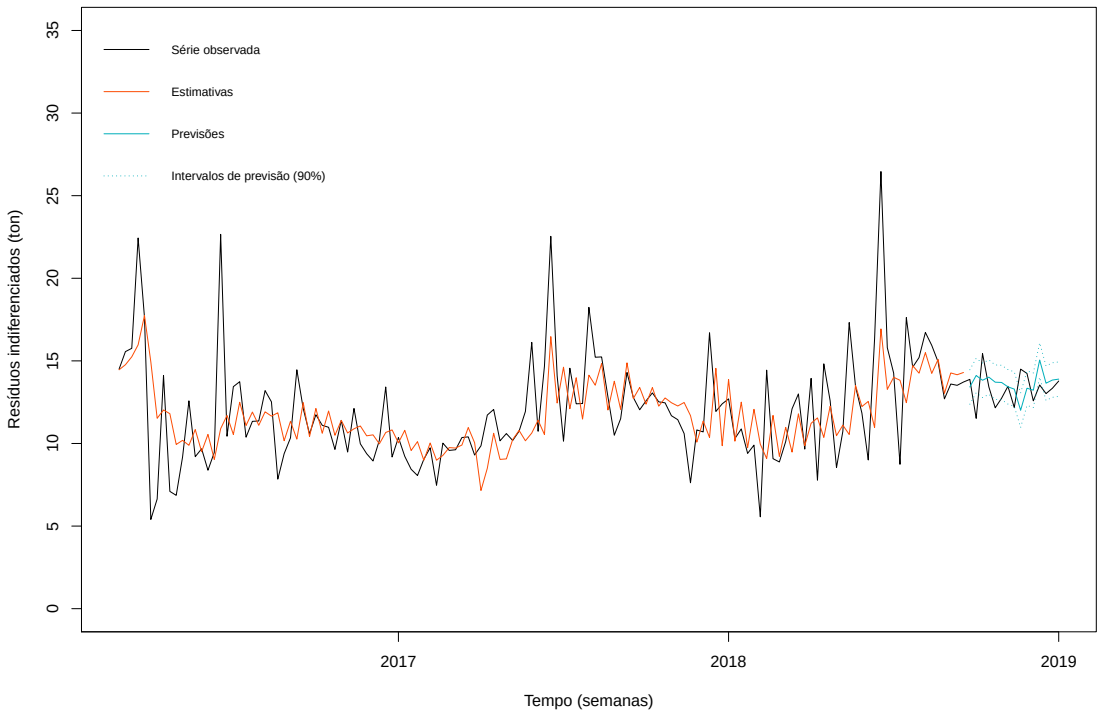


Figura 8.12: Previsões (no período de teste), pontuais e intervalares (90%), e estimativas pontuais (entre a 9.ª semana de 2016 e a 38.ª semana de 2018) obtidas através do modelo SARIMA, sobrepostas à série das quantidades de resíduos indiferenciados no CHG.

8.3 Avaliação dos Modelos de Previsão

Formulados os modelos de previsão aos dados fornecidos, é necessária a identificação dos modelos que melhor se adequaram à situação em estudo. Para efeitos comparativos são utilizadas quatro medidas de avaliação: o EQM e a sua correspondente na mesma escala dos dados, REQM, o EPAM, o e o EEAM. Para além destas medidas terem sido calculadas para a série de teste, para as respetivas 15 observações em cada série temporal, são também determinadas para a série de treino, a partir dos resíduos do modelo em questão. Os resultados podem ser consultados na Tabela 8.10.

Da análise da Tabela 8.10, verifica-se que o melhor modelo que melhor explica o comportamento dos dados (série de treino) foram os estimados para os resíduos seletivos na zona de implementação do PAYT pelo que, no entanto, em respeito à previsão (série de teste) é a que apresenta o pior resultado. Já em relação às séries dos resíduos indiferenciados em contentores de profundidade e na zona de implementação do PAYT, pode-se afirmar que estes detêm o melhor modelo ajustado à série de teste do que na série de treino.

Tabela 8.10: Medidas de avaliação calculadas para as séries estudadas, no período de treino e no período de teste respetivo, com base nos resultados obtidos na aplicação do método de previsão.

	Série de treino				Série de teste			
Série	EQM	REQM	EPAM	EEAM	EQM	REQM	EPAM	EEAM
CONT	271,187	16,468	3,511	0,639	206,806	14,381	3,571	0,467
SEL	**3,202**	**1,789**	**23,473**	**0,414**	**13,633**	**3,692**	**69,687**	**0,363**
IND	**7,804**	**2,793**	**16,488**	**0,579**	1,654	1,286	7,965	0,723

Uma vez avaliada a precisão das previsões realizadas (pontuais), é fundamental compreender a eficácia das previsões intervalares. Teoricamente, os intervalos de previsão são calculadas a uma confiança de 90%, o que significa que 90% dos intervalos deve incluir a observação observada (real). Com isto, considera-se que as previsões intervalares mais eficazes são aquelas cuja taxa de cobertura efetiva mais se aproxima de 90%. Note-se que os intervalos de previsão são obtidos com base na série de teste, para cada série distinta onde, neste estudo, contêm apenas 15 observações e, desta forma, a análise das taxas de cobertura deve ser cuidada.

Nas três séries em estudo – CONT, SEL e IND – são calculadas as taxas de cobertura de valor igual a 40%, 40% e 60%, respetivamente. Com isto é notório que o modelo formulado para a série correspondente à produção de resíduos indiferenciados na zona de implementação do sistema PAYT apresenta melhores resultados.

Tabela 8.11: Tabela com os respetivos intervalos de previsão, valores previstos e valores reais para cada série em estudo, respetivamente.

CONT			SEL			IND		
Intervalo de predição	Valor Previsto	Valor Real	Intervalo de predição	Valor Previsto	Valor Real	Intervalo de predição	Valor Previsto	Valor Real
(325,117 ; 337,321)	331,219	337,200	(2,300 ; 6,667)	4,484	2,360	(12,376 ; 14,463)	13,419	13,880
(327,413 ; 339,617)	333,515	314,020	(4,357 ; 8,725)	6,541	10,880	(13,072 ; 15,159)	14,115	11,520
(329,027 ; 341,231)	335,129	318,000	(3,757 ; 8,124)	5,940	3,800	(12,782 ; 14,869)	13,826	15,440
(326,765 ; 338,969)	332,867	318,660	(2,931 ; 7,299)	5,115	7,820	(12,972 ; 15,059)	14,015	13,420
(334,013 ; 346,217)	340,115	319,320	(2,236 ; 6,604)	4,420	3,880	(12,666 ; 14,753)	13,710	12,160
(330,435 ; 342,639)	336,537	312,400	(4,320 ; 8,688)	6,504	6,920	(12,649 ; 14,737)	13,693	12,740
(333,233 ; 345,639)	339,335	324,760	(0,921 ; 5,288)	3,105	2,060	(12,384 ; 14,471)	13,428	13,460
(332,958 ; 345,162)	339,060	326,380	(3,480 ; 7,848)	5,664	2,040	(12,249 ; 14,336)	13,293	12,180
(332,840 ; 345,044)	338,942	331,180	(0,768 ; 5,135)	2,951	5,640	(10,967 ; 13,054)	12,011	14,500
(333,118 ; 345,322)	339,220	337,220	(3,982 ; 8,350)	6,166	9,500	(12,303 ; 14,390)	13,347	14,240
(330,492 ; 342,696)	336,594	342,240	(0,588 ; 4,956)	2,772	1,840	(12,181 ; 14,268)	13,225	12,580
(335,077 ; 347,281)	341,179	353,560	(3,861 ; 8,228)	6,044	2,400	(14,008 ; 16,095)	15,052	13,540
(337,559 ; 349,763)	343,661	348,860	(2,092 ; 6,460)	4,276	10,880	(12,605 ; 14,693)	13,649	13,020
(335,336 ; 347,540)	341,438	363,520	(3,216 ; 7,584)	5,400	2,080	(12,796 ; 14,883)	13,839	13,340
(326,509 ; 338,713)	332,611	337,700	(1,224 ; 5,591)	3,408	11,620	(12,851 ; 14,939)	13,895	13,780

Capítulo 9

Conclusão

O presente estudo permitiu a criação de uma base de conhecimentos acerca da Gestão de Resíduos em contexto local, nomeadamente sobre a implementação de um sistema inovador denominado de *Pay-As-You-Throw* em que o cidadão paga unicamente aquilo que produz.

Este trabalho focou-se na análise e modelação de dados via modelos de Regressão Linear, nomeadamente no estudo dos fatores que influenciam as quantidades de resíduos indiferenciados e seletivos, respetivamente, produzidas na zona de implementação do sistema PAYT. Desta forma verificou-se a presença de uma tendência crescente na produção de resíduos indiferenciados e uma decrescente na produção de resíduos seletivos.

Realça-se que a produção de resíduos indiferenciados está relacionada de forma negativa com a venda de sacos de 50 litros, a utilizadores domésticos, com o número de fretes, isto é, o número de vezes que a viatura de desloca da zona de atuação à estação de triagem. Caso a VITRUS opte por não oferecer sacos para a reciclagem, a produção destes resíduos poderá aumentar significativamente, uma vez que há falta informação e de medidas pedagógicas para a importância da reciclagem.

A produção de resíduos seletivos está associada linearmente, de forma negativa, com o número de sacos de 15 e 30 litros vendidos a utilizadores domésticos. Em relação ao número de utilizadores que compraram sacos, realça-se a associação linear negativa com o número de utilizadores não domésticos, da tipologia A, que compram sacos, ou seja, estes utilizadores são grandes produtores de resíduos indiferenciados. Em relação aos das tipologias B, D e E verifica-se uma relação positiva, ou seja, são estes que influenciam de forma significativa a produção de resíduos seletivos. Já a quilometragem efetuada e o número de fretes realizados contribuem para o aumento da recolha de resíduos seletivos. O número de sacos de 30 litros para a reciclagem, oferecidos aos utilizadores, contribuem de forma positiva, contrariamente aos de 100 litros que detêm uma relação negativa com a produção de resíduos seletivos. Esta situação observada nos sacos de 100 litros poderá estar relacionada com possíveis quebras de *stock* ou outros fatores que influenciam a produção de uma forma direta.

Numa fase seguinte, foi modelada a sazonalidade dos dados, no período observado, onde numa primeira etapa foram formulados os modelos com, exclusivamente, variáveis indicatrizes correspondentes à componente sazonal. Desta forma, para os resíduos indiferenciados foram obtidos padrões sazonais nos meses de fevereiro, abril, junho e agosto. Estes meses podem estar relacionados com os festejos do Carnaval, Páscoa, Festas Gualterianas e outras atividades realizadas no CHG.

Nos resíduos seletivos, foram obtidos padrões sazonais nos meses de fevereiro, abril, julho, agosto, setembro e dezembro. Estes padrões também podem ser explicados devido às atividades realizadas no CHG que, como consequência, levam a uma produção de resíduos significativa. Porém, realça-se que, nestes meses, os utilizadores tendem a reciclar com frequência levando a empresa a ter especial atenção nestes meses. Ao aprofundar a modelação da sazonalidade, foram formulados modelos com a combinação das variáveis significativas obtidas por Regressão Linear Simples e, também, por Regressão Linear Múltipla, após aplicação do método de seleção autorregressivo. Desta forma, verifica-se que os modelos que detêm maior AIC são os que resultam da combinação das variáveis obtidas por Regressão Linear Múltipla com as variáveis indicatrizes da componente sazonal.

Numa última fase foram aplicados modelos de previsão em Séries Temporais a três séries distintas – `CONT`, `SEL` e `IND` – nas quais foram obtidos os respetivos modelos de previsão de forma a averiguar o comportamento da produção de resíduos nas diversas zonas afetas ao SHU. Então, para os resíduos indiferenciados em contentores de profundidade foi obtido o modelo SARIMA$(0,1,2)(0,0,1)_{52}$, para os resíduos seletivos no CHG considerou-se o modelo SARIMA$(0,0,1)(1,1,0)_{52}$ e os resíduos indiferenciados foram modelados com um SARIMA$(1,1,2)(1,0,0)_{52}$. Obtidos os modelos foram calculadas as estimativas pontuais e as respetivas estimativas intervalares de onde se conclui que a melhor taxa de cobertura obtida foi para os resíduos indiferenciados no CHG com um valor igual a 60%. Também foram calculadas as medidas de avaliação para verificação da precisão das previsões efetuadas de onde se conclui que os modelos obtiveram um melhor ajustamento na série de teste para as séries `CONT` e `SEL`. Contrariamente, a série `IND` obteve um melhor ajustamento na série de treino.

9.1 Sugestões para trabalho futuro

No decorrer do presente estudo poderiam ter sido incluídas variáveis correspondentes à caracterização da zona de implementação do PAYT, nomeadamente o número, por mês, de habitantes, de lojistas e de habitações com ou sem moradores, as habilitações escolares dos utilizadores e, também variáveis económicas que de uma forma geral influenciam a produção de resíduos. Desta forma não foi possível considerar tais variáveis nos modelos de regressão com a finalidade de inferir acerca das variáveis que mais influência têm na produção de resíduos no CHG. A metodologia aplicada apresenta algumas limitações uma vez que o sistema foi implementado em abril de 2016 e não é possível obter dados de

certas variáveis, por exemplo, com frequência semanal de forma a obter uma significativa quantidade de observações sobre as quais se poderia modelar e inferir.

Em relação à aplicação e modelos para previsão em Séries Temporais, surgiram algumas ideias de possíveis modelos que se poderiam aplicar aos dados em estudo, nomeadamente a aplicação de modelos de Alisamento Exponencial, modelos de Holt-Winters e, também, a aplicação de modelos de Regressão Múltipla para modelação de tendências e padrões sazonais a partir de séries temporais interrompidas. Com a aplicação desta diversidade de modelos poderia-se, entretanto, efetuar uma comparação de modelos com vista à seleção daquele que detenha uma melhor precisão nas previsões efetuadas.

Bibliografia

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Alpuim, M. (2013). *Notas de apoio à disciplina de Modelos Lineares, Mestrado MAEG*. FCUL.

Alpuim, T. (1998). *Séries Temporais*. Associação dos Estudantes da Faculdade de Ciências de Lisboa, 2ª ed.

Batllevell, Marta & Hanf, K. (2008). The fairness of payt systems: Some guidelines for decision-makers. *Waste management*, 28(12):2793–2800.

Bilitewski, B. (2008). From traditional to modern fee systems. *Waste management*, 28(12):2760–2766.

Box, George E.P. & Jenkins, G. (1970). Statistical models for forecasting and control.

Box, G. & Jenkins, G. . R. G. (2013). *Time series analysis: forecasting and control*. John Wiley and Sons.

Box, G. E. and Jenkins, Gwilym M. & Reinsel, G. C. . L. G. M. (2016). *Time series analysis: forecasting and control*. John Wiley & Sons, 5ª ed.

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

Canterbury, Janice & Newill, R. (2003). The pay-as-you-throw payoff. *American City & County*, 118(11):36–36.

Chatfield, C. (2000). *Time-series forecasting*. Chapman and Hall/CRC.

Chatfield, C. (2003). *The analysis of time series: an introduction*. Chapman and Hall/-CRC, 6ª ed.

Chatfield, C. (2004). *The analysis of time series: an introduction*. Chapman and Hall/-CRC, 5ª ed.

Chatterjee, Samprit & Hadi, A. . P. B. (2000). Regression analysis by example john wiley & sons. *Inc., New York*.

Chatterjee, Samprit & Hadi, A. S. (2009). *Sensitivity analysis in linear regression*, volume 327. John Wiley & Sons.

Cleveland, William S. & Terpenning, I. J. (1982). Graphical methods for seasonal adjustment. *Journal of the American Statistical Association*, 77(377).

Cordeiro, C. M. H. (2011). Métodos de reamostragem em modelos de previsão.

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

Dohogne, JJ & Labriga, L. . L. G. (2016). Cross-analysis of 'pay-as-you-throw'schemes in selected eu municipalities.

Fahrmeir, L. and Kneib, Thomas & Lang, S. . M. B. (2013). *Regression: models, methods and applications*. Springer Science & Business Media.

Freitas, D. d. G. S. M. (2013). *Implementação do Sistema Pay As You Throw-PAYT no Centro Histórico de Guimarães e Zona Envolvente*. PhD thesis, Universidade Fernando Pessoa.

Galton, F. (1889). *Natural Inheritance*, volume 42. Macmillan.

Gonçalves, A. Manuela & Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics*, 22(8):933–945.

Hyndman, Rob J. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J. (2019). Cran task view: Time series analysis.

Jebb, A. T. and Tay, Louis & Wang, W. . H. Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology*, 6:727.

Karagiannidis, Avraam & Xirogiannopoulou, A. . T. G. (2008). Full cost accounting as a tool for the financial assessment of pay-as-you-throw schemes: A case study for the panorama municipality, greece. *Waste Management*, 28(12):2801–2808.

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

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

Martinho, M. & Gonçalves, M. (1999). *Gestão de Resíduos*. Universidade Aberta.

Menezes, R. (2019). Sebenta de séries temporais.

Metcalfe, Andrew V. & Cowpertwait, P. S. (2009). *Introductory time series with R*. Springer.

Murteira, Bento & Müller, D. . T. K. F. (2000). *Análise de sucessões cronológicas*.

Navarro-Esbrí, Joaquín & Diamadopoulos, E. . G. D. (2002). Time series analysis and forecasting techniques for municipal solid waste management. *Resources, Conservation and Recycling*, 35:201–214.

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

Persons, W. M. (1919). *Indices of Business Conditions: An Index of General Business Conditions*, volume 1. Harvard University Press.

Reichenbach, J. (2008). Status and prospects of pay-as-you-throw in europe–a review of pilot research and implementation studies. *Waste Management*, 28(12):2809–2814.

Rimaitytė, I., Ruzgas, T., and Denafas, Gintaras & Račys, V. . M. D. (2012). Application and evaluation of forecasting methods for municipal solid waste generation in an eastern-european city. *Waste Management and Research*, 30(1):89–98.

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

Schindler, H. R., Schmalbein, N., Steltenkamp, V., and Cave, Jonathan & Wens, B. . A. A. (2012). Smart trash: Study on rfid tags and the recycling industry.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

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

Sen, Ashish & Srivastava, M. (2012). *Regression analysis: theory, methods, and applications*. Springer Science and Business Media.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Shumway, Robert H. & Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer.

Skumatz, Lisa A & Green, K. (2002). *Variable-rate Or"pay-as-you-throw"Waste Management: Answers to Frequently Asked Questions*. Reason Foundation.

Skumatz, L. A. (2008). Pay as you throw in the us: Implementation, impacts, and experience. *Waste management*, 12(28):2778–2785.

Song, Jingwei & He, J. (2014). A multistep chaotic model for municipal solid waste generation prediction. *Environmental engineering science*, 31(8):461–468.

Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Tsay, R. (2010). *Analysis of Financial Time Series*. John Wiley and Sons, 3ª ed.

VITRUS AMBIENTE, EM, S. (2018). Relatório do serviço de higiene urbana.

VITRUS AMBIENTE, EM, S. (2019a). Relatório da implementação do sistema payt em guimarães.

VITRUS AMBIENTE, EM, S. (2019b). Vitrus ambiente.

Wheelwright, Steven & Makridakis, S. . H. R. J. (1998). *Forecasting: methods and applications*. John Wiley & Sons.

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

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

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

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

Apêndice A

Circuitos de Recolha de Resíduos Urbanos

Tabela A.1: Distribuição dos circuitos de recolha indiferenciada pelas freguesias.

Circuito 1
Azurém; Fermentões; UF Oliveira, S. Paio e S. Sebastião; Creixomil; Nespereira; UF Selho S. Lourenço e Gominhães; Pencelo; Polvoreira; S. Torcato.
Circuito 2
Aldão; Azurém; Costa; Fermentões; Mesão Frio; Moreira de Cónegos; UF Conde e Gandarela; UF Oliveira, S. Paio e S. Sebastião.
Circuito 3
Barco; Brito; Caldelas; Fermentões; Ponte; Sande S. Martinho; Prazins Santa Eufémia; UF Sande Vila Nova e S. Clemente; UF Souto Santa Maria, Souto S. Salvador e Gondomar; Urgezes.
Circuito 4
Creixomil; Gondar; Guardizela; Pinheiro; Polvoreira; Selho S. Cristóvão; Selho S. Jorge; Serzedelo; UF Candoso Santiago e Mascotelos; Urgezes.
Circuito 5
Brito; Creixomil; Fermentões; Silvares; UF Candoso S. Tiago e Mascotelos; Urgezes.
Circuito 6
Ponte; Brito; Ronfe; Gondar.
Circuito 7
Caldelas; UF Briteiros Santo Estevão e Donim.
Circuito 8
Ponte.
Circuito 9
UF Sande S. Lourenço e Balasar; Longos; Briteiros Santa Leocádia.
Circuito 10
UF Souto Santa Maria, Souto S. Salvador e Gondomar; UF Arosa e Castelões.
Circuito 11
Sande São Martinho.
Circuito 12
Caldelas; Barco; Briteiros S. Salvador.

Apêndice B

Regressão Linear Simples

Tabela B.1: Regressão Linear Simples, tendo como variável resposta `papel`.

Regressão Linear Simples				
		papel		
Variável	Estimativas	$\hat{\sigma}$	p-valor	R^2
MTH	$\hat{\beta_0}$=6,029	0,769	<0,001	0,021
	$\hat{\beta_1}$= 0,032	0,039	0,009	
SAC15UD	$\hat{\beta_0}$=6,127	0,909	<0,001	0,010
	$\hat{\beta_1}$=0,001	0,002	0,589	
SAC15UND	$\hat{\beta_0}$=6,224	0,639	<0,001	0,015
	$\hat{\beta_1}$=0,007	0,011	0,498	
SAC30UD	$\hat{\beta_0}$=8,109	1,369	<0,001	0,042
	$\hat{\beta_1}$=-0,003	0,003	0,254	
SAC30UND	$\hat{\beta_0}$=7,161	1,087	<0,001	0,010
	$\hat{\beta_1}$=-0,002	0,004	0,572	
SAC50UD	$\hat{\beta_0}$=6,881	0,619	<0,001	0,012
	$\hat{\beta_1}$=-0,014	0,022	0,542	
SAC50UND	$\hat{\beta_0}$=5,126	1,556	0,002	0,029
	$\hat{\beta_1}$=0,001	0,001	0,344	
SAC100UD	$\hat{\beta_0}$=6,864	0,460	<0,001	**0,035**
	$\hat{\beta_1}$=-0,065	0,062	0,297	
SAC100UND	$\hat{\beta_0}$=5,243	1,393	0,001	0,031
	$\hat{\beta_1}$=0,002	0,002	0,327	
UDC	$\hat{\beta_0}$=7,500	2,453	0,005	0,005
	$\hat{\beta_1}$=-0,013	0,035	0,706	
TAC	$\hat{\beta_0}$=7,266	1,949	0,001	0,004
	$\hat{\beta_1}$=-0,037	0,102	0,722	
TBC	$\hat{\beta_0}$=3,321	2,460	0,187	0,055
	$\hat{\beta_1}$=0,136	0,101	0,190	
TCC	$\hat{\beta_0}$=6,640	0,855	<0,001	0,000
	$\hat{\beta_1}$=-0,004	0,054	0,936	
TDC	$\hat{\beta_0}$=4,725	1,253	0,001	0,072
	$\hat{\beta_1}$=0,244	0,158	0,132	
TEC	$\hat{\beta_0}$=6,184	1,307	<0,001	0,003
	$\hat{\beta_1}$=0,057	0,181	0,755	
KMS	$\hat{\beta_0}$=5,594	0,875	<0,001	0,047
	$\hat{\beta_1}$=0,001	0,000	0,223	
FRT	$\hat{\beta_0}$=7,055	1,853	0,001	0,002
	$\hat{\beta_1}$=-0,025	0,093	0,794	
DEP	$\hat{\beta_0}$=5,581	0,816	<0,001	0,057
	$\hat{\beta_1}$=0,004	0,003	0,181	
OFERECE	$\hat{\beta_0}$=6,717	0,528	<0,001	0,005
	$\hat{\beta_1}$=-0,285	0,758	0,709	
REC30	$\hat{\beta_0}$=6,569	0,465	<0,001	0,000
	$\hat{\beta_1}$=0,000	0,001	0,971	
REC50	$\hat{\beta_0}$=6,301	0,493	<0,001	0,023
	$\hat{\beta_1}$=0,001	0,001	0,400	
REC100	$\hat{\beta_0}$=6,470	0,491	<0,001	0,004
	$\hat{\beta_1}$=0,000	0,001	0,730	

Tabela B.2: Regressão Linear Simples, tendo como variável resposta `plastico`.

Regressão Linear Simples				
	plastico			
Variável	Estimativas	$\hat{\sigma}$	p-valor	R^2
MTH	$\hat{\beta_0}$=4,674	0,335	<0,001	0,008
	$\hat{\beta_1}$=-0,008	0,017	0,626	
SAC15UD	$\hat{\beta_0}$=4,797	0,392	<0,001	0,018
	$\hat{\beta_1}$=-0,001	0,001	0,460	
SAC15UND	$\hat{\beta_0}$=4,421	0,278	<0,001	0,008
	$\hat{\beta_1}$=0,002	0,005	0,628	
SAC30UD	$\hat{\beta_0}$=3,468	0,572	<0,001	0,107
	$\hat{\beta_1}$=-0,002	0,001	0,063	
SAC30UND	$\hat{\beta_0}$=3,941	0,459	<0,001	0,057
	$\hat{\beta_1}$=-0,002	0,002	0,181	
SAC50UD	$\hat{\beta_0}$=4,640	0,269	<0,001	0,008
	$\hat{\beta_1}$=-0,005	0,010	0,610	
SAC50UND	$\hat{\beta_0}$=2,605	0,583	<0,001	0,272
	$\hat{\beta_1}$=0,001	0,000	0,002	
SAC100UD	$\hat{\beta_0}$=4,465	0,202	<0,001	0,010
	$\hat{\beta_1}$=-0,015	0,027	0,585	
SAC100UND	$\hat{\beta_0}$=2,654	0,502	<0,001	0,031
	$\hat{\beta_1}$=0,003	0,001	0,001	
UDC	$\hat{\beta_0}$=3,761	1,055	0,001	0,017
	$\hat{\beta_1}$=0,011	0,015	0,466	
TAC	$\hat{\beta_0}$=3,268	0,813	0,000	0,075
	$\hat{\beta_1}$=0,067	0,043	0,124	
TBC	$\hat{\beta_0}$=3,017	1,060	0,008	0,063
	$\hat{\beta_1}$=0,063	0,044	0,159	
TCC	$\hat{\beta_0}$=4,165	0,363	<0,001	0,039
	$\hat{\beta_1}$=0,026	0,023	0,270	
TDC	$\hat{\beta_0}$=3,781	0,545	<0,001	0,062
	$\hat{\beta_1}$=0,098	0,069	0,161	
TEC	$\hat{\beta_0}$=4,353	0,565	<0,001	0,003
	$\hat{\beta_1}$=0,026	0,078	0,746	
KMS	$\hat{\beta_0}$=4,891	0,381	<0,001	0,034
	$\hat{\beta_1}$=0,000	0,000	0,304	
FRT	$\hat{\beta_0}$=1,621	0,600	0,011	0,442
	$\hat{\beta_1}$=0,150	0,030	<0,001	
DEP	$\hat{\beta_0}$=4,893	0,356	<0,001	0,040
	$\hat{\beta_1}$=-0,001	0,001	0,263	
OFERECE	$\hat{\beta_0}$=4,485	0,229	<0,001	0,003
	$\hat{\beta_1}$=0,094	0,328	0,776	
REC30	$\hat{\beta_0}$=4,486	0,199	<0,001	0,005
	$\hat{\beta_1}$=0,000	0,000	0,696	
REC50	$\hat{\beta_0}$=4,490	0,216	<0,001	0,003
	$\hat{\beta_1}$=0,000	0,000	0,776	
REC100	$\hat{\beta_0}$=4,665	0,209	<0,001	0,032
	$\hat{\beta_1}$=0,000	0,000	0,317	

Tabela B.3: Regressão Linear Simples, tendo como variável resposta `vidro`.

Regressão Linear Simples				
vidro				
Variável	Estimativas	$\hat{\sigma}$	p-valor	R^2
MTH	$\hat{\beta_0}$=13,140	1,391	<0,001	0,010
	$\hat{\beta_1}$=0,039	0,071	0,586	
SAC15UD	$\hat{\beta_0}$=13,060	1,636	<0,001	0,008
	$\hat{\beta_1}$=0,002	0,004	0,618	
SAC15UND	$\hat{\beta_0}$=13,350	1,154	<0,001	0,008
	$\hat{\beta_1}$=0,009	0,019	0,626	
SAC30UD	$\hat{\beta_0}$=13,600	2,515	<0,001	0,000
	$\hat{\beta_1}$=-0,000	0,006	0,933	
SAC30UND	$\hat{\beta_0}$=8,861	1,723	<0,001	0,232
	$\hat{\beta_1}$=0,017	0,006	0,005	
SAC50UD	$\hat{\beta_0}$=13,663	1,121	<0,001	0,001
	$\hat{\beta_1}$=-0,007	0,040	0,871	
SAC50UND	$\hat{\beta_0}$=3,128	2,040	0,135	0,484
	$\hat{\beta_1}$=0,008	0,001	<0,001	
SAC100UD	$\hat{\beta_0}$=14,298	0,829	<0,001	0,010
	$\hat{\beta_1}$=-0,112	0,111	0,322	
SAC100UND	$\hat{\beta_0}$=3,157	1,592	0,056	0,609
	$\hat{\beta_1}$=0,015	0,002	<0,001	
uUDC	$\hat{\beta_0}$=11,801	4,409	0,012	0,007
	$\hat{\beta_1}$=0,029	0,063	0,648	
TAC	$\hat{\beta_0}$=1,503	2,696	0,581	0,411
	$\hat{\beta_1}$=0,656	0,141	<0,001	
TBC	$\hat{\beta_0}$=-0,466	3,740	0,902	0,325
	$\hat{\beta_1}$=0,595	0,154	0,001	
TCC	$\hat{\beta_0}$=12,082	1,499	<0,001	0,051
	$\hat{\beta_1}$=0,121	0,094	0,208	
TDC	$\hat{\beta_0}$=9,510	2,196	<0,001	0,119
	$\hat{\beta_1}$=0,565	0,276	0,049	
TEC	$\hat{\beta_0}$=15,254	2,339	<0,001	0,013
	$\hat{\beta_1}$=-0,209	0,324	0,523	
KMS	$\hat{\beta_0}$=13,500	1,611	<0,001	0,001
	$\hat{\beta_1}$=0,000	0,000	0,834	
FRT	$\hat{\beta_0}$=0,023	2,177	0,992	0,5746
	$\hat{\beta_1}$=0,709	0,110	<0,001	
DEP	$\hat{\beta_0}$=14,975	1,494	<0,001	0,024
	$\hat{\beta_1}$=-0,004	0,005	0,389	
OFERECE	$\hat{\beta_0}$=14,109	0,949	<0,001	0,007
	$\hat{\beta_1}$=-0,619	1,362	0,652	
REC30	$\hat{\beta_0}$=13,249	0,809	<0,001	0,046
	$\hat{\beta_1}$=0,002	0,002	0,231	
REC50	$\hat{\beta_0}$=12,880	0,859	<0,001	0,082
	$\hat{\beta_1}$=0,002	0,001	0,105	
REC100	$\hat{\beta_0}$=13,570	0,882	<0,001	0,006
	$\hat{\beta_1}$=0,000	0,002	0,669	

Apêndice C

Regressão Linear Múltipla

C.1 Resíduos de papel/cartão

Tabela C.1: Modelo de regressão linear múltipla para a produção de resíduos de papel/cartão.

PAPEL/CARTÃO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = -0,106$	3,309	0,975
SAC15UD	$\hat{\beta_1} = -0,013$	0,004	0,012
SAC30UD	$\hat{\beta_2} = -0,019$	0,007	0,014
SAC30UND	$\hat{\beta_3} = -0,014$	0,005	0,010
UDC	$\hat{\beta_4} = 0,141$	0,063	0,036
TAC	$\hat{\beta_5} = -0,246$	0,138	0,089
TDC	$\hat{\beta_7} = 0,471$	0,167	0,010
TEC	$\hat{\beta_8} = 0,562$	0,212	0,015
KMS	$\hat{\beta_9} = 0,002$	0,001	0,011
FRT	$\hat{\beta_{10}} = 0,302$	0,135	0,036
OFERECE:1	$\hat{\beta_{11}} = 2,643$	1,470	0,087
REC50	$\hat{\beta_{11}} = 0,002$	0,001	0,050
$R_a^2 = 0,526$			

C.2 Resíduos de vidro

Tabela C.2: Modelo de regressão linear múltipla para a produção de resíduos de vidro.

VIDRO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = 0,499$	2,718	0,856
SAC15UD	$\hat{\beta_1} = -0,009$	0,004	0,034
SAC30UD	$\hat{\beta_2} = -0,0236$	0,006	0,001
SAC30UND	$\hat{\beta_3} = 0,009$	0,004	0,023
UDC	$\hat{\beta_4} = 0,155$	0,056	0,011
TDC	$\hat{\beta_5} = 0,476$	0,156	0,006
FRT	$\hat{\beta_7} = 0,705$	0,094	<0,001
DEP	$\hat{\beta_8} = -0,010$	0,004	0,013
REC30	$\hat{\beta_9} = 0,002$	0,001	0,086
REC100	$\hat{\beta_{10}} = -0,002$	0,001	0,010
$R_a^2 = 0,8571$			

C.3 Resíduos de plástico

Tabela C.3: Modelo de regressão linear múltipla para a produção de resíduos de plástico.

PLÁSTICO			
	Estimativas	$\hat{\sigma}$	**Valor de prova**
	$\hat{\beta_0} = 2,371$	0,967	0,023
MTH	$\hat{\beta_1} = -0,048$	0,025	0,070
SAC15UD	$\hat{\beta_2} = -0,003$	0,001	0,016
SAC30UD	$\hat{\beta_3} = -0,004$	0,001	0,019
UDC	$\hat{\beta_4} = 0,030$	0,014	0,041
TAC	$\hat{\beta_5} = -0,092$	0,030	0,005
TDC	$\hat{\beta_7} = 0,158$	0,037	<0,001
TEC	$\hat{\beta_8} = 0,157$	0,040	<0,001
FRT	$\hat{\beta_9} = 0,273$	0,029	<0,001
OFERECE:1	$\hat{\beta_{10}} = -2,332$	0,624	0,001
REC50	$\hat{\beta_{11}} = -0,001$	0,000	0,012
REC100	$\hat{\beta_{12}} = -0,002$	0,000	<0,001
$R_a^2 = 0,877$			

C.4 Modelos de Regressão para a sazonalidade

Tabela C.4: Valores calculados a partir do modelo sazonal inicial, dos resíduos indiferenciados e seletivos, respetivamente.

Resíduos indiferenciados			**Resíduos seletivos**		
Estimativas	$\hat{\sigma}$	**Valor de prova**	**Estimativas**	$\hat{\sigma}$	**Valor de prova**
$\hat{\beta_0}$=50,801	1,363	<0,001	$\hat{\beta_0}$=24,566	0,620	<0,001
$\hat{\beta_1}$=-6,974	4,286	0,119	$\hat{\beta_1}$=-3,493	1,950	0,088
$\hat{\beta_2}$=18,199	4,286	<0,001	$\hat{\beta_2}$=1,714	1,950	0,389
$\hat{\beta_3}$=6,473	4,286	0,146	$\hat{\beta_3}$=5,781	1,950	0,007
$\hat{\beta_4}$=7,653	4,286	0,089	$\hat{\beta_4}$=9,348	1,950	<0,001
$\hat{\beta_5}$=2,599	4,286	0,551	$\hat{\beta_5}$=4,528	1,950	0,030
$\hat{\beta_6}$=0,553	4,286	0,899	$\hat{\beta_6}$=0,394	1,950	0,842
$\hat{\beta_7}$=1,479	4,286	0,733	$\hat{\beta_7}$=-2,726	1,950	0,177
$\hat{\beta_8}$=-2,147	4,286	0,622	$\hat{\beta_8}$=-4,359	1,950	0,036
$\hat{\beta_9}$=-3,681	5,160	0,483	$\hat{\beta_9}$=-3,106	1,950	0,200
$\hat{\beta_{10}}$=-15,121	5,160	0,008	$\hat{\beta_{10}}$=-5,616	2,347	0,026
$\hat{\beta_{11}}$=-8,521	5,160	0,114	$\hat{\beta_{11}}$=-2,886	2,347	0,233
$R^2 = 0,626$			$R^2 = 0,714$		