



**Ana Raquel
Ferreira Martins**

**Poluentes Atmosféricos e o seu Impacto a Curto
Prazo na Saúde: Um Estudo de Séries Temporais**

**Air Pollutants and their Short-Term Impact on
Health: A Time Series Analysis Approach**



**Ana Raquel
Ferreira Martins**

**Poluentes Atmosféricos e o seu Impacto a Curto
Prazo na Saúde: Um Estudo de Séries Temporais**

**Air Pollutants and their Short-Term Impact on
Health: A Time Series Study**

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica de Sónia Gouveia, Professora do Departamento de Matemática da Universidade de Aveiro.

Este trabalho foi desenvolvido no Instituto de Engenharia Electrónica e Informática de Aveiro (IEETA) e foi financiado pelos projetos UID/CEC/00127/2013 e UID/CEC/00127/2019 no âmbito da instituição de I & D.

o júri / the jury

presidente / president

Eugénio Rocha

Professor Professor Auxiliar da Universidade de Aveiro (por delegação da Reitora da Universidade de Aveiro)

vogais / examiners committee

Sónia Gouveia

Professora Auxiliar Convidada da Universidade de Aveiro (orientadora)

Argentina Soeima Leite

Professora Auxiliar do Departamento de Matemática da Universidade de Trás-os-Montes e Alto Douro (arguente)

agradecimentos / acknowledgements

À Professora Sónia Gouveia quero agradecer a orientação que me proporcionou. Em particular, agradeço a autonomia que me permitiu ter na tomada de decisões ao longo do desenvolvimento deste trabalho científico. Estou-lhe grata por todo o tempo que dispensou para refletir comigo nas questões colocadas, pelo seu auxílio na resolução de problemas e acima de tudo, por me ensinar no verdadeiro sentido da palavra. Não posso ainda deixar de lhe agradecer, por todos os outros momentos de conversa e pela partilha experiências. A Professora Sónia é, sem dúvida, um exemplo de uma excelente investigadora, professora e ser humano.

À Alexandra Monteiro, Investigadora Principal no Centro de Estudos do Ambiente e do Mar da Universidade de Aveiro, agradeço o apoio técnico que proporcionou relativamente à qualidade do ar em Portugal, nomeadamente a sua disponibilidade para esclarecer qualquer questão prontamente. Uma nota de agradecimento a todos os professores do mestrado que apesar do meu percurso distinto, me receberam com entusiasmo e se mostraram sempre disponíveis para me auxiliarem.

À minha colega Letícia por ter sido uma verdadeira companheira nos momentos de estudo, pela entreaajuda e pela amizade.

À Diana e ao João, por me acolherem no IEETA, obrigada por todas as conversas e companheirismo.

Não posso deixar de agradecer ao IEETA, a instituição de acolhimento, pelas condições proporcionadas para desenvolver este trabalho. E acima de tudo, agradeço a oportunidade de desenvolver este trabalho no âmbito de uma bolsa de investigação.

Olhando à espera privada, às minhas queridas amigas e amigos, por validarem a minha realidade, por serem pessoas extraordinárias e pelo seu apoio incondicional.

Aos meus pais, por me apoiarem nas minhas decisões, pela sua paciência e pelo altruísmo que tão bem vos caracteriza, um muito obrigada. Aos meus irmãos mais novos, Diogo e Rodrigo, porque não sou apenas eu um exemplo para vocês, também o são vocês para mim.

À minha madrinha Augusta, pelo exemplo de determinação que és, por trazeres luz nos momentos de maior indeterminação.

E por fim, mas não menos importante, ao Diogo, por continuares a ser minha constante, na inconstância que a vida é.

Resumo

Há já algumas décadas que a associação entre poluentes atmosféricos e a saúde foi estabelecida. Atualmente, é estimado que ocorram cerca de 9 mil milhões de mortes prematuras em todo o mundo (16% todas as mortes) devido à poluição. Para além do impacto na mortalidade, há também evidência que a poluição atmosférica está associada com *outcomes* de morbilidade, como, por exemplo, as admissões hospitalares. No sentido de controlar a poluição atmosférica, têm sido impostos limites legais por todo o mundo. No entanto, na Europa, incluindo Portugal, os limites legais para alguns poluentes são superiores aos recomendados pela Organização Mundial de Saúde. Apesar da comunidade científica estudar a poluição atmosférica há mais de 30 anos, novas evidências continuam a surgir relativamente aos impactos negativos que esta tem na saúde humana e também no ambiente. Em Portugal poucos estudos têm sido realizados para avaliar o impacto da poluição atmosférica na saúde. Assim, o objetivo deste estudo é, precisamente, estudar a relação entre os poluentes atmosféricos e a saúde humana, em particular nas admissões hospitalares resultantes de causas respiratórias no hospital de Aveiro entre 2013 e 2016. Adicionalmente, pretende-se produzir previsões das admissões hospitalares, o que é de especial importância para gerir os recursos hospitalares. Assim, para cumprir com estes objetivos, modelos SARIMA e ARFIMA foram combinados (SARFIMA) de modo a descrever os poluentes atmosféricos, enquanto que modelos lineares generalizados foram utilizados para modelar as admissões hospitalares utilizando os poluentes como covariáveis. Diferentes estratégias para realizar as previsões das admissões hospitalares foram exploradas, em particular recorreu-se aos dados originais dos poluentes e também aos valores previstos para os poluentes segundo os modelos SARFIMA.

Abstract

For some decades now, the association between air pollution and health has been established. Currently, it is estimated that there are around 9 million premature deaths worldwide (16% of all deaths) due to pollution. Besides the impact on mortality, there is also evidence that air pollution is associated with morbidity outcomes, such as hospital admissions. Efforts to control air pollution have been made worldwide by implementing legal limits. Nevertheless, in Europe, Portugal included, the limits for some air pollutants are higher than the recommended by the World Health Organisation. Despite the scientific community being studying air pollution effects on health for over 30 years, new evidence of its harmful impact on human health and the environment continues to emerge. In Portugal, few studies on the impact of air pollution on health have been conducted. Hence, the goal of this work is to study the relationship between air pollution and health, particularly on respiratory hospital admissions in Aveiro hospital between 2013 and 2016. Furthermore, this study aims at producing hospital admissions forecasts, which is of considerable importance to manage hospital resources. To accomplish these goals, SARIMA and ARFIMA models (SARFIMA) were combined to describe air pollutants data, whilst generalised linear models were used to model hospital admissions using air pollutants as covariables. Different hospital admissions forecasts strategies were explored, namely the use of the observed air pollutants' data and the predicted pollutants' values according to the SARFIMA models.

Contents

Contents	i
Acronyms	iii
List of Figures	v
List of Tables	ix
1 Introduction	1
1.1 Air Pollution & Health	3
1.2 Objectives	14
1.3 Data	15
1.3.1 Air Pollutants	15
1.3.2 Hospital Admissions	19
2 Methods	21
2.1 Continuous Time Series	23
2.1.1 Theoretical Aspects	23
2.1.2 Practical Implementation Aspects	38
2.2 Discrete Time Series	47
2.2.1 Theoretical Aspects	47
2.2.2 Practical Implementation Aspects	49
3 Results	53
3.1 Air Pollutants	55
3.1.1 Descriptive Results	55
3.1.2 Building a Modelling Framework	61
3.1.3 Model Forecast	77
3.2 Hospital Admissions in Aveiro	81

3.2.1	Descriptive Results	81
3.2.2	Model Fit	89
3.2.3	Model Forecast Strategies	93
4	Conclusions	95
	Bibliography	99
	Appendices	105
A	Introduction	107
A.1	Air Pollutants Legal Limits	109
A.2	Air Pollutants Data Preprocessing	115
B	Methods	119
B.1	ARFIMA estimation	121
C	Results	125
C.1	Air Pollutants Descriptive Statistics	127
C.2	ARFIMA Models	129
C.3	SARFIMA Models	131

Acronyms

ACF Autocorrelation Function.

AR Autoregressive.

ARCH Autoregressive Conditional Heteroskedasticity.

ARFIMA Fractional Autoregressive Integrated Moving Average.

ARIMA Autoregressive Integrated Moving Average.

ARMA Autoregressive Moving Average.

CCF Cross-correlation Function.

CO Carbon Monoxide.

DHR Dynamic Harmonic Regression.

GARCH Generalised Autoregressive Conditional Heteroskedasticity.

GLM Generalised Linear Model.

INARMA Integer Autoregressive Integrated Moving Average.

INGARCH Integer-valued Generalised Autoregressive Conditional Heteroskedasticity.

MA Moving Average.

NO₂ Nitrogen Dioxide.

NO_x Nitrogen Oxides.

O₃ Ozone.

PACF Partial Autocorrelation Function.

PM₁₀ Particulate Matter with a diameter inferior to $10\mu m$.

PM_{2.5} Particulate Matter with a diameter inferior to $2.5\mu m$.

SARFIMA Seasonal Fractional Autoregressive Integrated Moving Average.

SARIMA Seasonal Autoregressive Integrated Moving Average.

SARMA Seasonal Autoregressive Moving Average.

SO₂ Sulphur Dioxide.

WHO World Health Organisation.

List of Figures

1.1	Approximate weekly mortality and SO ₂ concentrations for Greater London, 1952-1953, just before and during the episode. Reproduced from [7].	3
1.2	NO ₂ exceedances (2005-2016). (a) number of hourly exceedances per year (b) annual average. The red line identifies the legal limit.	7
1.3	NO _x annual average (2005-2016). The red line is the critical level established.	7
1.4	PM ₁₀ exceedances (2005-2016). (a) number of hourly exceedances per year. (b) annual average. The red line identifies the legal limit.	8
1.5	PM _{2.5} annual average (2005-2016). The red line is the target value in 2010 and mandatory legal limit since 2015, whereas the grey line is the mandatory legal limit for 2020.	9
1.6	O ₃ exceedances (2005-2016). (a) number of daily maximum 8h-mean exceedances, (b) one hour exceedances of the information threshold. The red line represents the maximum number of exceedances allowed.	10
1.7	Estimated excess mortality attributed to air pollution in Europe, and the contributing disease categories. Reproduced from [37].	11
1.8	Geographic distribution of air pollutants stations.	17
1.9	Boxplot of the ratio between the mean of the imputed series and the mean of the series with missing data (upper panel). Same representation for the standard deviation (lower panel). (a) SO ₂ , (b) PM ₁₀	18
1.10	Example of an imputed time series with the 1-NN method.	19
1.11	Location of Aveiro hospital and the nearest air pollution stations.	20
2.1	Gaussian white noise time series.	23
2.2	ACF of <i>varve</i> time series from the <i>astsa</i> R package. (a) original time series, (b) differenced time series. The blue lines represent the limits $\frac{\pm 2}{\sqrt{n}}$, in which lie 95% of observations if the time series resembles a white noise. (To be continued)	30
2.3	Box-Jenkins methodology for model selection. Adapted from [45].	32

3.1	Time series of air pollutants with background influence. (a) rural , (b) suburban and (c) urban. (To be continued)	55
3.2	Distribution of time series mean according to type of background and environment of the 134 stations. (a) NO ₂ , (b) NO _x , (c) PM ₁₀ , (d) PM _{2.5} , (e) O ₃ , (f) SO ₂ , (g) CO. (To be continued)	59
3.3	Time Series of CO at Aveiro station. (a) Complete series, (b) Weekly data from 1 January 2005 to 7 January 2005.	61
3.4	Characteristics of CO time series at Aveiro station. (a) ACF, (b) PACF. . . .	62
3.5	CO Aveiro station transformed time series. (a) ACF, (b) PACF. <i>Upper Panel</i> - Differenced time series, <i>Lower Panel</i> - Box-Cox transformation and differenced time series.	62
3.6	Residuals analysis of SARIMA model. (a) distribution, (b) ACF, (c) PACF. . .	64
3.7	Analysis of CO DHR model residuals at Aveiro station. (a) distribution, (b) ACF, (c) PACF.	67
3.8	Periodogram of CO time series at Aveiro station. (a) original time series, (b) DHR fitted time series.	68
3.9	Analysis of CO ARFIMA model residuals at Aveiro station.(a) distribution, (b) ACF, (c) PACF.	70
3.10	Distribution of d parameter for all time series according to type of environment and influence. (a) NO ₂ , (b) NO _x , (c) PM ₁₀ , (d) PM _{2.5} , (e) O ₃ , (f) SO ₂ , (g) CO. (To be continued)	71
3.11	Periodogram of ARFIMA fitted time series of CO at Aveiro station.	73
3.12	Analysis of CO SARFIMA model residuals at Aveiro station. (a) distribution, (b) ACF, (c) PACF.	75
3.13	Periodogram of SARFIMA fitted values of CO at Aveiro station.	76
3.14	Forecasts of SARFIMA model for Aveiro station, CO pollutant.	77
3.15	MASE Distribution according to type of environment and influence. (a) NO ₂ , (b) NO _x , (c) PM ₁₀ , (d) PM _{2.5} , (e) O ₃ , (f) SO ₂ , (g) CO. (To be continued) . .	79
3.16	Annual distributions of the daily number of hospital admissions (2013-2016) for each ICD-9 code (Table 1.4)	82
3.17	Time series of daily hospital admissions (all ICD-9 codes combined).	83
3.18	Daily hospital admission. (a) ACF, (b) PACF.	83
3.19	Daily maximum value air pollutants time series. (a) NO ₂ , (b) NO _x , (c) PM ₁₀ , (d) PM _{2.5} , (e) O ₃ , (f) SO ₂ , (g) CO. (To be continued)	84
3.20	CCF between Aveiro hospital admissions and air pollutants in the nearest stations. (a) NO ₂ , (b) NO _x , (c) PM ₁₀ , (d) PM _{2.5} , (e) O ₃ , (f) SO ₂ , (g) CO. . . .	86

3.21	Analysis of residuals of hospital admissions. (a) Link Function - Identity, (b) Link Function - Logarithmic.	91
3.22	Prediction intervals at 95%. (a) using observed NO _x and CO, (b) using forecasts from NO _x and CO.	93
3.23	Time series of daily hospital admissions and forecasts by air pollutants covariates with observed covariates and predicted SARFIMA covariates.	94
A.1	Boxplot of the ratio between the mean of the imputed series and the mean of the series with missing data (upper panel). Same representation for the standard deviation (lower panel). (a) NO ₂ , (b) NO _x , (c) PM ₁₀ , (d) O ₃ , (e) CO. (To be continued)	116

List of Tables

1.1	Recommended air quality limit values by the WHO [62].	5
1.2	Air quality legal limits established in the Directive 2008/50/EC [2] and DL 102/2010 [3].	6
1.3	Characteristics and number of included time series.	16
1.4	Description of the International Classification of Diseases 9th Revision codes [59].	20
2.1	ACF and PACF behaviour for purely ARMA models. Reproduced from [50].	33
2.2	ACF and PACF behaviour for purely SARMA models. Reproduced from [50].	33
3.1	Coefficients of SARIMA model for CO time series at Aveiro station.	63
3.2	Best ARIMA model and Fourier terms obtained by the established framework.	66
3.3	DHR coefficients for CO time series at Aveiro station.	68
3.4	Hurst Exponent for Aveiro station, CO air pollutant.	69
3.5	Coefficients of ARFIMA model for CO time series at Aveiro station.	69
3.6	Coefficients of SARFIMA model CO time series at Aveiro station.	76
3.7	Performance measures and percentage of observations within the prediction intervals for CO at Aveiro station.	78
3.8	Pearson correlation between air pollutants lagged time series. Bold indicates correlations higher than 0.50.	88
3.9	Coefficients and adequacy measures for the generalised linear models. Model I - Identity Link Function, Model L - Logarithmic Link Function.	90
3.10	AIC and BIC information criteria for the optimal model and 3 alternative models.	92
3.11	Performance measures and percentage of observations within the prediction intervals of one month forecasts for each model.	94
A.1	Number of one hour limit value exceedances of NO ₂ for all stations (2005-2016). Bold numbers surpass the annual maximum number of exceedances (18). . . .	109

A.2	Mean annual NO ₂ value for all stations (2005-2016). Bold numbers exceed the calendar year value ($40\mu g/m^3$).	110
A.3	Number of one hour limit value exceedances of PM ₁₀ for all stations (2005-2016). Bold numbers surpass the annual maximum number of exceedances (35).	111
A.4	Mean annual PM ₁₀ value for all stations (2005 and 2016). Bold numbers exceed the calendar year value ($40\mu g/m^3$).	112
A.5	Number of daily maximum 8h-mean O ₃ exceedances for all stations (2005-2016). Bold numbers surpass the annual maximum value allowed (25).	113
A.6	Number of one hour exceedances of the Information Alert of O ₃ ($180\mu g/m^3$), for all stations (2005-2016).	114
B.1	Model coefficients, standard error, <i>z-value</i> and <i>p-value</i> for NO _x at Arcos Station.	122
B.2	Model coefficients, standard error, <i>z-value</i> and <i>p-value</i> for NO ₂ at Custóias Station.	123
B.3	Model coefficients, standard error, <i>z-value</i> and <i>p-value</i> for NO ₂ at Chamusca Station.	123
C.1	Mean (\bar{x}) and standard deviation (σ) of each pollutant time series.	128
C.2	Hurst Exponent for all pollutants time series at each station.	130
C.3	SARFIMA models coefficients for each station - NO ₂	132
C.4	SARFIMA models coefficients for each station - NO _x	133
C.5	SARFIMA models coefficients for each station - PM ₁₀	134
C.6	SARFIMA models coefficients for each station - PM _{2.5}	135
C.7	SARFIMA models coefficients for each station - O ₃	136
C.8	SARFIMA models coefficients for each station - SO ₂	137
C.9	SARFIMA models coefficients for each station - CO.	137
C.10	SARFIMA performance measures for each station - NO ₂	138
C.11	SARFIMA performance measures for each station - NO _x	139
C.12	SARFIMA performance measures for each station - PM ₁₀	140
C.13	SARFIMA performance measures for each station - PM _{2.5}	140
C.14	SARFIMA performance measures for each station - O ₃	141
C.15	SARFIMA performance measures for each station - SO ₂	142
C.16	SARFIMA performance measures for each station - CO.	142

Chapter 1

Introduction

Air pollution impact on health has been studied for the last couple of decades and its negative impact on health is well-established. Nonetheless, new evidence of the harmful impacts of air pollution on health continues to emerge even at low levels, such as those recommended by the World Health Organisation (WHO). Some of the current legal limits in Europe, Portugal included, are higher than those proposed by the WHO. The literature on air pollution impact on health is scarce in Portugal. Hence the objective of this research work is to study the short-term impact of air pollution on respiratory hospital admissions in Aveiro, Portugal. An additional goal is to explore forecast strategies of hospital admissions, which are of particular interest to manage hospital resources. To accomplish these goals a framework to model air pollutants using ARIMA-based models is developed and generalised linear (GLM) models are used to model and forecast hospital admissions.

1.1 Air Pollution & Health

Evidence that air quality is associated with increased mortality arose in the early XX century as a result of winter smog episodes. Previously, it was thought that the increase in mortality resulted solely from low temperatures. A particularly important episode that established such association was the prolonged and intense air pollution episode that occurred in London between 5 and 9 of December of 1952. This episode was characterised by a lengthy temperature inversion with little to no wind, which enabled the accumulation of pollutants from domestic coal burning. The smoke concentrations were around 3- to 5-fold higher than the expected for this period [6]. Figure 1.1 shows the approximate weekly total mortality and weekly average SO_2 concentrations for Greater London during the episode. Later studies on this smog episode estimate that there was an excess of 12,000 deaths from December 1952 until February 1953 due to its acute and persisting effects [7]. Bell and Davis (2001) reported that the average number of deaths before the smog episode were 1,570 per week. By 13 of December, the number of deaths was over two-fold higher, reaching nearly 5,000 deaths per week and an increase of about 50% in mortality persisted until late February. Note that SO_2 was not the only pollutant affecting air quality but, at this time, only SO_2 was routinely monitored [7].

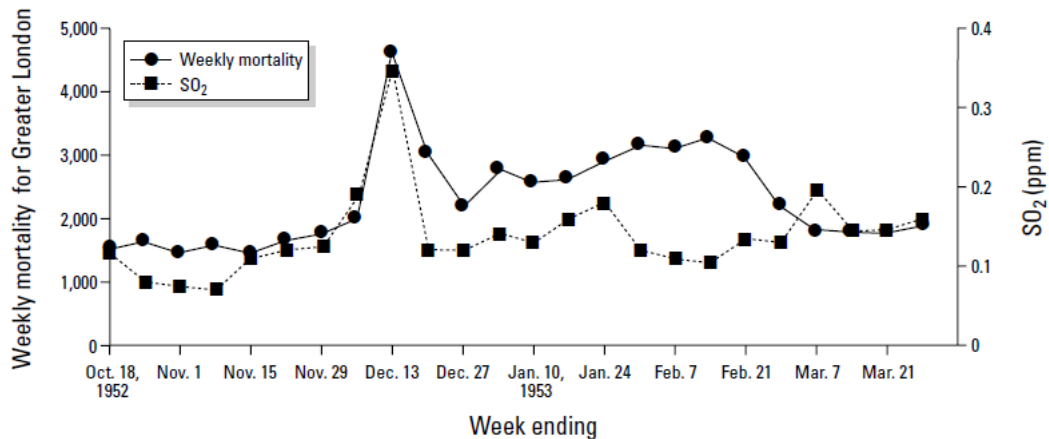


Figure 1.1: Approximate weekly mortality and SO_2 concentrations for Greater London, 1952-1953, just before and during the episode. Reproduced from [7].

Afterwards, the interest in studying the association between air pollution and health outcomes, namely mortality, rose considerably. In 1987, the World Health Organisation (WHO) published the first edition of the Air Quality Guidelines for Europe, in which the health risks of 27 pollutants were assessed. Among these pollutants were the classical air pollutants, NO_2 , PM_{10} , $\text{PM}_{2.5}$, O_3 and SO_2 , and the organic pollutant CO . These guidelines aimed at provid-

ing a basis for protecting public health from the adverse effects of environmental pollutants, including air pollutants, in order to eliminate or reduce to a minimum air pollutants' exposure likely to be hazardous to human health or well-being [58]. Later on, due to new developments in risk assessment and new scientific data on air pollution toxicology and epidemiology, the WHO decided it was necessary to review and update the Air Quality Guidelines. Therefore, in 1993 a Planning Meeting took place to define a framework to review the available data, which was completed by 2000, resulting in the publication of the second edition of the guidelines.

In the 2000 edition, besides reviewing mortality data, the impact of air pollution on morbidity was also evaluated. For instance, there was evidence that an increase of $10\mu g/m^3$ in PM_{10} increased the relative risk of use of bronchodilators, and a 20% increase in the number of hospital admissions were reported for an increase of $120\mu g/m^3$ of O_3 [60]. Research on air quality continued to grow rapidly due to the awareness of the impact of air pollution on health among scientists and policy-makers. Furthermore, the WHO carried out the project "Systematic review of health aspects of air pollution in Europe" under the European Commission's Clean Air for Europe (CAFE) programme [61]. The results of this project showed that it was necessary to conduct a further revision of the second edition of the air quality guidelines, particularly to NO_2 , PM, O_3 and SO_2 . In fact, this review of the guidelines considered data at a global scale and not only data from Europe, since air pollution research had grown considerably worldwide. Hence, a new update to the previous guidelines was completed in 2005 focusing on the previous air pollutants at a global scale [62].

Despite WHO developing guidelines for air quality, countries may or may not embody these guidelines in their legislation. The following definitions are essential in order to understand the level of scientific knowledge associated with each legal limit established and the possible impact of surpassing the established air pollutants levels:

- Limit Value - a level fixed on the basis of scientific knowledge, with the aim of avoiding, preventing or reducing harmful effects on human health and/or the environment as a whole, to be attained within a given period and not to be exceeded once attained;
- Alert Threshold - a level beyond which there is a risk to human health from brief exposure for the population as a whole and at which immediate steps are to be taken by the Member States;
- Information Threshold - a level beyond which there is a risk to human health from brief exposure for particularly sensitive sections of the population and for which immediate and appropriate information is necessary;
- Critical Level - a level fixed on the basis of scientific knowledge, above which direct

adverse effects may occur on some receptors, such as trees, other plants or natural ecosystems but not on humans.

These definitions here presented are in accordance to the 2008/50/EC Directive on ambient air quality and cleaner air for Europe [2].

Table 1.1 shows the recommended limit values by WHO for PM₁₀, PM_{2.5} and NO₂ (calendar year), O₃ and CO (daily maximum 8h-mean) and SO₂ (daily average), in order to mitigate their impact on human health [62]. In contrast, Table 1.2 shows the legal limits established for air pollutants by the European Union. Bold numbers correspond to the limits recommended by WHO in Table 1.1. It is clear that the current European legal limits, established in the Directive 2008/50/EC, are higher than the recommended by WHO for all pollutants, except for NO₂ and CO. The calendar year legal limit, both for PM₁₀ and PM_{2.5}, is twice higher than the recommended by the WHO. Also, O₃ daily maximum 8h-mean established by the EU is 20% higher than the guideline proposed by the WHO. Lastly, for SO₂, the current legal limit corresponds to the interim goal established by the WHO in 2005 ($125\mu g/m^3$) [62]. Hence, it is clear that further efforts must be done in order to meet the recommended WHO guidelines. Additionally, it is noteworthy that, due to the recent body of knowledge published regarding the effects of ambient air pollutant on health, the WHO has organised a global consultation meeting to seek an expert opinion as to whether or not re-evaluate the evidence of air quality impact on health. This group of experts considered that there were new evidence that justified the review of the guidelines for PM₁₀, PM_{2.5}, O₃, NO₂, SO₂ and CO. Hence, a new revision of the guideline is currently taking place, thus, it is possible that the current legal limits become even more outdated.

	NO ₂ ^a	PM ₁₀ ^a	PM _{2.5} ^a	O ₃ ^b	SO ₂ ^c	CO ^b
Recommended Limit	40 $\mu g/m^3$	20 $\mu g/m^3$	10 $\mu g/m^3$	100 $\mu g/m^3$	20 $\mu g/m^3$	10 mg/m^3

^a Calendar year ^b Daily maximum 8h-mean. ^c Daily average

Table 1.1: Recommended air quality limit values by the WHO [62].

Pollutant	Limit Value				Alert Threshold 1h	Information Threshold 1h ^c	Critical Level 1 year ^b
	1h	8h-mean ^a	1 day	1 year ^b			
NO ₂	200 $\mu\text{g}/\text{m}^3$ (^d)			40 $\mu\text{g}/\text{m}^3$	400 $\mu\text{g}/\text{m}^3$		
NO _x							30 $\mu\text{g}/\text{m}^3$
PM ₁₀	50 $\mu\text{g}/\text{m}^3$ (^e)			40 $\mu\text{g}/\text{m}^3$			
PM _{2.5}				25 $\mu\text{g}/\text{m}^3$ (^f) (20 $\mu\text{g}/\text{m}^3$)(^g)			
O ₃		120 $\mu\text{g}/\text{m}^3$ (^h)			240 $\mu\text{g}/\text{m}^3$	180 $\mu\text{g}/\text{m}^3$	
SO ₂	350 $\mu\text{g}/\text{m}^3$ (ⁱ)		125 $\mu\text{g}/\text{m}^3$ (^j)		500 $\mu\text{g}/\text{m}^3$		20 $\mu\text{g}/\text{m}^3$
CO		10 mg/m^3					

^a Daily maximum 8h-mean ^b Calendar year

^c The exceedance of the threshold is to be measured or predicted for three consecutive hours.

^d Not to be exceeded more than 18 times a calendar year.

^e Not to be exceeded more than 35 times a calendar year.

^f Target value in 2010 and mandatory limit in 2015. ^g Limit value for 2020.

^h Not to be exceeded more than 25 times a calendar year.

ⁱ Not to be exceeded more than 24 times a calendar year.

^j Not to be exceeded more than 3 times a calendar year.

Table 1.2: Air quality legal limits established in the Directive 2008/50/EC [2] and DL 102/2010 [3].

Air pollutants legal limits in Portugal are laid down in the Portuguese Law DL 102/2010 [3], which were transposed from the European Directive 2008/50/EC [2]. Figure 1.2 presents the behaviour of 8 monitoring stations in Portugal mainland for NO₂. Figure 1.2(a) depicts the number of hourly measurements above the limit value (200 $\mu\text{g}/\text{m}^3$) each year. The number of exceedances above 18 has been decreasing, in fact, in 2016 the ceiling was not exceeded. On the contrary, annual mean values surpass the established ceiling throughout the entire time period for Av. Liberdade, Fr. Sá Carneiro and Fr. Bartolomeu stations [Figure 1.2(b)]. These monitoring stations are located in urban areas with a large traffic influence and since combustion processes release NO₂, it is no surprise that these locations have such high annual mean values [62]. Regarding the threshold alert, this value is never exceeded, thus no graphical representation is presented.

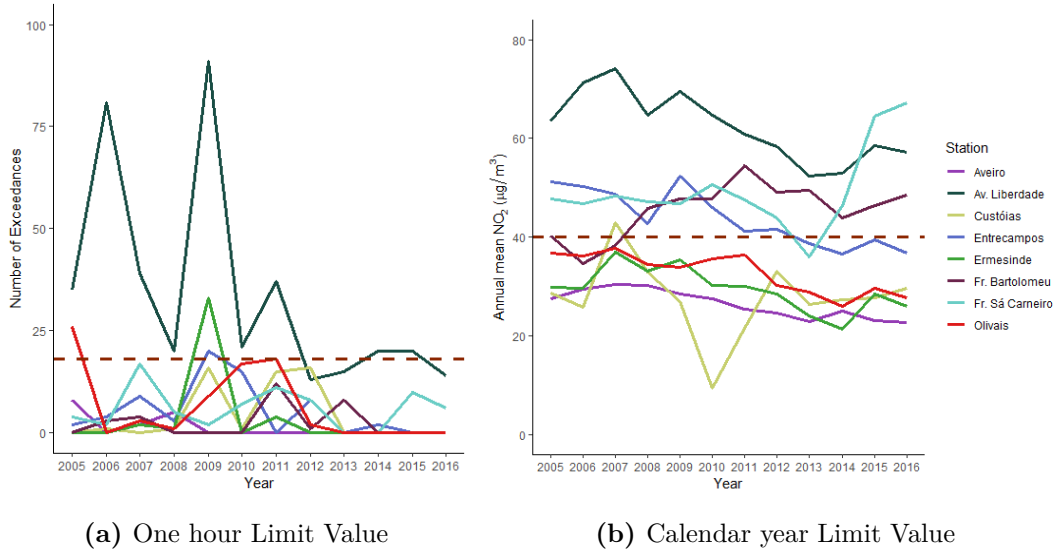


Figure 1.2: NO₂ exceedances (2005-2016). (a) number of hourly exceedances per year (b) annual average. The red line identifies the legal limit.

Similarly to NO₂, results for NO_x are presented in Figure 1.3 for the same stations as above. The exceedances can be two to three times higher than the critical level ($30\mu\text{g}/\text{m}^3$), particularly for Av. Liberdade, Custóias, Entrecampos, Fr. Bartolomeu and Fr. Sá Carneiro (Figure 1.3), which are located in areas with large traffic influence, as already mentioned. Even though the exceedances are quite large for critical values of NO_x, according to scientific evidence, these do not have a direct impact on human health and well-being [2, 3].

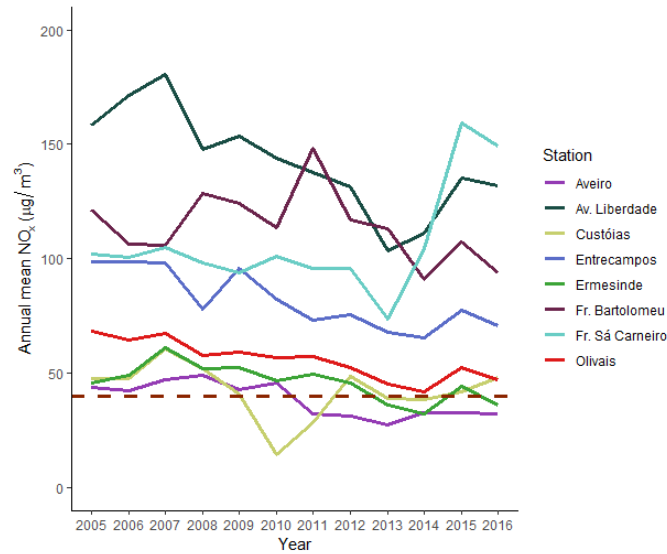


Figure 1.3: NO_x annual average (2005-2016). The red line is the critical level established.

Regarding PM_{10} , results for 10 mainland stations are presented in Figure 1.4. Despite the observed exceedances at Av. Liberdade and Aveiro stations, an overall decreasing trend is observed from 2005 up to 2016. In fact, in the latter year, there were less than 35 exceedances above $50\mu g/m^3$ in all stations. With respect to the annual mean, there is also a diminishing trend of its values and, since 2013 all stations complied with the established limit of $40\mu g/m^3$ [Figure 1.4(b)]. However, if one considers the limit value recommended by the WHO ($20\mu g/m^3$) in Figure 1.4(b) the annual mean is above this threshold up to 2012 for all stations. From this year onward, only half of the presented stations comply with the WHO guideline, with Aveiro, Av. Liberdade, Meco, Entrecampos and Estarreja still not complying with the recommended WHO guideline.

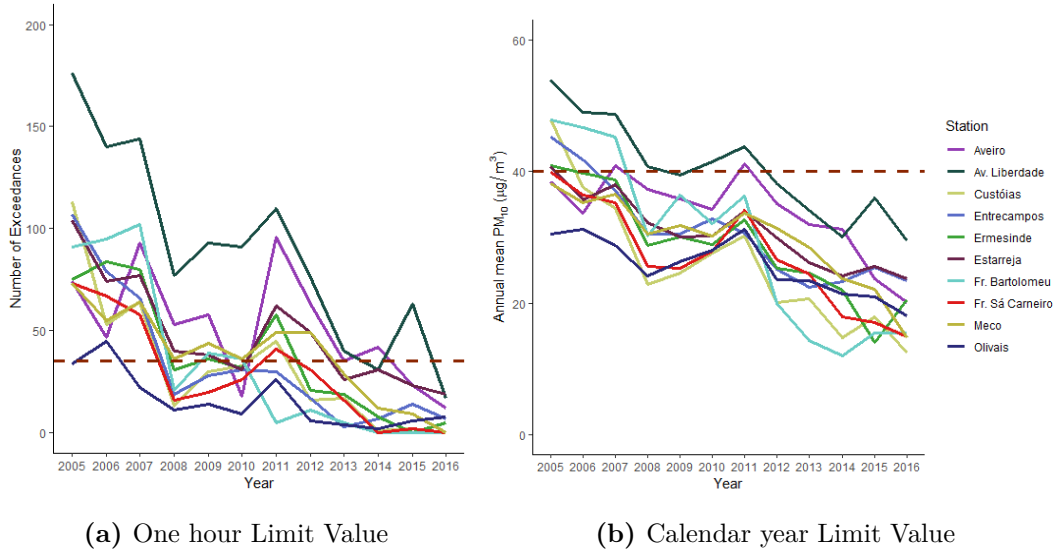


Figure 1.4: PM_{10} exceedances (2005-2016). (a) number of hourly exceedances per year. (b) annual average. The red line identifies the legal limit.

As for $PM_{2.5}$, Figure 1.5 shows the annual mean for 7 stations. One can observe that since 2007 all selected stations comply with the target value for 2010, which is the current mandatory limit value. Furthermore, annual means are in compliance with the legal limit to be established for 2020. Nevertheless, if one were to consider the guideline from WHO ($10\mu g/m^3$) at least half of the stations at a given year would be non-compliant.

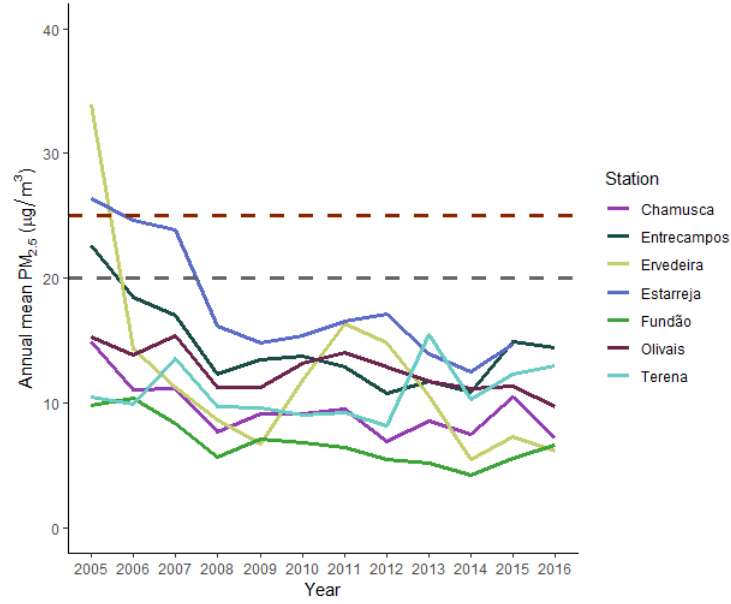


Figure 1.5: PM_{2.5} annual average (2005-2016). The red line is the target value in 2010 and mandatory legal limit since 2015, whereas the grey line is the mandatory legal limit for 2020.

Ozone legal limits are established for daily maximum 8h-mean, information and alert thresholds (Table 1.2). Figure 1.6(a) shows the number of exceedances above $120\mu\text{g}/\text{m}^3$ of O₃ for 8 mainland stations. At Chamusca and Ervedeira, the limit is frequently exceeded. As for the information threshold, the highest value of annual exceedances is found for Ílhavo station in 2006, for which the information alert was exceeded nearly 60 times [Figure 1.6(b)]. Overall, a decreasing trend is seen from 2005 up to 2015, year in which these stations do not exceed the O₃ information alert. However, in 2016, in Fornelo do Monte, Chamusca, Ervedeira, Estarreja and Ílhavo the legal limit was exceeded. Curiously, all exceedances occurred mainly between June and September. As for the alert threshold, this was surpassed only once, at Estarreja and Fornelo do Monte stations, in 2005 and 2016, respectively.

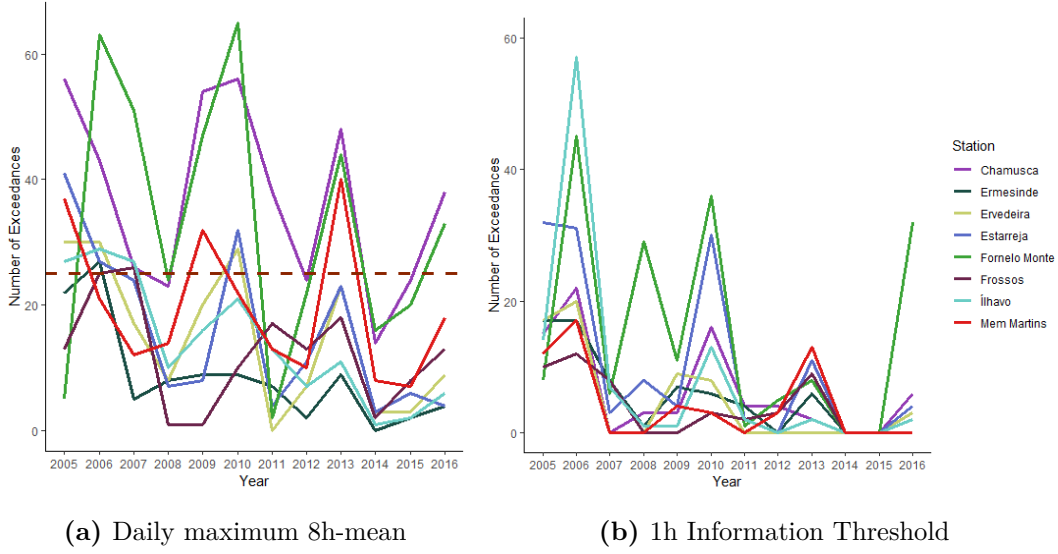


Figure 1.6: O_3 exceedances (2005-2016). (a) number of daily maximum 8h-mean exceedances, (b) one hour exceedances of the information threshold. The red line represents the maximum number of exceedances allowed.

Regarding SO_2 , of the 12 mainland monitoring stations considered, only one exceeded the legal limits established in Table 1.1. This was Lavradio station, which exceeded the hourly limit 52 and 29 times in 2007 and 2008, respectively. The daily limit value was surpassed in 2005, 2007 and 2008, respectively, 4, 12 and 5 times. Furthermore, the alert threshold and the critical level were exceeded once only, in 2007 at Lavradio. Lastly, for CO the legal limit established in Table 1.1 was analysed for eight stations. In this time period no station surpassed the ceiling value. Therefore, it is fair to say that air pollutants levels in Portugal are generally within the established by the European directive and the Portuguese law [2, 3]. Nevertheless, air pollutants levels are higher than the guidelines recommended by WHO for some pollutants, such as $PM_{2.5}$ and PM_{10} . Detailed information of air pollutants exceedances can be found in Table A.1 up to Table A.6.

In 2015, diseases caused by pollution were responsible for an estimated 9 million of premature deaths worldwide, which corresponds to 16% of all deaths. This number is three times higher than the premature deaths caused by AIDS, tuberculosis and malaria combined [36]. In Europe alone, the excess mortality rate due to air pollution is estimated to be near 800 000 deaths [37]. Figure 1.7 shows the estimated percentage of deaths due to different causes as a result of air pollution. Forty percent of the deaths attributed to air pollution in Europe are due to cardiovascular outcomes. Respiratory and cardiovascular outcomes were the first to be explored due to the clear link between air pollution and these biological systems, for which

evidence is compelling. For instance, a systematic review and meta-analysis study found that short-term exposure to a $10\mu\text{g}/\text{m}^3$ increment of ambient $\text{PM}_{2.5}$ is associated with increased Chronic Obstructive Pulmonary Disease (COPD) hospitalisation and mortality [38]. Furthermore, another systematic review found that an increase of $1\text{mg}/\text{m}^3$ of CO was associated with COPD admissions [44]. Additionally, it was also found a mortality increase of 6% in the European Union as a result of a short-term exposure to a PM_{10} increment of $10\mu\text{g}/\text{m}^3$, while hospital admission increased by 1% [51]. Regarding asthma, NO_2 , SO_2 , O_3 , PM_{10} , $\text{PM}_{2.5}$ and CO were significantly associated with increased risk of asthma-related emergency room visits and hospitalisations [70]. With respect to cardiovascular outcomes, there is strong evidence that increased levels of PM_{10} , NO_2 and CO can have adverse effects on cardiovascular function in people living with coronary heart disease [55]. A systematic review and meta-analysis has shown that hospitalisation and death resulting from heart failure are associated with significant increases in CO, NO_2 , $\text{PM}_{2.5}$ and PM_{10} [49].

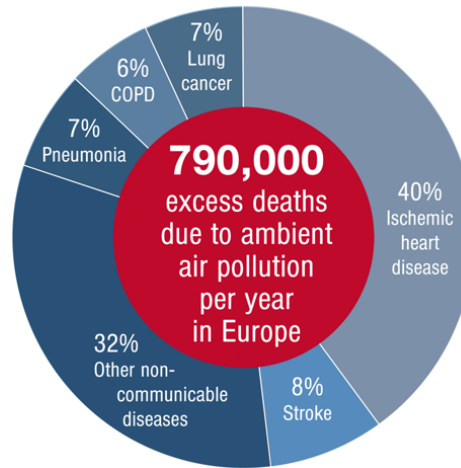


Figure 1.7: Estimated excess mortality attributed to air pollution in Europe, and the contributing disease categories. Reproduced from [37].

More recently, besides exploring respiratory and cardiovascular outcomes, researchers have extended their interest into other diseases, such as neurological disorders [20], mental health [69], reproductive health [11], rheumatic diseases [53], blood pressure [66], kidney diseases [65], thrombosis [47] and even adverse birth outcomes [68, 46, 52]. Significant associations have been found between some air pollutants and many of these diseases. For example, $\text{PM}_{2.5}$ is significantly associated with dementia, stroke, Alzheimer’s disease and also Parkinson’s disease [20]. On the contrary, no conclusive statistical evidence has been found between O_3 and mental health [69]. It is of the utmost importance to mention that the International Agency for Research on Cancer (IARC) has recently classified air pollution and the PM

mixture as carcinogenic, with evidence of increased risk of cancer even at levels below the current WHO PM_{2.5} guideline [40, 63]. Furthermore, as Figure 1.7 indicates, 32% of deaths due to ambient air pollution are from non-communicable diseases, which further justifies the study of these diseases.

Despite the compelling evidence between air pollution exposure and both mortality and morbidity, very few studies have been performed in Portugal. In fact, a fairly extensive search found only a couple of studies performed in Portugal and all of these were limited to Lisbon area [4, 15, 22, 14]. Alves *et al.* (2010) studied daily hospital admissions due to cardiorespiratory diseases and levels of PM₁₀, SO₂, CO, NO, NO₂, and O₃ between 1999 and 2004. They found significant positive associations, lagged by 1 or 2 days, between cardiocirculatory hospital admissions and both CO and NO₂ for all age groups [4]. Garret and Casimiro (2011) studied the impact of ambient O₃ and PM_{2.5} concentrations on daily mortality, between 2004 and 2006, in Lisbon and found that the population group aged over 65 years old had increased relative risk of cardiovascular mortality for each 10 $\mu\text{g}/\text{m}^3$ of PM_{2.5} increase. They also reported that O₃ exposure was associated with increased all-cause mortality in the population group ≥ 65 years old and the general population [22]. Cruz *et al.* (2015) aimed at studying the influence of air pollutants (CO, NO, NO₂, SO₂, O₃, PM₁₀ and PM_{2.5}) on respiratory and cardiovascular hospital admissions in Lisbon city from 2006 to 2008. Positive significant associations between cardiovascular diseases for ages between 15 and 64 and the pollutants CO, NO, NO₂, SO₂ and PM₁₀ were found. In particular, hospital admissions due to stroke for people aged over 64 years old increased with a 10 $\mu\text{g}/\text{m}^3$ increase of NO₂. Furthermore, associations between respiratory admissions for ages below 15 and CO, NO, NO₂, SO₂ and PM₁₀ were also found [14].

Therefore, there is a lack of scientific knowledge on this particular subject in Portugal, which has a large potential to be filled up, since there exist many air quality monitoring stations in Portugal as a result of the requirements from the European Directive 2008/50/EC and the Portuguese law DL 102/2010 [2, 3]. Also, even though the aforementioned studies are fairly recent, the hospital admission data used is over 10 years old. The Portuguese health ministry collects data on cause of hospital admissions and also cause of mortality, thus it is possible to use more recent data, which is of greater interest. Hence, there is data available so that new scientific evidence can be added to the current body of knowledge of air pollution research and its impacts on health. Even though air pollution is a topic that has been studied since around the 1950s, it is still a subject of enormous interest, with growing numbers of publications in the last few years. Furthermore, recent evidence shows that pollution levels' as low as the WHO recommends still have negative impacts on health. As shown, in Portugal air pollution levels generally comply with the established limits but

not with the WHO recommended values, which further justifies the emergence in studying the impact of air quality on health in Portugal.

In order to study the relationship between air pollution exposure and health outcomes, time series analysis is essential, since it is imperative to understand their relationship throughout time. Time series analysis can be used to simply describe the behaviour of a time series, to build a model that explains the variation of a time series as a result of other time series used as covariables or even to make forecasts, that is, to predict the near future observations. In this context, it is of interest to understand the overall trend of air pollutants and hospital admissions from a descriptive point of view. Nevertheless, this information is truly useful when combined, that is, when hospital admissions are studied as a result of air pollutants variation. Forecasts of hospital admissions can be particularly helpful to manage hospital resources and planning strategies.

1.2 Objectives

The goal of this research work is to study the short-term impact of air pollution on respiratory hospital admissions in Aveiro, Portugal between 2013 and 2016. In order to achieve this goal the following objectives are defined:

1. Perform a descriptive study of air pollutants;
2. Develop a comprehensive and flexible framework to model and estimate air pollutants data in Portugal mainland;
3. Produce forecasts of air pollutants;
4. Carry out a descriptive study of hospital admissions due to respiratory causes at Aveiro hospital;
5. Model hospital admissions at Aveiro hospital using air pollutants as covariates;
6. Explore forecast strategies for hospital admissions at Aveiro hospital.

As previously mentioned, forecasts can be quite useful for hospital resource management and planning at moments of increased air pollution, such as forest fires. In order to obtain hospital admissions forecasts, two strategies will be used: i) using observed air pollutants data, ii) using forecasts from air pollutants models. The latter strategy may be particularly useful when no observed air pollutants data is available.

It is also worth to note that an overall goal of this project is to improve the knowledge on time series analysis theory in order to further continue to develop research work in this scientific area.

1.3 Data

1.3.1 Air Pollutants

Air pollution data was downloaded from QualAr database (Base de Dados Online sobre a Qualidade do Ar) available at <https://qualar.apambiente.pt/qualar/>. The Portuguese Environmental Agency (Agência Portuguesa do Ambiente) is responsible for maintaining and validating the data [1].

Hourly measurements from all Portuguese mainland stations with information regarding NO_2 , NO_x , PM_{10} , $\text{PM}_{2.5}$, O_3 , SO_2 and CO from 2005 up to 2016 were retrieved. In total, there were 46 time series of NO_2 and NO_x , 35 time series of O_3 , 39 time series of PM_{10} , 20 time series of SO_2 and 12 time series of CO and $\text{PM}_{2.5}$ with data for the selected time period. Nevertheless, only time series with at least 85% of observations were included in the study, which resulted in including the time series shown in Table 1.3, which also shows the type of environment and the type of influence of each station. An urban background station is a place in an urban area where levels are representative of the exposure of the general urban population. Similarly, equivalent definitions can be used to describe the remaining environments and influence surroundings, since environment type (urban, suburban, rural) refers to the environment on a scale of several kilometres and station influence (traffic, industrial, background) refers to the impact (or absence) of near-by emissions [2]. Furthermore, it is displayed the number of included stations (less than 15% of missing data) of each pollutant.

ID	Station	Environment	Influence	NO ₂	NO _x	PM ₁₀	PM _{2.5}	O ₃	SO ₂	CO
1	Arcos	Urban	Background	✓	✓			✓		✓
2	Aveiro	Urban	Traffic	✓	✓	✓				✓
3	Av. Liberdade	Urban	Traffic	✓	✓	✓				✓
4	Beato	Urban	Background	✓	✓			✓		
5	Chamusca	Rural	Background	✓	✓	✓	✓	✓		
6	Custóias	Suburban	Background	✓	✓	✓		✓		
7	Entrecampos	Urban	Traffic	✓	✓	✓	✓	✓	✓	✓
8	Ermesinde	Urban	Background	✓	✓	✓		✓		
9	Ervedeira	Rural	Background	✓	✓	✓	✓	✓	✓	
10	Escavadeira	Urban	Industrial	✓	✓	✓		✓	✓	
11	Estarreja	Suburban	Background	✓	✓	✓	✓	✓	✓	
12	Fornelo Monte	Rural	Background	✓		✓		✓		
13	Fr. Bartolomeu	Urban	Traffic	✓	✓	✓				
14	Fr. Sá Carneiro	Urban	Traffic	✓	✓	✓				✓
15	Frossos	Suburban	Traffic	✓	✓	✓		✓		
16	Fundão	Rural	Background	✓	✓	✓	✓	✓	✓	
17	Ílhavo	Suburban	Background	✓	✓	✓		✓	✓	
18	Instituto Geofísico	Urban	Background	✓	✓	✓		✓		
19	Laranjeiro	Urban	Background	✓	✓	✓		✓		✓
20	Lavradio	Urban	Industrial						✓	
21	Loures	Urban	Background	✓	✓	✓		✓		
22	Meco	Suburban	Industrial			✓		✓		
23	Mem Martins	Urban	Background	✓	✓	✓		✓	✓	
24	Monte Chãos	Suburban	Industrial	✓				✓		
25	Monte Velho	Rural	Background			✓				
26	Olivais	Urban	Background	✓	✓	✓	✓	✓	✓	✓
27	Paços Ferreira	Urban	Background					✓		
28	Paio Pires	Suburban	Industrial	✓	✓			✓	✓	
29	Pe Moreira Neves	Urban	Traffic	✓	✓					
30	Quebedo	Urban	Traffic	✓	✓	✓			✓	✓
31	Quinta Marquês	Urban	Background		✓					
32	Restelo	Urban	Background	✓	✓			✓		
33	Sonega	Rural	Industrial	✓	✓			✓	✓	
34	Terena	Rural	Background				✓	✓		
35	Vermoim	Urban	Traffic	✓	✓	✓				
36	VNTelha	Urban	Traffic					✓		
Total		-	-	29	28	24	7	26	12	8

Table 1.3: Characteristics and number of included time series.

Over 20 stations are found for NO₂, NO_x, PM₁₀ and O₃, while SO₂ has only 12 time series and PM_{2.5} and CO only have, 7 and 8 time series, respectively. Carbon monoxide and PM_{2.5} have fewer series for different reasons. Carbon monoxide results essentially from combustion processes (e.g. motor vehicles combustion, forest fires), but it is mostly consumed in the process of ozone formation [62], therefore its levels are currently monitored only in metropolitan areas where there is a large traffic intensity. In contrast, PM_{2.5} has less time series because in 2005 there were only 12 monitoring stations measuring this pollutant. In the subsequent years, more stations started to monitor PM_{2.5} and, currently, there are 24

monitoring stations measuring this pollutant in Portugal mainland [1].

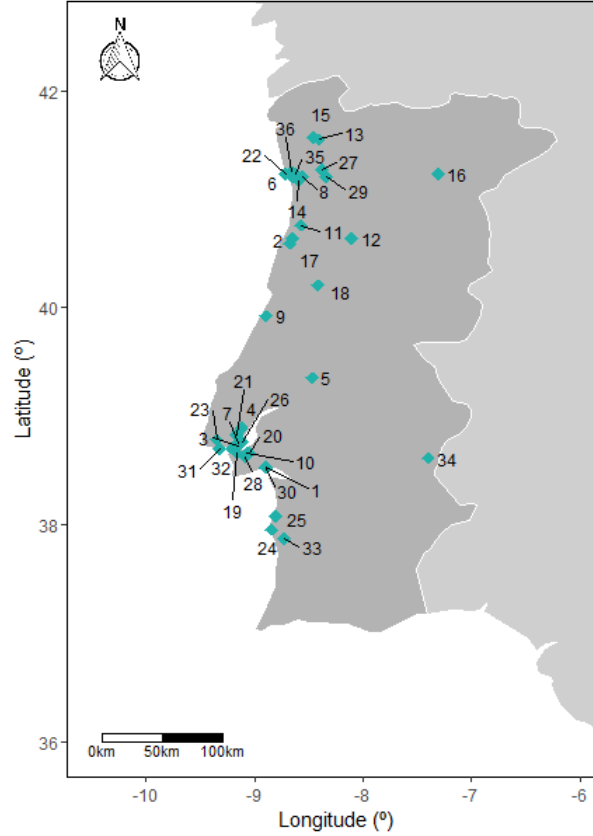


Figure 1.8: Geographic distribution of air pollutants stations.

Figure 1.8 shows the geographic distribution of the air pollutants stations in Portugal mainland. Most stations (50%) are located in the metropolitan areas of Lisbon and Porto, which is no surprise since these are the areas with more traffic and also industrial areas, that is with higher number of polluting sources. On the contrary, very few of the included stations are located in the hinterland. It is noteworthy that no station in the south of Portugal was included.

As mentioned, only time series with at most 15% of missing observations were included, because in time series analysis it is essential to have complete data so that the time series dependence is fully accounted for. In order to have complete data, missing imputation is needed. Hence, to ensure that the missing imputation does not lead to any changes in the dependency structure of the time series, a maximum of 15% of missing data has been used in the literature [25]. Missing information in these time series are most likely due to devices malfunction. Therefore, the k-nearest neighbours (kNN) method, particularly, the 1-NN method ($k = 1$) is especially useful to impute the missing data. The value $k = 1$ was

selected because it allows maintaining the data structure (i.e., mean and standard deviation) [8]. Figure 1.9 shows the ratio between the mean and the standard deviation of the original time series and the imputed time series according to the number of k -neighbours for SO_2 and PM_{10} . Selecting $k = 1$ allows maintaining the data structure and a small computational effort. Figure A.1 shows these results for the remaining pollutants.

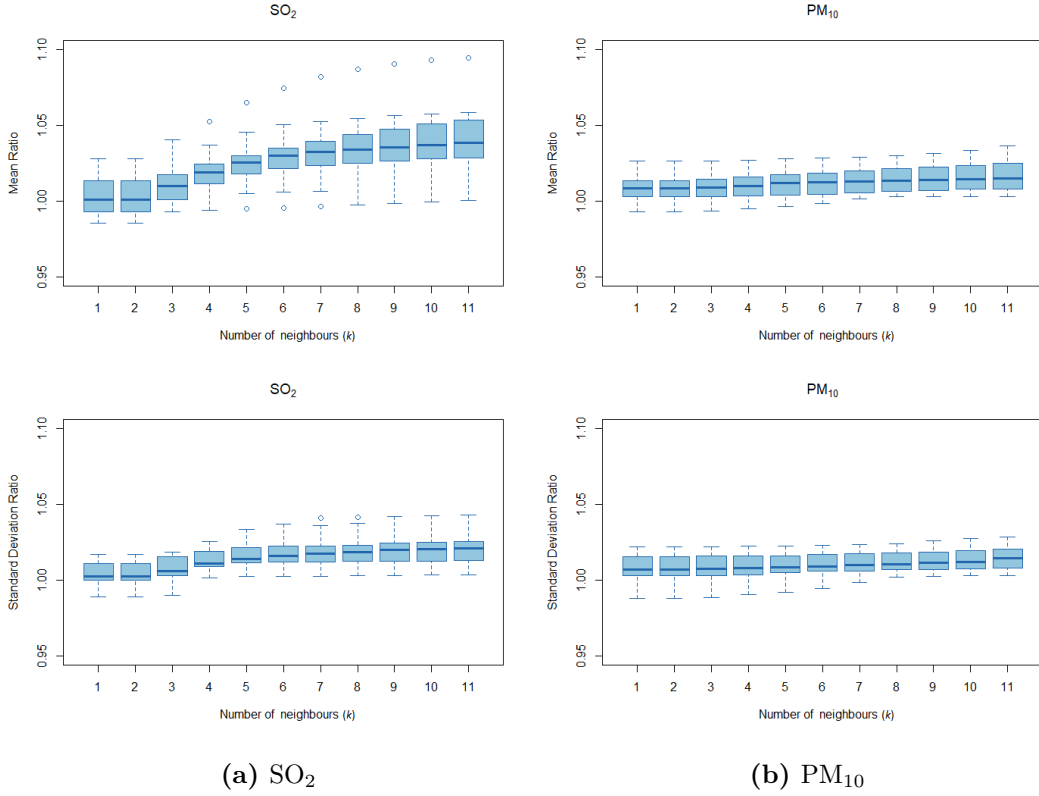


Figure 1.9: Boxplot of the ratio between the mean of the imputed series and the mean of the series with missing data (upper panel). Same representation for the standard deviation (lower panel). (a) SO_2 , (b) PM_{10} .

In the 1-NN method, the missing data is replaced by the value of the nearest neighbour, that is, the time series with the most similar values using the Heterogeneous Euclidean-Overlap Metric (HEOM) [64]. Figure 1.10 shows the imputed PM_{10} Estarreja time series and shows that the imputed values keep the seasonal structure of the series.

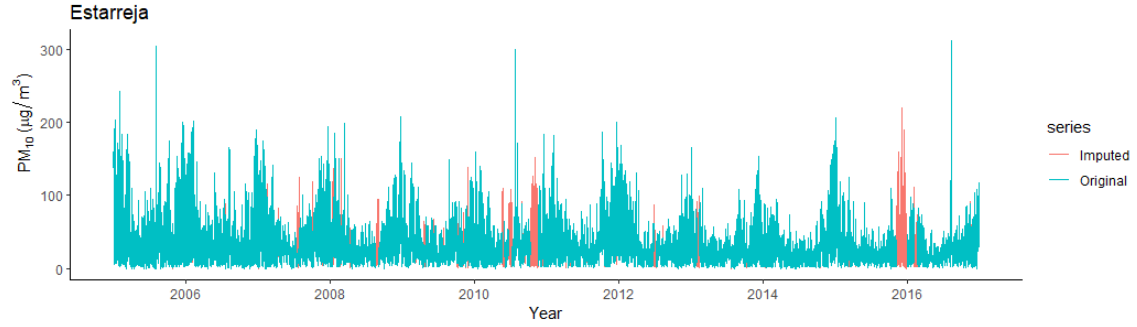


Figure 1.10: Example of an imputed time series with the 1-NN method.

For each station with simultaneous PM_{10} and $PM_{2.5}$ measurements, it is worth to mention that, time series of $PM_{2.5}$ suffered a specific preprocessing. Since PM_{10} series include measurements of all particles with less than $10\mu m$, including those of size inferior to $2.5\mu m$, values of $PM_{2.5}$ higher than PM_{10} were removed, which added at most 6% of additional missing values in each time series of $PM_{2.5}$. This procedure was implemented under the advice of air quality experts as these errors are common to occur for this pollutant. Imputation of missing values was then performed as previously described.

1.3.2 Hospital Admissions

Data on emergency hospital admissions due to respiratory causes were provided by Aveiro hospital for the period between 2013 and 2016. Specifically, admissions with the following ICD-9 (International Classification of Diseases 9th Revision) codes [59] were considered: 460-462, 464-465, 485, 490-493. Table 1.4 shows the detailed information for each code. These codes were selected under medical professionals advice and suggestion since these outcomes are the most likely to be influenced by air pollution levels.

ICD-9 code	Description
460	Acute nasopharyngitis (common cold)
461	Acute sinusitis
462	Pharyngitis, acute
464	Acute laryngitis and tracheitis
465	Acute upper respiratory infection of multiple or unspecified sites
466.1	Acute bronchiolitis
485	Bronchopneumonia, organism unspecified
490	Bronchitis, not specified as acute or chronic
491	Chronic bronchitis
491.21	Obstructive chronic bronchitis with (acute) exacerbation
491.22	Obstructive chronic bronchitis with acute bronchitis
493	Asthma
493.92	Asthma, unspecified type, with (acute) exacerbation

Table 1.4: Description of the International Classification of Diseases 9th Revision codes [59].

Figure 1.11 shows the location of Aveiro hospital and the surrounding air pollution stations. The closest station to Aveiro hospital, as expected, is Aveiro station followed by Ílhavo and Estarreja. Nevertheless, each station may be suitable to describe the association between hospital admissions and air pollution, as populations from all these cities are likely to attend Aveiro hospital.

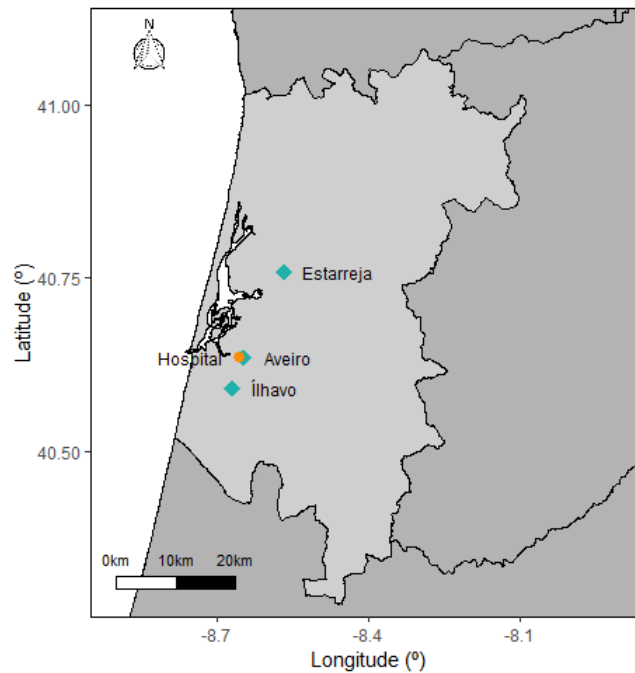


Figure 1.11: Location of Aveiro hospital and the nearest air pollution stations.

Chapter 2

Methods

In this chapter, a detailed description of the methodology used is presented for both continuous time series and discrete time series. Theoretical aspects are discussed firstly and practical implementation questions in R software are discussed afterwards. For continuous time series, SARIMA, DHR, ARFIMA and SARFIMA models are introduced in order to develop a framework to model and forecast air pollutants data. Regarding discrete time series, generalised linear models are described.

2.1 Continuous Time Series

In this work the focus is on ARIMA-based models and its extensions due to its simplicity and capacity to deal with a multitude of scenarios and also due to the relevance of these models in time series theory [45].

2.1.1 Theoretical Aspects

A time series is defined as a collection of random variables indexed according to the order in which they are obtained over time. For instance, a time series can be a sequence of random variables x_1, x_2, x_3 , and so on, where the random variable x_1 is the value of the time series at $t = 1$, and the variable x_2 denotes the value taken by the time series for the second time period ($t = 2$), and so on [50]. Generally, a collection of random variables $\{x_t\}$ indexed by t is referred to as a stochastic process, while the observed values of the stochastic process are referred to realisations of that process. The term time series will be used to refer both to the process and to a particular realisation, thus, no distinction between the two concepts will be made. It is common for time series to be represented as X_t , such as in [10], nevertheless the notation x_t of [50] is adopted in this work.

The simplest example of a time series is a *white noise*, which is a collection of uncorrelated random variables (w_t) with zero mean and finite variance (σ_w^2). The white noise process is denoted as $w_t \sim \text{wn}(0, \sigma_w^2)$. Sometimes it is required that the white noise is independent and identically distributed (iid), thus this process is distinguished by denoting it as $w_t \sim \text{iid}(0, \sigma_w^2)$ and is referred to as a Gaussian white noise process (Figure 2.1).

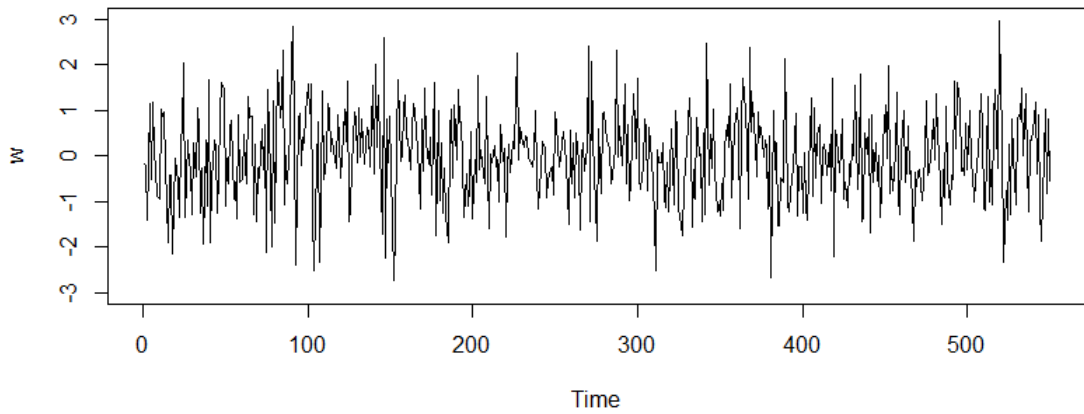


Figure 2.1: Gaussian white noise time series.

To adequately study time series, there are some particular tools to analyse time series and some definitions that need to be introduced first. Considering the stochastic process $\{x_t : t = 0, \pm 1, \pm 2, \dots\}$, as mentioned, for each t , x_t is a random variable, and it has a distribution function [45], given by

$$F_t(x) = P\{x_t \leq x\}, \quad -\infty < x < +\infty, \quad (2.1)$$

and, the density function [45] is

$$f_t(x) = \frac{\partial F_t(x)}{\partial x}. \quad (2.2)$$

These marginal functions, when they exist are quite informative in examining the behaviour of the time series [50]. The definitions used in this section are based on [50, 45, 10], for further details please refer to it.

The mean function is another informative marginal descriptive measure.

Definition 2.1.1. The mean function, if exists, is defined as

$$\mu_t = E(x_t) = \int_{-\infty}^{+\infty} x f_t(x) dx, \quad (2.3)$$

where E denotes the usual expected value operator.

The autocovariance measures the linear dependence between two points of the time series, observed at different instants. Very smooth time series have autocovariance functions that stay large even when t and s are far apart, while rough series exhibit autocovariance functions that are nearly zero for observations very further apart.

Definition 2.1.2. The autocovariance function is defined as the second moment product

$$\gamma_x(s, t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \quad \forall s, t. \quad (2.4)$$

Note that $\gamma_x(s, t) = \gamma_x(t, s)$ for all time points s and t . It is clear that for $s = t$, the autocovariance reduces to the (assumed finite) variance, since

$$\gamma_x(t, t) = \text{cov}(x_t, x_t) = E[(x_t - \mu_t)^2] = \text{var}(x_t). \quad (2.5)$$

Definition 2.1.3. The autocorrelation function (ACF) is defined as

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}. \quad (2.6)$$

The ACF is one of the most useful and important tools in time series analysis. This function measures to what extent an observation at time t depends on s . In other words, it

measures the linear predictability of the series at time t , say x_t , using only the value x_s . The autocorrelation function ranges between 1 and -1 ($-1 \leq \rho(s, t) \leq 1$).

When it is of interest to study the predictability of one series (x_t) from another time series (y_t), and assuming that both have finite variances, the following definitions can be derived.

Definition 2.1.4. The cross-covariance function between two time series x_t and y_t is given by

$$\gamma_{xy}(s, t) = \text{cov}(x_s, y_t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})]. \quad (2.7)$$

Definition 2.1.5. The cross-correlation function (CCF) is defined as

$$\rho_{xy}(s, t) = \frac{\gamma_{xy}(s, t)}{\sqrt{\gamma_x(s, s)\gamma_y(t, t)}}. \quad (2.8)$$

Another function of interest is the partial autocorrelation function (PACF), this function shows the autocorrelation between x_t and x_{t-h} , after removing any linear dependence on $x_1, x_2, \dots, x_{t-h+1}$. Therefore, it shows to what extent two observations of the time series are correlated if the observations in between are removed. The PACF correlation of lag h denoted by ϕ between x_t and x_{t+k} is given by

$$\phi_{11} = \text{corr}(x_{t+1}, x_t) \quad (2.9)$$

$$\phi_{hh} = \text{corr}[x_{t+k} - P_{t,k}(x_{t+k}), x_t - P_{t,k}(x_t)], \quad k \geq 2, \quad (2.10)$$

where $P_{t,k}$ are the collection of observations from $x_{t+1}, \dots, x_{t+k-1}$.

The above definitions are completely general, however, time series analysis requires special assumptions, namely, stationarity [45]. Stationarity of a time series can be defined in a strict or weak sense.

Definition 2.1.6. A strictly stationary time series is one for which the probabilistic behaviour of every collection of values $\{x_{t1}, x_{t2}, \dots, x_{tk}\}$, is identical to that of the time shifted set $\{x_{t1+h}, x_{t2+h}, \dots, x_{tk+h}\}$. That is,

$$P\{x_{t1} \leq c_1, \dots, x_{tk} \leq c_k\} = P\{x_{t1+h} \leq c_1, \dots, x_{tk+h} \leq c_k\}, \quad (2.11)$$

for all $k = 1, 2, \dots$, all time points t_1, t_2, \dots, t_k , all numbers c_1, c_2, \dots, c_k , and all time shifts $h = 0, \pm 1, \pm 2, \dots$.

An important case where stationarity implies strict stationarity is a Gaussian time series. However, Definition 2.1.6 is too strong for most real applications. Furthermore, stationarity in the strong sense is very difficult to assess from a single data set. Nevertheless, a less restrictive definition of stationarity can be employed to be adequate to real data, this is the concept of weak stationarity [50].

Definition 2.1.7. A weakly stationary time series (x_t) is a finite variance process such that,

1. the mean value function, defined in (Equation 2.3) is constant and does not depend on time t , and,
2. the autocovariance function, defined in (Equation 2.4) depends on s and t only through their difference $|s - t|$.

Onwards, the term stationarity is employed to describe a weakly stationary time series.

Since the time series are stationary, and the mean and the autocovariance of the process do not depend on time, the definitions of mean, autocovariance and the autocorrelation function can be presented as below [50]. The mean function $E(x_t) = \mu_t$, is now simply written as $\mu_t = \mu$. Also, the notation of the autocovariance can be simplified. Let $s = t + h$, where h represents the time shift or lag. Then, $\gamma(t + h, t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_h, x_0) = \gamma(h, 0)$, because the time difference between $t + h - h$ is the same as the time difference between $h - 0$. Hence, the autocovariance function of a stationary time series does not depend on time. For convenience, we drop the second argument of $\gamma(h, 0)$ and the autocovariance function is written as $\gamma(h)$.

Definition 2.1.8. The autocovariance function of a stationary time series can be written as

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu)(x_t - \mu)]. \quad (2.12)$$

Definition 2.1.9. The ACF of a stationary time series using (Equation 2.6) is given by

$$\rho(h) = \frac{\gamma(t + h, t)}{\sqrt{\gamma(t + h, t + h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}. \quad (2.13)$$

A white noise process has mean $\mu = 0$ and autocovariance

$$\gamma_h(s, t) = \text{cov}(w_{t+h}, w_t) = \begin{cases} \sigma_w^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases} \quad (2.14)$$

Therefore, white noise satisfies the conditions of (Definition 2.1.7) and is weakly stationary. If the white noise were to be Gaussian, the time series would also be strictly stationary, by evaluating (Definition 2.1.6) using the fact that the noise would also be independent and identically distributed.

Definition 2.1.10. The PACF of a stationary process, x_t , denoted ϕ_{hh} , for $h = 1, 2, \dots$, is

$$\phi_{11} = \text{corr}(x_{t+1}, x_t) = \rho(1), \quad (2.15)$$

and,

$$\phi_{hh} = \text{corr}(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t), \quad h \geq 2. \quad (2.16)$$

The notion of stationarity is also important when studying the relationship between two time series, thus the cross-covariance function and the cross-correlation function are written as follows:

Definition 2.1.11. Two time series, x_t and y_t , are said to be jointly stationary if they are each stationary, and the cross-covariance function is a function only of lag h ,

$$\gamma_{xy}(h) = \text{cov}(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)]. \quad (2.17)$$

Definition 2.1.12. The CCF of jointly stationary time series, x_t and y_t , is defined as,

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}. \quad (2.18)$$

It should be noted that the CCF, generally, is not symmetric about zero, that its, $\rho_{xy}(h) \neq \rho_{xy}(-h)$. However, $\rho_{xy}(h) = \rho_{yx}(-h)$.

The theoretical ACF and CFF functions are useful for describing the properties of hypothesised models, but most of the analyses are performed using sampled data. This limitation implies that usually, there is only one realisation of the process. Thus, it is necessary to use averages over this single realisation to estimate the population mean and covariance functions, for which the assumption of stationarity is mandatory [50]. Therefore the sample mean (random variable) is

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t, \quad (2.19)$$

and the standard error of the estimate is the square root of $\text{var}(\bar{x})$, which is given by

$$\begin{aligned} \text{var}(\bar{x}) &= \text{var}\left(\frac{1}{n} \sum_{t=1}^n x_t\right) = \frac{1}{n^2} \text{cov}\left(\sum_{t=1}^n x_t, \sum_{s=1}^n x_s\right) \\ &= \frac{1}{n^2} \left(n\gamma_x(0) + (n-1)\gamma_x(1) + (n-2)\gamma_x(2) + \dots + \gamma_x(n-1) \right. \\ &\quad \left. + (n-1)\gamma_x(-1) + (n-2)\gamma_x(-2) + \dots + \gamma_x(1-n) \right) \\ &= \frac{1}{n} \sum_{h=-n}^n \left(1 - \frac{|h|}{n}\right) \gamma_x(h) \end{aligned} \quad (2.20)$$

Definition 2.1.13. The estimator of the sample autocovariance function is given by

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}), \quad (2.21)$$

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \dots, n-1$.

Definition 2.1.14. The sample ACF is

$$\hat{\rho} = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \quad (2.22)$$

There is a property of the sample autocorrelation that is important to mention. Under some conditions (see [50]) if a time series x_t is white noise, then for a large n the sample ACF, $\hat{\rho}_x(h)$, for $h = 1, 2, \dots, H$, where H is fixed but arbitrary, is approximately normally distributed with zero mean and standard deviation given by,

$$\sigma_{\hat{\rho}_x}(h) = \frac{1}{\sqrt{n}}. \quad (2.23)$$

This is a rough method to assess whether peaks in the autocorrelation are statistically significant, by determining whether or not the observed peak is outside the interval $\frac{\pm 2}{\sqrt{n}}$. For a white noise time series, about 95% of the observations lie within these limits.

Definition 2.1.15. Lastly, the sample cross-covariance function and the sample CCF are given, respectively, by

$$\hat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \quad (2.24)$$

where $\hat{\gamma}_{xy}(-h) = \hat{\gamma}_{yx}(h)$ determines the function for negative lags. And,

$$\hat{\rho}_{xy} = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}. \quad (2.25)$$

For which, the same property of the ACF applies if at least one of the processes is an independent white noise [50].

SARIMA

An autoregressive process of order p [AR(p)], satisfies the following equation

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \dots - \phi_p x_{t-p} = \varepsilon_t, \quad (2.26)$$

where $\varepsilon_t \sim \text{wn}(0, \sigma_w^2)$, $\sigma_w^2 > 0$ and $\phi_1, \phi_2, \dots, \phi_p$, ($\phi_p \neq 0$) are the model parameters, the time series is stationary and has zero mean. Equation 2.26 can be written more concisely as

$$\phi_p(B)x_t = \varepsilon_t, \quad (2.27)$$

where B is the backshift operator defined as $Bx_t = x_{t-1}$ and the autoregressive operator is defined as $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$. This operator can be extended to powers, for instance $B^2 x_t = B(Bx_t) = x_{t-2}$. The autoregressive process is stationary when the roots of the polynomial $\phi_p(z) \neq 0$ lie outside the unit circle, that is $|z| \leq 1$ [45].

In (Equation 2.26) the model has zero mean, however, if the mean of the process considered is not zero, one just needs to replace x_t by $x_t - \mu$ and write the equation as

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} - \dots + \phi_p x_{t-p} + \varepsilon_t, \quad (2.28)$$

where $c = \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p)$.

On the contrary, moving average processes [MA(q)], are of the following form

$$x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}, \quad (2.29)$$

where $\varepsilon_t \sim \text{wn}(0, \sigma_w^2)$, $\sigma_w^2 > 0$ and $\theta_1, \theta_2, \dots, \theta_q$, ($\theta_q \neq 0$) are the parameters. Similarly to the AR model, the MA model can be expressed as

$$x_t = \theta_q(B) \varepsilon_t, \quad (2.30)$$

where the moving average operator is $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q$.

The autoregressive and the moving average models can be combined to produce autoregressive moving average models - ARMA(p, q). These models are stationary and are represented by the following equation

$$x_t - \phi_1 x_{t-1} - \dots - \phi_p x_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad (2.31)$$

where $\phi_p \neq 0$, $\theta_q \neq 0$ and $\varepsilon_t \sim \text{wn}(0, \sigma_w^2)$. If x_t has non-zero mean, the model can be written in an identical form to (Equation 2.28). Note that an ARMA(0, q) is simply a moving average model of order q , whereas if the model is an ARMA($p, 0$) is just an AR(p) model. The above model can be written more concisely as

$$\phi_p(B) x_t = \theta_q(B) \varepsilon_t. \quad (2.32)$$

For further details on ARMA conditions and proofs [50] is an excellent resource.

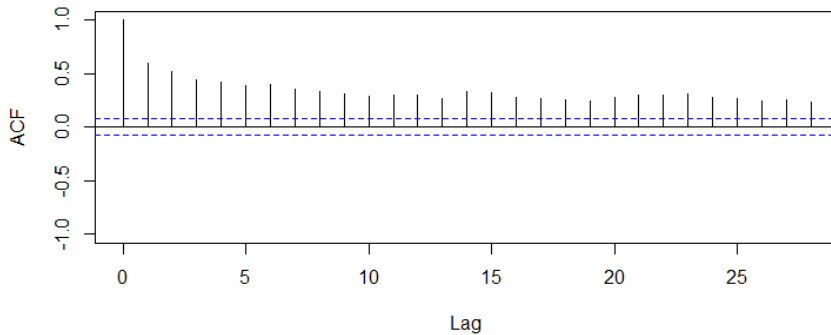
A generalisation of ARMA models can be performed to obtain seasonal models, the seasonal autoregressive and moving average models, denoted as SARMA(p, q)(P, Q) $_S$, where P and Q are the autoregressive and moving average orders of the seasonal component, respectively. These models are defined as

$$\Phi_P(B^S) \phi_p(B) x_t = \Theta_Q(B^S) \theta_q(B) \varepsilon_t, \quad (2.33)$$

where B^S is the backshift operator of the seasonal component, Φ_P is the AR seasonal parameter $\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \dots - \Phi_P B^{PS}$, Θ_Q is the MA seasonal parameter $\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \dots + \Theta_Q B^{QS}$ and S is the value of the seasonal component. For instance, if $S = 24$ it means that, a time series with a hourly resolution, has a

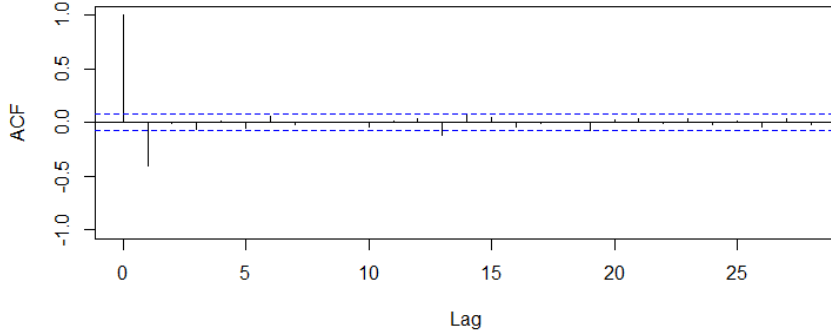
daily periodicity. Similarly, to ARMA models, SARMA models can be purely $\text{SAR}(P, 0)_S$ or $\text{SMA}(0, Q)_S$.

The time series models presented so far, require that the time series is stationary, but truth is that in real applications it is highly unlikely to have stationarity. Hence, transformations to the time series are usually required to achieve stationarity. Time series can be non-stationary in mean, variance or both [45]. Non-stationarity in mean and variance are dealt with differently. If a series is non-stationary in mean, the most common way to try to achieve stationarity is to perform differencing of the series, for which the lag difference (∇x_t) operator is particularly useful. The lag difference operator is defined as $\nabla x_t = (1 - B)x_t = x_t - x_{t-1}$. Similarly to the backshift operator B , this operator can also be extended to powers, for example, $\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}$. As an example, Figure 2.2 shows the autocorrelation function for *varve* time series of the *astsa* R package. As it can be seen in the ACF of the original time series [Figure 2.2(a)], all lags present significant correlations, thus indicating that the series is non-stationary. On Figure 2.2(b), a first differencing was applied, that is $\nabla^1 x_t = x_t - x_{t-1}$. After applying the first order differencing only a significant peak at lag 1 is found, which suggests that an MA(1) model may be adequate to describe the data, after differencing.



(a) Original series.

Figure 2.2: ACF of *varve* time series from the *astsa* R package. (a) original time series, (b) differenced time series. The blue lines represent the limits $\frac{\pm 2}{\sqrt{n}}$, in which lie 95% of observations if the time series resembles a white noise. (To be continued)



(b) Differenced series.

Figure 2.2: ACF of *varve* time series from the *astsa* R package. (a) original time series, (b) differenced time series. The blue lines represent the limits $\frac{\pm 2}{\sqrt{n}}$, in which lie 95% of observations if the time series resembles a white noise. (Continued)

ARMA models that allow differencing are denoted as $\text{ARIMA}(p, d, q)$ where the I stands for integrated and is represented by $d \in \mathbb{N}_0$, which is the differencing parameter. If $d = 0$, then the ARIMA is simply an $\text{ARMA}(p, q)$ model. The model is written as

$$\phi_p(B)\nabla^d x_t = \theta_q(B)\varepsilon_t, \quad (2.34)$$

In practice, a differencing parameter $d > 2$ is unlikely to be used due to the increasing model complexity with increasing d and the estimation effort [30].

Generalisation to seasonal models is straightforward and give origin to $\text{SARIMA}(p, d, q)(P, D, Q)_S$ defined as

$$\Phi_P(B^S)\phi_p(B)\nabla_S^D\nabla^d x_t = \Theta_Q(B^S)\theta_q(B)\varepsilon_t, \quad (2.35)$$

where D is the seasonal differencing parameter, hence $\nabla_S^D = (1 - B^S)^D$ and $\varepsilon_t \sim \text{wn}(0, \sigma_w^2)$. For instance, in a hourly time series with daily seasonality, that is $S = 24$ and $D = 1$, $\nabla_S^D x_t = \nabla_{24}^1 x_t = x_t - x_{t-24}$, making a first order differencing on the seasonal component, implies to remove to each observation the value of the observation of the previous day. When dealing with seasonal data it is advisable to first apply the seasonal differencing and only afterwards perform the non-seasonal differencing, if necessary [30]. $\text{SARIMA}(p, d, q)(P, D, Q)_S$ processes are stationary if $d + D < 0.5$, $D < 0.5$ and $\phi_p(z)\Phi_P(z^S) \neq 0$, for $|z| \leq 1$ [9].

To deal with non-stationarity in variance of the times series, the Box-Cox transformation can be used to stabilise the variance. However, this implies that the modelled time series will be on a different scale from the original data, thus, there might be a loss of interpretability of the model due to the change of physical units in x_t . The Box-Cox transformation is defined

as

$$y_t = \begin{cases} \frac{(x_t^\lambda - 1)}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x_t) & \text{if } \lambda = 0. \end{cases} \quad (2.36)$$

In 1970, Box and Jenkins proposed a methodology to model ARIMA, and by extension SARIMA models. Figure 2.3 shows the proposed methodology by these statisticians in order to achieve an *optimal model*, that is a model which includes the smallest number of estimated parameters needed to adequately fit the patterns of the data [45]. This methodology consists essentially in 3 stages: (i) Identification, (ii) Estimation and (iii) Diagnostic Assessment.

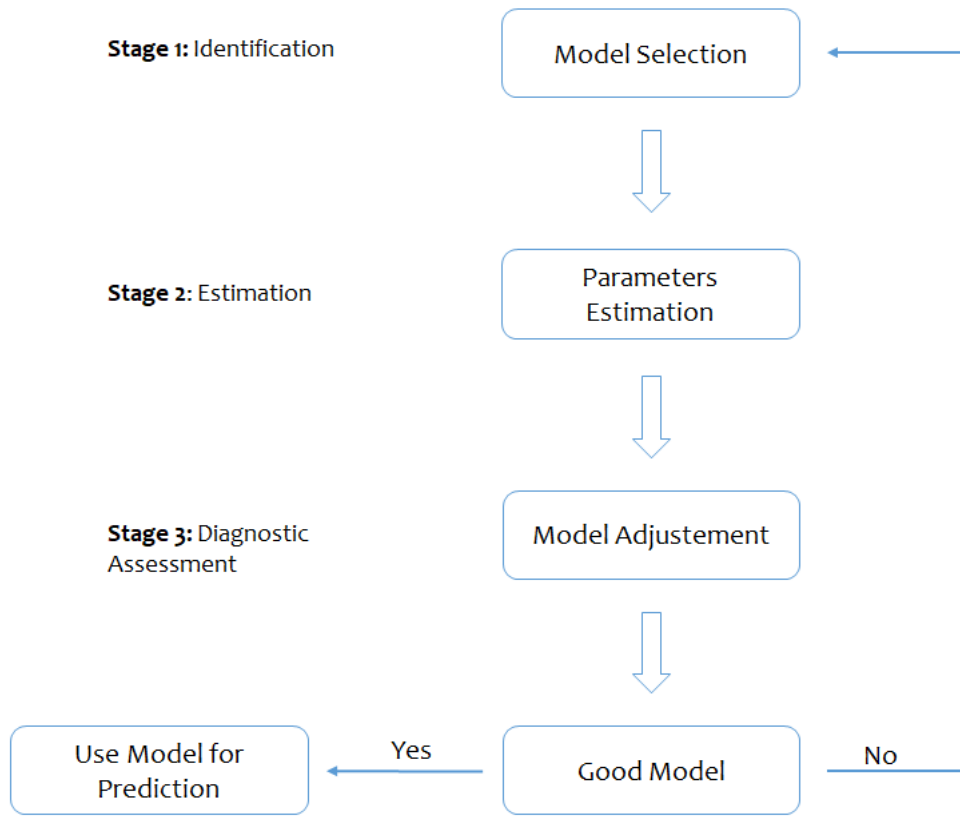


Figure 2.3: Box-Jenkins methodology for model selection. Adapted from [45].

On the first stage, it is proposed to identify an ARIMA or SARIMA model. To do so, it is necessary to study the properties of the time series, that is its stationarity. Hence, this stage is a two-step process in which, first the time series must be stationarised and afterwards, based on their ACF and PACF a model is selected. Tables 2.1 and 2.2 describe the behaviour of the ACF and PACF, for ARMA and SARMA processes, to facilitate model selection.

Process	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off at lags
PACF	Cuts off after lag p	Tails off at	Tails off

Table 2.1: ACF and PACF behaviour for purely ARMA models. Reproduced from [50].

Process	AR(P) $_S$	MA(Q) $_S$	ARMA(P, Q) $_S$
ACF*	Tails off at lags kS , $k = 1, 2, \dots$	Cuts off after lags QS	Tails off at lags kS
PACF*	Cuts off after lags PS	Tails off at lags kS , $k = 1, 2, \dots$	Tails off at lags kS

*The values at non-seasonal lags $h \neq kS$, for $k = 1, 2, \dots$, are zero.

Table 2.2: ACF and PACF behaviour for purely SARMA models. Reproduced from [50].

Once the model is selected it can now be estimated. Nevertheless, the estimation of ARIMA or SARIMA models is not simple. In fact, it has been shown that only for the simpler models is possible to define explicit formulas to obtain the parameter estimators. Generally, the estimators are the solution of a complex system of non-linear equations, which require numerical analysis techniques and computational calculations to be solved [45]. For an ARMA(p, q) model, let $\beta = (\mu, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)^T$, be the $(p + q + 1)$ -dimensional vector of the model parameters. In this case, the maximum likelihood function is

$$L(\beta, \sigma_w^2) = \prod_{t=1}^n f(x_t | x_{t-1}, \dots, x_1) \quad (2.37)$$

The conditional distribution of x_t given x_{t-1}, \dots, x_1 is Gaussian with mean x_t^{t-1} and variance P_t^{t-1} , where $P_t^{t-1} = \gamma(0) \prod_{j=1}^{t-1} (1 - \phi_{jj}^2)$ [50]. For ARMA models, $\gamma(0) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2$ and it can be written as

$$P_t^{t-1} = \sigma_w^2 \left\{ \left[\sum_{j=0}^{\infty} \psi_j^2 \right] \left[\prod_{j=1}^{t-1} (1 - \phi_{jj}^2) \right] \right\} \stackrel{(def)}{=} \sigma_w^2 r_t, \quad (2.38)$$

where r_t is the term in brackets and ψ_j^2 are the *psi*-weights, that is the MA representation of the ARMA model. Note that the r_t terms are functions only of the regression parameters and that they may be computed recursively as $r_{t+1} = (1 - \phi_{tt}^2)r_t$, with initial condition $r_1 = \sum_{j=0}^{\infty} \psi_j^2$. The likelihood of the data can now be written as

$$L(\beta, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} [r_1(\beta)r_2(\beta)\dots r_n(\beta)]^{-1/2} \exp \left[-\frac{S(\beta)}{2\sigma_w^2} \right], \quad (2.39)$$

where

$$S(\beta) = \sum_{t=1}^n \left[\frac{(x_t - x_t^{t-1}(\beta))^2}{r_t(\beta)} \right]. \quad (2.40)$$

Note that, x_t^{t-1} and r_t are function of β alone. Maximum likelihood estimation can now proceed by maximising (Equation 2.39) with respect to β and σ_w^2 , that is,

$$\hat{\sigma}_w^2 = n^{-1} S(\hat{\beta}), \quad (2.41)$$

where $\hat{\beta}$ is the value of β that minimises the concentrated likelihood

$$l(\beta) = \log[n^{-1} S(\beta)] + n^{-1} \sum_{t=1}^n \log r_t(\beta). \quad (2.42)$$

After estimating the model parameters it is necessary to assess the model goodness of fit. There are several points that can be evaluated. For instance, one can evaluate whether or not all model parameters are significant and, if not, probably these are not necessary to describe the data, and a model with fewer parameters will be more suitable to the data. Also, it is necessary to evaluate the errors of the model. The error must have zero mean, be uncorrelated and normally distributed. These characteristics can be assessed by visual inspection of the ACF of the residuals and their histogram. Nevertheless, there are also statistical tests, such as the Kendal and Stuart test or the Portmanteau test, which evaluate the null hypothesis of the residuals resembling a white noise distribution.

Dynamic Harmonic Regression

Dynamic regression models (DHR) are models of the form,

$$y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_k x_{k,t} + \varepsilon_t, \quad (2.43)$$

where y_t is a linear combination of the predictor variables $(x_{1,t}, \dots, x_{k,t})$ and ε_t , instead of being an uncorrelated error term is allowed to contain autocorrelation, in particular $\varepsilon_t \sim \text{ARIMA}(p, d, q)$. Let $\varepsilon_t \sim \text{ARIMA}(p, d, q) = \eta_t$, so this is not confused with ε_t which is a white noise error [30].

It is a well-known fact that any signal can be decomposed into a combination of sinusoidal waves [50], in other words, a time series can be expressed in terms of

$$x_t = \sum_{j=0}^k \left[\alpha_k \cos(\lambda_j t) + \beta_k \sin(\lambda_j t) \right], \quad (2.44)$$

where x_t is a Fourier series, $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k$ are the unknown parameters and $\lambda_1, \dots, \lambda_k$ are fixed frequencies, where $\lambda = \frac{2\pi k}{S}$, with S being the seasonal period of x_t . Therefore, a dynamic harmonic regression can be defined as

$$y_t = \beta_0 + x_t + \eta_t, \quad (2.45)$$

where x_t is of the form (Equation 2.44).

Dynamic harmonic regression models were introduced by Young and colleagues in 1980 [67] and are particularly well-suited to describe data with long seasonal periods or with multiple seasonality. SARIMA models are only able to deal with single seasonality and current computer implementation cannot deal with seasonal periods larger than 350 ($S = 350$), which can be quite an inconvenient if one has high-sampled data, such as hourly data or higher [30]. To describe a time series with multiple seasonality it is necessary to use as many Fourier terms (k) as the seasonality considered at different frequencies. The order of the Fourier terms can be at most half of the seasonal period, as suggested by $\lambda = \frac{2\pi k}{S}$. While the long memory of the model is dealt with the Fourier series, the short memory can be handled by the ARIMA error. The disadvantage of this model, compared to a SARIMA model is that the seasonality is assumed to be fixed in DHR models. Nevertheless, in practice, seasonality is usually remarkably constant so this is not a big disadvantage, except maybe for long time series. Another disadvantage may be the computation time if one has a model with high order Fourier terms. Additionally, it is necessary to select the order of the Fourier terms, which can be accomplished by comparing different models' performance after model estimation [30].

ARFIMA

Alternative models to describe the long memory component of time series are fractionally integrated autoregressive and moving average processes (ARFIMA or FARIMA), which are stationary processes with a much slower decreasing autocorrelation function. These processes are also called long memory processes or long-range dependent and were introduced by Hosking (1981) [28] and Granger and Joyeux (1980) [26] as an intermediate compromise between the short memory ARMA models and the fully integrated non-stationary processes in the Box-Jenkins class [50]. An ARFIMA(p, d, q) is characterised by allowing $d \in \mathbb{R}$. Hence this is the difference between ARIMA and ARFIMA models, while the first only allows $d \in \mathbb{N}$, the latter allows for fractional values of d . An ARFIMA is stationary if $d < 0.5$ and all the roots of the equation $\phi(z) = 0$ lie outside the unit circle [28]. Nevertheless, the ARFIMA process only describes the long-memory component if $0 < d < 0.5$ [9]. An ARFIMA(p, d, q) satisfy the following equation [45].

$$\phi_p(B)\nabla^d x_t = \theta_q(B)\varepsilon_t, \quad (2.46)$$

where $\phi(B)$ and $\theta(B)$ are the AR and MA operators, respectively, and $\varepsilon_t \sim \text{wn}(0, \sigma_w^2)$. The operator $(1 - B)^d$ is defined by the binomial expansion

$$\nabla^d = (1 - B)^d = \sum_{j=0}^{\infty} \pi_j B^j, \quad (2.47)$$

where $\pi_0 = 1$ and

$$\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} = \prod_{0 < k \leq j} \frac{k-1-d}{k}, \quad j = 0, 1, 2, \dots, \quad (2.48)$$

where $\Gamma(u)$ is the well-known Gamma function,

$$\Gamma(u) \begin{cases} \int_0^\infty e^{-z} z^{u-1} dz & u > 0, \\ \infty & u = 0, \\ u^{-1} \Gamma(1+u) & u < 0. \end{cases} \quad (2.49)$$

The autocorrelation function of an ARFIMA(p, q, d) process with $0 < |d| < 0.5$, has the following property

$$\rho(h)h^{1-2d} \rightarrow c \quad \text{as } h \rightarrow \infty, \quad (2.50)$$

which implies that the autocorrelation function converges to zero at a much slower rate than the ACF of an ARMA process. Thus, the latter are known as short-memory process in contrast to the ARFIMA processes.

Instead of being written like in (Equation 2.46), an ARFIMA(p, d, q) process can also be regarded as an ARMA(p, q) process driven a by fractionally integrated noise, that is,

$$\phi_p(B)x_t = \theta_q(B)w_t, \quad (2.51)$$

where

$$(1-B)^d w_t = \varepsilon_t. \quad (2.52)$$

The w_t process is called fractionally integrated white noise and can be shown to have variance

$$\gamma_w(0) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma^2(1-d)}, \quad (2.53)$$

Following the previous definitions it is possible to define the autocorrelation and autocovariance functions in terms of the model parameters' (p, d, q). The ACF is given by

$$\rho_w(h) = \frac{\Gamma(h+d)\Gamma(1-d)}{\Gamma(h-d+1)\Gamma(d)} = \prod_{0 < k \leq h} \frac{k-1-d}{k-d}, \quad h = 1, 2, \dots, \quad (2.54)$$

where Γ is the gamma function (see Equation 2.49). In this notation, the autocovariance of the ARFIMA(p, d, q) process x_t is

$$\gamma_x(h) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \psi_j \psi_k \gamma_w(h+j-k), \quad (2.55)$$

where $\sum_{i=0}^{\infty} \psi_i z^i = \frac{\theta(z)}{\phi(z)}$, with $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$, $\phi_p \neq 0$ and $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$, $\phi_q \neq 0$ are the AR and MA polynomials, respectively, and $|z| \leq 1$, where z is a complex number, γ_w is the autocovariance of the w_t with parameter d and σ^2 , that is,

$$\gamma_w(h) = \gamma_w(0) \rho_w(h), \quad (2.56)$$

with $\gamma_w(0)$ and $\rho_w(h)$ as in (Equation 2.53) and (Equation 2.54).

The spectral density of x_t is given by

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2} |1 - e^{-i\lambda}|^{-2d} \quad (2.57)$$

The calculation of the exact Gaussian likelihood of observations x_1, \dots, x_n of a fractionally integrated ARMA process is very slow and demanding in terms of computer memory. Therefore, instead of estimating the parameters $d, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$ by maximising the exact Gaussian likelihood, it is much simpler, for example, to maximise the Whittle approximation L_W [10], defined by

$$-2\ln(L_w) = n\ln(2\pi) + 2n\ln\sigma + \sigma^{-2} \sum_j \frac{I_n(\omega_j)}{g(\omega_j)} + \sum_j \ln g(\omega_j), \quad (2.58)$$

where I_n is the periodogram, $\sigma^2 g/(2\pi)$ ($= f$) is the model spectral density, and \sum_j denotes the sum over all non-zero Fourier frequencies $\omega_j = 2\pi j/n \in (-\pi, \pi)$. There are several techniques available to estimate the d parameter, either in the time or in the frequency domain. One well-known technique is the Geweke and Porter-Hudak (GPH) estimator, given by

$$\begin{aligned} \ln\{I(\lambda_{j,T})\} &= \ln\{\sigma^2 f_u(0)/2\pi\} - d \ln\{4\sin^2(\lambda_{j,T}/2)\} + \\ &\quad \ln\{f_u(\lambda_{j,T})/f_u(0)\} + \ln\{I(\lambda_{j,T})/f(\lambda_{j,T})\}, \end{aligned} \quad (2.59)$$

where $f(\lambda) = (\sigma^2/2\pi)\{4\sin^2(\lambda)\}^{-d} f_u(\lambda)$.

Equation 2.59 shows the GPH method, where $\lambda_{j,T} = 2\pi j/T$ ($j = 0, \dots, T-1$) and $I(\lambda_{j,T})$ represents the periodogram at the selected harmonic ordinates. This equation resembles a simple linear regression estimation, where $\ln\{I(\lambda_{j,T})\}$ is the dependent variable, $\ln\{4\sin^2(\lambda_{j,T}/2)\}$ is the explanatory variable, $\ln\{I(\lambda_{j,T})/f(\lambda_{j,T})\}$, $\ln\{I(\lambda_{j,T})/f(\lambda_{j,T})\}$ is the disturbance, $-d$ is the slope, and the intercept term is $\ln\{\sigma^2 f_u(0)/2\pi\}$ plus the mean of $\ln\{I(\lambda_{j,T})/f(\lambda_{j,T})\}$. As for the term $\ln\{f_u(\lambda_{j,T})/f_u(0)\}$, this becomes negligible as attention is confined to harmonic frequencies nearer zero. For further details on Geweke and Porter-Hudak method, see reference [23].

The long-range dependence of a time series can be characterised by the Hurst exponent, which is named after its inventor. Hurst investigated the behaviour of the rescaled adjusted

range analysis (R/S), a statistic defined for a time series as

$$R/S = s^{-1} \max(U_1, \dots, U_c), \quad (2.60)$$

where $U_j = \sum_{t=1}^n (x_t - \bar{x})$, $\bar{x} = \sum_{t=1}^n x_t$ and $s^2 = n^{-1} \sum_{t=1}^n (x_t - \bar{x})^2$ [29]. A value of the Hurst exponent higher than 0.5 suggests that the time series has long-term positive autocorrelation, whereas a Hurst exponent value < 0.5 indicates a time series with long-term switching between high and low values. When the Hurst exponent is equal to 0.5 it means that the time series do not has long-range dependency. The Hurst exponent and the fractional differencing parameter are related through the following expression $d = H - 0.5$ [43].

SARFIMA

In the literature one can find models with the designation of SARFIMA, where these models are allowed to have $0 < |d| < 0.5$ and $0 < |D| < 0.5$. Nevertheless, by SARFIMA we consider the following model

$$\Phi_P(B^S) \nabla_S^D x_t = \Theta_Q(B^S) z_t, \quad (2.61)$$

where $\Phi_P(B^S)$ is the AR seasonal operator, $\Theta_Q(B^S)$ is the MA seasonal operator, S is the seasonality, ∇_S^D is the seasonal differencing operator with $D \in \mathbb{N}_0$ and z_t , is an autocorrelated error with an ARFIMA(p, q, d) structure, that is,

$$\phi_p(B) \nabla^d z_t = \theta_q(B) \varepsilon_t. \quad (2.62)$$

Therefore, these models are a pure seasonal SARIMA with a given seasonality, combined with an error that follows an ARFIMA(p, d, q) process, which allow a suitable decomposition of a time series of high-frequency sampling (hourly frequency or higher) with a long and short memory component. SARFIMA(p, d, q)(P, D, Q) is a stationary process if $d + D < 0.5$ and $\phi_p(z) \Phi_P(z^S) \neq 0$, for $|z| \leq 1$. Furthermore the stationary process has long-memory property if $0 < d + D < 0.5$, $0 < D < 0.5$ and $\phi_p(z) \Phi_P(z^S) \neq 0$, for $|z| \leq 1$ [9].

2.1.2 Practical Implementation Aspects

SARIMA

As mentioned previously, Box and Jenkins established a framework to choose the order of ARIMA model parameters [50, 45]. However, this approach has some drawbacks, namely it is not entirely objective, its implementation requires careful examination of the data by a knowledgeable and experienced analyst, furthermore, it may fail to unambiguously identify a model [30]. To overcome the subjectivity associated with the Box-Jenkins methodology,

several attempts have been made to automate ARIMA modelling in the last decades. One of the most recent attempts is the implementation in the *forecast* package for R software [31].

Following (Equation 2.35), the $\text{SARIMA}(p, d, q)(P, D, Q)_S$ process can be expanded to

$$\Phi_P(B^S)\phi_p(B)(1 - B^S)^D(1 - B)^d y_t = c + \Theta_Q(B^S)\theta_q(B)\varepsilon_t \quad (2.63)$$

However, R uses a different model parameterisation

$$\Phi_P(B^S)\phi_p(B)(y'_t - \mu) = \Theta_Q(B^S)\theta_q(B)\varepsilon_t, \quad (2.64)$$

where $y'_t = (1 - B)^d(1 - B^S)^D y_t$ and μ is the mean of y'_t . To convert the R parameterisation to (Equation 2.63) it is sufficient to set $c = \mu\Phi_P(B^S)\phi_p(B)$.

Hyndman and Khandakar [31] present the algorithm implemented at *auto.arima* function. In summary, they define a methodology to estimate all model parameters, starting with the differencing parameters (d and D). In the case of SARIMA models, the D parameter is firstly estimated, followed by the d parameter. The default method to estimate the D parameter is a measure of seasonal strength computed from an STL decomposition [32]. As for the non-seasonal parameter, d , this is selected using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit roots test. For further details on the implementation of these parameters estimation, please refer to [31, 32, 35]. After setting D and d , a step-wise algorithm is followed to determine the order of the remaining parameters based on the Akaike Information Criteria

$$\text{AIC} = -2\log(L) + 2(z), \quad (2.65)$$

where $z = p + q + P + Q + k$, which are the model parameters in the case of SARIMA models, $k = 1$ if $c \neq 0$ and $k = 0$ otherwise, L is the maximised likelihood of the model fitted to the differenced data $(1 - B)^d(1 - B^S)^D y_t$.

Another measure that can be used to model selection is the Bayesian Information Criteria

$$\text{BIC} = \log(n)(z) - 2\log(L), \quad (2.66)$$

where $z = p + q + P + Q + k$ for SARIMA models, L is the maximised likelihood of the fitted model and n are the number of observations.

Below the step-wise algorithm used in this work is described:

- **Step 1:** Start by trying four possible models
 - $\text{ARIMA}(2, d, 2)$ if $S = 1$ and $\text{SARIMA}(2, d, 2)(1, D, 1)_S$ if $S > 1$
 - $\text{ARIMA}(0, d, 0)$ if $S = 1$ and $\text{SARIMA}(0, d, 0)(0, D, 0)_S$ if $S > 1$

- ARIMA(1, d , 0) if $S = 1$ and SARIMA(1, d , 0)(1, D , 0) $_S$ if $S > 1$
- ARIMA(0, d , 1) if $S = 1$ and SARIMA(0, d , 1)(0, D , 1) $_S$ if $S > 1$

Of the four models, the one with the smallest AIC value is selected. This model is designated as the *current* model, which is denoted by ARIMA(p, d, q) if $S = 1$ or SARIMA(p, d, q)(P, D, Q) $_S$ if $S > 1$. Note that, when $d + D \leq 1$, the models are fitted with $c \neq 0$, otherwise $c = 0$.

• **Step 2: Consider up to thirteen variations of the *current* model**

- where one of p , q , P and Q is allowed to vary by ± 1 from the *current* model;
- where p and q both vary by ± 1 from the *current* model;
- where P and Q both vary by ± 1 from the *current* model;
- where the constant c is included if the current model has $c = 0$ or excluded if the current model has $c \neq 0$.

At each step, the model with lower AIC becomes the *current* model and the procedure is repeated until the algorithm is not able to find a model with lower AIC than the *current* model. In this algorithm, models are computed using the maximum likelihood estimation method. To achieve convergence of the fitted models some constraints were imposed on the algorithm to avoid problems with convergence or near unit-roots. Further details on the algorithm and its implementation can be found at [31, 30].

Dynamic Harmonic Regression

Dynamic Harmonic Regression models are also fitted using the *auto.arima* function. This function has a parameter *xreg*, which allows specifying the external regressions, that is, the coefficients of the linear regression. Furthermore, the package also includes a function designated *fourier* that computes the Fourier terms of a given order for the time series.

Since these models are estimated with *auto.arima*, the same algorithm described in the previous section is used to obtain the optimal orders (p, d, q) of the model. Models with seasonality are disregarded as the seasonality is now described by the regression coefficients, i.e, the Fourier terms. It is also worth to mention that the *auto.arima* function estimates the joint-model by maximum-likelihood and not separately [21]. This means that the orders of the model, as well as the regression coefficients, are obtained together and not sequentially. Finally, model selection is performed based on AIC.

ARFIMA

The Hurst exponent characterises the long-range dependence. To assess the long-range dependence of the air pollutants time series, the Rescaled Range (R/S) analysis and the

adjusted (R/S) analyses were computed. First, it is necessary to divide the time series of length L into k subseries of length n . Afterwards, for each subseries $m = 1, \dots, k$ it is necessary to:

1. compute the mean (E_m) and standard deviation (S_m);
2. normalise the data ($Z_{i,m}$) by subtracting the sample mean $X_{i,m} = Z_{i,m} - E_m$ for $i = 1, \dots, n$;
3. build a cumulative time series $Y_{i,m} = \sum_{j=1}^i X_{j,m}$ for $i = 1, \dots, n$
4. find the range $R_m = \max\{Y_{1,m}, \dots, Y_{n,m}\} - \min\{Y_{1,m}, \dots, Y_{n,m}\}$
5. rescale the range R_m/S_m

Finally, calculate the mean value of the rescaled range for all subseries of length n

$$(R/S)_n = \frac{1}{k} \sum_{m=1}^k R_m/S_m. \quad (2.67)$$

Since, it is known that R/S statistics asymptotically follows the relation

$$(R/S)_n \sim cn^H, \quad (2.68)$$

the Hurst exponent is given by

$$\log(R/S)_n = \log(c) + H\log(n) \quad (2.69)$$

Equivalently, a plot of the $(R/S)_n$ statistics against n on a double-logarithmic paper. If the time series is a white noise process, the plot will be a straight line with slope 0.5. If the process is persistent the slope is > 0.5 , whereas if the process is anti-persistent the slope is < 0.5 . However, it should be noted that for small n there is a significant deviation from the 0.5 slope. For this reason, the values of the R/S statistics can be approximated by

$$E(R/S)_n = \begin{cases} \frac{n-\frac{1}{2}}{n} \frac{\Gamma((n-1)/2)}{\sqrt{\pi}\Gamma(n/2)} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{if } n \leq 340, \\ \frac{n-\frac{1}{2}}{n} \frac{1}{\sqrt{n\pi/2}} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{if } n > 340. \end{cases} \quad (2.70)$$

where Γ is the Euler Gamma function. This is called the corrected R/S method [57]. The Hurst exponent was computed with function *hurstexp* from the *pracma* package version 2.1.4 for R software.

In order to estimate the ARFIMA models, we used the function *arfima* from the *forecast* package. This function combines the *fracdiff* function of the same named package,

with the above-presented algorithm (*auto.arima* function) to estimate the ARFIMA models. The *fracdiff* function uses the maximum likelihood function to estimate the parameters of a fractionally-differenced ARIMA model [19]. This package implements a likelihood approximation using the fast and accurate method of Haslett and Raftery (1989). In summary, an approximation of the conditional means and variances using only the partial autocorrelations for the fractionally differenced ARIMA(0, d , 0) process is done, followed by finding an approximation of maximum likelihood estimators of μ and σ^2 analytically. Then, a concentrated likelihood function is found and, finally, the partial linear regression coefficients of the ARIMA(0, d , 0) process is approximated. For further details of the maximum likelihood approximation method, please refer to [27].

The *arfima* function is implemented to firstly assume an AR(2) model in order to determine the initial value of the fractional differencing parameter (d). Then, after the initial d is selected the *auto.arima* function with the algorithm described in (Section 2.1.2) is applied to determine the AR and MA orders. However, it was decided to change the estimation method of the initial value of the fractional differencing parameter, d , as this assumes an autocorrelation structure that may not be the best to describe the data. The Geweke and Porter-Hudak (GPH) method was used to estimate the value of the initial d parameter as this is based on the regression equation using the periodogram function as an estimate of the spectral density, and thus, takes into account the observed information of the data [23].

Finally, the model is re-estimated using the AR and MA orders to estimate the final d value. After obtaining the final model with the AR and MA orders and the value of the d parameter, the coefficients were refined using exact maximum likelihood estimation, which is estimated through the *arima* function from base R package *stats*. In this function, the exact likelihood is computed via a state-space representation of the ARIMA process, and the innovations and their variance are found by a Kalman filter [54].

There is an additional question regarding the ARFIMA model estimation that it is necessary to discuss, the computation of the Hessian matrix (*fracdiff* function). In fact, what is computed is a finite difference approximation to the Hessian. To do so, it is necessary to establish the size of the finite difference interval for numerical derivatives (h). This h parameter influences the covariance matrix, the correlation matrix and the standard error of the estimated parameters, but not the parameters' coefficients. It is important to obtain the standard error measures to infer regarding the significance of the coefficients of the model. The h parameter is estimated by default as

$$h = \min\{0.1, \sqrt{a} \times (1 + |\text{Log}(\text{Likelihood})|)\}, \quad (2.71)$$

where $a = .Machine\$double.neg.eps = 1.110223e^{-16}$, which is a small positive floating-point

number x such that $1 - x \neq 1$. Nevertheless, for most time series of air pollutants, using the default h value did not allow the computation of the standard error and, as a consequence it was not possible to obtain the coefficients' p -values. Since this is a limitation that the authors of the package know that is possible to occur, another function within the package is available to recompute the Hessian matrix with different h values (*fracdiff.var*). We developed an algorithm, using this function, to determine the h' value (if it exists) closest to h , for which the Hessian matrix is stable and it is possible to compute the standard error of the coefficients. The algorithm is presented in Appendix B, as well as some illustrative results of the impact of h value on the standard error estimates. In summary, we verified that if h was changed it was not reliable to conclude regarding the statistical significance of the coefficients, particularly for the fractional differencing parameter (d). Therefore, it was decided that the model would be computed with the default h value, and no discussion regarding the significance of the coefficients would be performed. The model will reflect the influence of the most important coefficients, and the remaining will have little impact on the estimated values.

SARFIMA

SARFIMA models were implemented by combining the abovementioned strategies. Firstly, a $\text{SARIMA}(p, d, q)(P, D, Q)_S$ model was fitted to the observed air pollution time series following the algorithm previously described for SARIMA models. Afterwards, an ARFIMA(p, d, q) model was fitted to the residuals of the SARIMA model, which results in a model as described in (Equation 2.61). Thus, the resulting model is a SARIMA model with errors that follow an ARFIMA(p, d, q), instead of following a white noise distribution.

In order to produce forecasts with these models, firstly, the residuals of the SARIMA model are replaced by the fitted values of the ARFIMA, that is,

$$\Phi(B^S)(1 - B^S)^D x_t = c + \Theta(B^S)\hat{z}_t, \quad (2.72)$$

where $\phi(B)(1 - B)^d \hat{z}_t = \theta(B)\varepsilon_t$, $\varepsilon_t \sim \text{wn}(0, \sigma_w^2)$, $d \in]0, 0.5[$ and $D \in \mathbb{N}_0$.

Afterwards the forecast of the SARIMA model is performed with the *forecast* function. Point forecast of x_t can be obtained by expanding (Equation 2.72) and write it in order to x_t , replace $t = T + 1$, and lastly, by replacing the future observations with their forecasts, future errors with zero, and past errors with the corresponding errors. Multi-step-ahead point forecasts are obtained recursively from the previous point forecast [30]. Bellow is presented an example for an ARIMA(1, 1, 1). First, the model equation is written

$$(1 - \hat{\phi}_1 B)(1 - B)x_t = (1 + \hat{\theta}_1 B)\varepsilon_t, \quad (2.73)$$

that can be expanded and re-written as

$$x_t = x_{t-1} + \hat{\phi}_1 x_{t-1} - \hat{\phi}_1 x_{t-2} + \varepsilon_t + \hat{\theta}_1 \varepsilon_{t-1}, \quad (2.74)$$

which is equivalent to,

$$x_t = (1 + \hat{\phi}_1)x_{t-1} - \hat{\phi}_1 x_{t-2} + \varepsilon_t + \hat{\theta}_1 \varepsilon_{t-1}. \quad (2.75)$$

Now, the equation looks like a ARMA(2, 1) process. Replacing $t = T + 1$, we have

$$x_{T+1} = (1 + \hat{\phi}_1)x_T - \hat{\phi}_1 x_{T-1} + \varepsilon_{T+1} + \hat{\theta}_1 \varepsilon_T. \quad (2.76)$$

Assuming that we have observation up to time T

$$\hat{x}_{T+1|T} = (1 + \hat{\phi}_1)x_T - \hat{\phi}_1 x_{T-1} + \varepsilon_{T+1} + \hat{\theta}_1 \varepsilon_T. \quad (2.77)$$

With the exception of ε_{T+1} and ε_T , all remaining values are known. The future error (ε_{T+1}) is replaced by zero and ε_T is replaced with the last observed residual. For $t = T + 2$, the same process is performed. Similarly to the previous forecast, $\varepsilon_{T+2} = \varepsilon_{T+1} = 0$. Now, x_{T+1} is not known but can be replaced by $\hat{x}_{T+1|T}$, as follows

$$\hat{x}_{T+2|T} = (1 + \hat{\phi}_1)\hat{x}_{T+1|T} - \hat{\phi}_1 x_T + \varepsilon_{T+2} + \hat{\theta}_1 \varepsilon_{T+1}. \quad (2.78)$$

This process continues recursively up to the horizon defined for the forecasts.

Since any ARIMA(p, d, q), can be written as an ARMA(p, q) process, the h-step-ahead prediction interval, for $h \geq 1$ with a $(1 - \alpha)$ probability can be obtained from

$$\left[\hat{x}_{t+h} - q_{\alpha/2} \hat{\sigma}^2(h); \hat{x}_{t+h} + q_{1-\alpha/2} \hat{\sigma}^2(h) \right], \quad (2.79)$$

where q_α is the quantile of the error distribution, $\hat{\sigma}^2$ is the estimated error variance and $\hat{\psi}_j$ are the estimated coefficients of the MA representation of the ARMA(p,q) process.

$$\hat{\sigma}^2(h) = \hat{\sigma}^2 \sum_{j=0}^{h-1} \hat{\psi}_j^2. \quad (2.80)$$

For further details on the prediction intervals estimates, please refer to [10]. The prediction intervals for ARIMA models are computed based on the assumption that the residuals are uncorrelated and normally distributed. If these assumptions do not hold, the prediction intervals may be incorrect. Therefore, if any of these assumptions are not met, an alternative is to compute the prediction intervals via bootstrap. The implementation in the forecast package computes bootstrap with 5000 bootstrapped sample paths. The estimation of the

forecasts and of the prediction intervals in the *forecast* function is done with a Kalman Filter. For further details on Kalman filters refer to [34].

To assess the performance of the model forecasts, the following measures were used: mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean percent error (MPE), mean absolute percent error (MAPE) and mean absolute scaled error (MASE) [33]. Let x_t and f_t denote the observation and the forecast, respectively, at time t . Then, the forecast error (e_t) is defined as

$$e_t = x_t - f_t. \quad (2.81)$$

The mean error is

$$\text{ME} = \frac{1}{h} \sum_{t=1}^h e_t, \quad (2.82)$$

the root mean square error is expressed as

$$\text{RMSE} = \sqrt{\left(\frac{1}{h} \sum_{t=1}^h e_t^2 \right)} \quad (2.83)$$

and, lastly, the mean absolute error is given by

$$\text{MAE} = \frac{1}{h} \sum_{t=1}^h |e_t|. \quad (2.84)$$

All of the above measures have the advantage of being scale-dependent, therefore it is possible to conclude regarding the error magnitude towards the distribution of the time series values. However, the ME and the RMSE are more sensitive to outliers than the MAE, which makes the later preferable compared to the former measures [33].

Regarding, MPE and MAPE, these measures are scale-independent as they are percentages. Therefore, their interpretation is also more straightforward, as it shows the mean deviation of the forecast in percentage compared to the true observation. These measures are particularly interesting when the goal is to compare the forecast performance of different methods and/or models [33]. The mean percent error is defined as

$$\text{MPE} = \frac{1}{h} \sum_{t=1}^h \frac{e_t}{x_t} \times 100\%, \quad (2.85)$$

whereas the mean absolute percent error is given by

$$\text{MAPE} = \frac{1}{h} \sum_{t=1}^h \frac{|e_t|}{x_t} \times 100\%. \quad (2.86)$$

Despite their advantages, these measures have the disadvantage of being infinite or undefined if $x_t = 0$, for any t . Furthermore, these measures also present a skewed distribution if any x_t is close to zero [33].

Lastly, the MASE was introduced by [33]. This measure is a scaled error, which means that this measure tries to remove the effect of the scale of the data, by comparing the method's performance to some benchmark forecast method, in particular, the naive forecast method. MASE scales the error based on the in-sample MAE from the random walk forecast method. A scaled error is defined as

$$q_t = \frac{e_t}{\frac{1}{n-1} \sum_{i=2}^n |x_t - x_{t-1}|}. \quad (2.87)$$

Scaled errors are less than one if they arise from better forecasts than the average one-step random walk forecast computed in-sample. On the other hand, if the forecast is greater than one, it means that the forecast is worse than the average one-step naive forecast computed in-sample [33]. The mean absolute scaled error is given by

$$\text{MASE} = \frac{1}{h} \sum_{t=1}^h q_t. \quad (2.88)$$

All continuous time series models were fitted with the *forecast* and *fracdiff* packages, versions 8.4 and 1.4-2, respectively. Air pollutants time series were fitted with observations from 2005 up to 2015. The year of 2016 was forecasted and model performance was conducted by comparing the 2016 forecasts with the observed values of that year.

2.2 Discrete Time Series

Count time series, integer-valued time series, discrete-valued time series are all terms to refer to time series of counts in which observations are non-negative. Count time series appear naturally when a number of events per time period are observed over time. An excellent example of such series is hospital admissions. These time series are non-negative, hence models should adequately capture their non-negativity and should also be flexible enough to suitably capture the dependence of the observations. Similarly to continuous time series there are several classes of models that may be considered. For instance, there are integer ARMA (INARMA) models, which are the integer-valued counterpart to the conventional ARMA models. The INARMA adapts the ARMA recursion to integer values using an appropriate thinning operation [56]. Another approach to model count time series is with generalised linear models (GLM) due to its convenience and flexibility. In fact, in hospital admissions studies the most common approach found were generalised additive models (GAM) [6], which are an extension of GLMs. Models based on a GLM approach have the following advantages comparatively to thinning-based operators: these models are able to describe covariate effects and negative correlations in a straightforward way and there is a rich toolkit available for this class of models [39].

2.2.1 Theoretical Aspects

Let $\{y_t : t \in \mathbb{N}\}$ denote a count time series, $\{\mathbf{X}_t : t \in \mathbb{N}\}$ a time-varying r -dimensional covariate vector, that is $\mathbf{X}_t = (X_{t,1}, \dots, X_{t,r})^T$. The conditional mean $E(y_t | \mathcal{F}_{t-1})$ of the count time series is a process $\{\lambda_t : t \in \mathbb{N}\}$, such that $E(y_t | \mathcal{F}_{t-1}) = \lambda_t$. Denote by \mathcal{F} the history of the joint process $\{y_t, \lambda_t, X_{t+1} : t \in \mathbb{N}\}$ up to time t including the covariate at time $t + 1$. Models of the following form are considered

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^p \beta_k \tilde{g}(y_{t-i_k}) + \sum_{\ell=1}^q \alpha_\ell g(\lambda_{t-j_\ell}) + \boldsymbol{\eta}^T \mathbf{X}_t, \quad (2.89)$$

where $g : \mathbb{R}^+ \rightarrow \mathbb{R}$ is a link function and $\tilde{g} : \mathbb{N}_0 \rightarrow \mathbb{R}$ is a transformation function. The vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_r)^T$ are the covariates effects. In the terminology of GLMs, $g(\lambda_t)$ is called the linear predictor. The set $P = \{i_1, i_2, \dots, i_p\}$ and the integers $0 < i_1 < i_2 < \dots < i_p < \infty$, with $p \in \mathbb{N}_0$, allow for the regression on arbitrary past observations of the response variable (y_t), whereas set $Q = \{j_1, j_2, \dots, j_q\}$ and the integers $0 < j_1 < j_2 < \dots < j_q < \infty$, with $q \in \mathbb{N}_0$, allow for regression on lagged conditional means $\lambda_{t-j_1}, \lambda_{t-j_2}, \dots, \lambda_{t-j_q}$. The specification of these sets orders is similar to the one done in ARIMA models, by analysing the empirical autocorrelation functions of the observed data [39]. Hence, the P set can account for the serial correlation, while set Q can deal with the seasonality of the time series.

Now, let us consider the case when $g(x) = \tilde{g}(x) = x$, that is the identity, and $P = \{1, \dots, p\}$, $Q = \{1, \dots, q\}$ and $\eta = 0$, which means that there are no covariates, model (Equation 2.89) can be simplified to

$$\lambda_t = \beta_0 + \sum_{k=1}^p \beta_k y_{t-k} + \sum_{\ell=1}^q \alpha_\ell \lambda_{t-\ell} \quad (2.90)$$

Further assuming that y_t given the past follows a Poisson distribution, one obtains an IN-GARCH model [39] - integer-valued generalised autoregressive conditional heteroskedasticity - which are derived from GARCH models. The ARCH models were first introduced by Engle (1982) in order to study the variability or volatility of a time series. Unlike ARMA models, which were built to model the mean of a process with constant variance, the ARCH models allow for the variance to change over time. These models were later extended to generalised ARCH (GARCH) models where the variance of the model is assumed to follow an ARMA error [50].

Considering the logarithmic link function $g(x) = \log(x)$, $\tilde{g}(x) = \log(x + 1)$, model (Equation 2.89) is written as

$$v_t = \beta_0 + \sum_{k=1}^p \beta_k \log(y_{t-k} + 1) + \sum_{\ell=1}^q \alpha_\ell v_{t-\ell}, \quad (2.91)$$

where $v_t = \log(\lambda_t)$. Model (Equation 2.90) is able to accommodate positive serial correlation only, as its covariance function is strictly positive [18]. On the contrary, model (Equation 2.91) allows to include negative serial correlation. Furthermore, it is easier to accommodate covariates in (Equation 2.91) than in (Equation 2.90) since the log-linear model implies positivity of the conditional mean process, while the identity model can include exclusively covariates which result in a positive regression term since otherwise, the mean of the Poisson process becomes negative [18]. Hence, the linear model (Equation 2.90) along with covariates should be fitted with caution because it is limited to positive effects only. Lastly, the covariates effect is additive for model (Equation 2.90) and multiplicative for model (Equation 2.91) [39]. The log-linear model has been extensively studied by Fokianos and Tjøstheim (2011), and the authors show that the use of a constant (c) in $\log(x + c)$ to avoid zero values do not affect inference. Furthermore, the authors argue that $c = 1$ is a reasonable value for the constant [18].

Model (Equation 2.89) and the Poisson assumption, that is, $y_t | \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$, imply that

$$P(y_t = y | \mathcal{F}_{t-1}) = \frac{\lambda_t^y \exp(-\lambda_t)}{y!}, \quad y = 0, 1, \dots \quad (2.92)$$

Therefore, it holds that $\text{var}(y_t|\mathcal{F}_{t-1}) = E(y_t|\mathcal{F}_{t-1}) = \lambda_t$. On the contrary, when the Negative Binomial distribution is selected to model the process, the conditional variance is allowed to be larger than the conditional mean, due to the overdispersion coefficient [39]. Following Christou and Fokianos (2014), it is assumed that $y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$, which is the Negative Binomial Distribution parametrised in terms of its mean (λ_t) with an additional dispersion parameter $\phi \in (0, \infty)$ [12], that is,

$$P(y_t = y|\mathcal{F}_{t-1}) = \frac{\Gamma(\phi + y)}{\Gamma(y + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_t} \right)^\phi \left(\frac{\lambda_t}{\phi + \lambda_t} \right)^y, y = 0, 1, \dots \quad (2.93)$$

Hence, the variance is $\text{var}(y_t|\mathcal{F}_{t-1}) = \lambda_t + \lambda_t^2/\phi$, which indicates that the conditional variance of the process increases quadratically with λ_t by $\frac{1}{\phi}$.

The Negative Binomial distribution is a mixed Poisson process. A mixed Poisson process is specified by setting $y_t = n_t(0, z_t\lambda_t]$, where $\{n_t\}$ are independent and identically distributed Poisson processes with $\mu_n = \text{var}_n = 1$ and $\{z_t\}$ are i.i.d. random variables with $\mu_z = 1$ and $\text{var}_z = \sigma^2$, which are independent of $\{y_t\}$ [18]. When $\{z_t\}$ are i.i.d. Gamma processes of random variables, one obtains the negative binomial process where $\sigma^2 = \frac{1}{\phi}$ [39]. One refers to σ^2 as the overdispersion coefficient since this is proportional to the extent of the overdispersion of the conditional distribution. When $\sigma^2 = 0$, there is no overdispersion, i.e., $\text{var}(y_t|\mathcal{F}_{t-1}) = \lambda_t$, which is the Poisson distribution.

2.2.2 Practical Implementation Aspects

To estimate the daily hospital admissions using the air pollutants as covariates, the recent package *tscount* (version 1.4.1) was used. This package was chosen due to its flexibility, since it allows to deal with dependent data, and to use the Poisson or the Negative Binomial Distributions. Furthermore, the package also allows the use of the identity or the logarithmic function as link functions. All of these functionalities are operationalised with the *tsglm* function [39].

The *tscount* package fits models of the form (Equation 2.89) by quasi conditional maximum likelihood estimation (ML). If the data follows a Poisson Distribution, an ordinary ML estimation is obtained. Nonetheless, if the distribution is a mixed Poisson, a quasi-ML estimator is retrieved. Let $\theta = (\beta_0, \beta_1, \dots, \beta_p, \alpha_1, \dots, \alpha_q, \eta_1, \dots, \eta_r)^T$ be the vector of the regression parameters. Despite the distribution considered, the parameter space of the INGARCH model (Equation 2.90) is given by

$$\Theta = \left\{ \theta \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, \dots, \beta_p, \alpha_1, \dots, \alpha_q, \eta_1, \dots, \eta_r \geq 0, \sum_{k=1}^p \beta_k + \sum_{\ell=1}^q \alpha_\ell < 1 \right\}. \quad (2.94)$$

The intercept (β_0) must be positive, and the remaining parameters must be non-negative to ensure positivity of the conditional mean (λ_t). The last condition ensures that the fitted model has a stationary and ergodic solution with moments of any order [39]. With the parametrisation of the Negative Binomial distribution as presented in (Equation 2.93) the estimation of the regression parameters ($\boldsymbol{\theta}$) does not depend on the dispersion parameter (ϕ), which allows to employ a quasi-maximum likelihood approach based on the Poisson likelihood. The dispersion parameter is estimated afterwards.

The log-likelihood, score vector and information matrix are all derived conditionally on pre-sample values of the time series and the conditional mean of the process (λ_t) at \mathcal{F}_0 . Considering a vector of observations $\mathbf{y} = (y_1, \dots, y_n)^T$, the conditional quasi log-likelihood function up to a constant is given by

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^n \log p_t(y_t; \boldsymbol{\theta}) = \sum_{t=1}^n (y_t \ln(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta})), \quad (2.95)$$

where $p_t(y; \boldsymbol{\theta}) = P(y_t = y | \mathcal{F}_{t-1})$ is the probability density function of a Poisson distribution as defined in (Equation 2.92). The conditional score function is the $(p+q+r+1)$ -dimensional vector

$$S_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^n \left(\frac{y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (2.96)$$

The conditional information matrix is given by

$$G_n(\boldsymbol{\theta}; \sigma^2) = \sum_{t=1}^n \text{cov} \left(\frac{\partial \ell(\boldsymbol{\theta}; y_t)}{\partial \boldsymbol{\theta}} \middle| \mathcal{F}_{t-1} \right) = \sum_{t=1}^n \left(\frac{1}{\lambda_t(\boldsymbol{\theta})} + \sigma^2 \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T. \quad (2.97)$$

If the Poisson distribution is used, then $\sigma^2 = 0$, otherwise, if the Negative Binomial distribution is considered $\sigma^2 = \frac{1}{\phi}$. Let $G_n^*(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}; 0)$ denote the Poisson conditional information matrix.

Assuming, that the quasi-maximum likelihood estimator ($\hat{\boldsymbol{\theta}}_n$) of $\boldsymbol{\theta}$ exists, it is the solution of the optimisation problem

$$\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}_n = \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}) \quad (2.98)$$

Let $\hat{\lambda}_t = \lambda_t(\hat{\boldsymbol{\theta}})$ denote the fitted values. Since the parameters have been obtained, now it is possible to estimate the dispersion parameter. According to [12], the dispersion parameter (ϕ) of the Negative Binomial distribution can be estimated by solving the following equation

based on the Pearson's χ^2 -statistics

$$\sum_{t=1}^n \frac{(y_t - \hat{\lambda}_t)^2}{\hat{\lambda}_t + \hat{\lambda}_t^2 / \hat{\phi}} = n - (p + q + r + 1), \quad (2.99)$$

and the variance parameter is estimated by $\hat{\sigma}^2 = \frac{1}{\hat{\phi}}$. In the particular Poisson case, $\hat{\sigma}^2 = 0$.

Regarding the inference of the parameters, this is based on the asymptotic normality of the quasi-maximum likelihood (QMLE) estimator for models without covariates [12]. For well-behaved models with covariates, the package authors' hypothesise that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N_{p+q+r+1}(\mathbf{0}, G_n^{-1}(\hat{\boldsymbol{\theta}}_n; \hat{\sigma}^2) G_n^*(\hat{\boldsymbol{\theta}}_n) G_n^{-1}(\hat{\boldsymbol{\theta}}_n; \hat{\sigma}^2)), \quad (2.100)$$

as $n \rightarrow \infty$, where $\boldsymbol{\theta}_0$ is the value of the true parameter and $\hat{\sigma}^2$ is a consistent estimator of σ^2 . The authors suppose that these assumptions hold under the same assumptions usually made for the ordinary linear regression model. The package authors' did not derive these theoretical properties for QMLE, instead, they have performed simulations studies of its asymptotic normality. Nonetheless, the authors offer an alternative approach to derive the standard errors and the confidence intervals: a parametric bootstrap procedure. This approach does not require the normal approximation. Therefore, in order to obtain these estimates, we opted by performing the bootstrap procedure, considering $B = 500$, which the authors claim to achieve stable results [39].

Regarding forecasts, the optimal one-step-ahead predictor \hat{y}_{t+1} for y_{t+1} , in terms of the mean square error given the past of the process up to time t and potential covariates at time $t+1$, is the conditional expectation λ_{t+1} given in (Equation 2.89). Analogously, a h -step-ahead prediction \hat{y}_{t+h} for y_{t+h} is obtained by recursively compute the 1-step-ahead prediction. The distribution of the h -step-ahead prediction is not known analytically, but can be approximated by a parametric bootstrap procedure based on B simulations of the realisations $y_{t+1}^{(b)}, \dots, y_{t+h}^{(b)}$ from the fitted model, $b = 1, \dots, B$. The prediction intervals of the h -step-ahead prediction (\hat{y}_{t+h}) computed with a coverage rate of $1 - \alpha$ are designed to include the true observation y_{t+h} with a $1 - \alpha$ probability. Furthermore, it is possible to compute prediction intervals with a global coverage rate of $1 - \alpha$, which is achieved by a Bonferroni adjustment of the individual coverage rates to $\frac{1-\alpha}{h}$. Therefore, the predictions (forecasts) of daily hospital admissions considering the covariates were performed considering a global coverage rate of 0.95 and the prediction intervals were derived with a parametric bootstrap of $B = 1000$. The forecasts were obtained using the *predict* function with the method *predict.glm*. Performance of the forecasts was evaluated using the same measures presented for the continuous time series in Section 2.1.2. For further details on the estimation method, prediction or implementation, please refer to [39].

Chapter 3

Results

This chapter is divided in two sections, the first describes air pollutants results and the second presents the results obtained for hospital admissions at Aveiro. Air pollution levels are generally higher in urban environment stations and stations with traffic influence. Of the approaches tested, SARFIMA models, which resulted from fitting a SARIMA model to the observed air pollution time series followed by fitting an ARFIMA model to the residuals of the SARIMA model, was the best to describe hourly pollutants data. One-year forecasts of SARFIMA models are presented. In general, there is good model performance for all pollutants studied. The forecasts converge to the time series mean in about four months.

Regarding hospital admissions in Aveiro, the optimal model built under empirical assumptions showed that CO and NO_x, lagged 5 and 6 days, respectively, were significantly associated with all-cause respiratory hospital admissions. Hospital admissions forecasts using observed pollutants data have good performance. Additionally, the results show that 10-days hospital admissions forecast using air pollutants SARFIMA forecasts are as good as the hospital admissions forecasts using the observed pollutants data.

3.1 Air Pollutants

3.1.1 Descriptive Results

Figure 3.1 displays air pollutants levels at rural, urban and suburban areas. Whenever possible Ervedeira, Estarreja and Olivais are used as examples of rural, suburban and urban stations, respectively. Despite having different environments (rural, suburban or urban), these stations have the same type of influence: background. The pollutants are presented in the following order: NO_2 , NO_x , PM_{10} , $\text{PM}_{2.5}$, O_3 , SO_2 and CO. For carbon monoxide, the time series is displayed only at Olivais (urban), since monitoring of this pollutant is performed mainly at urban settings.

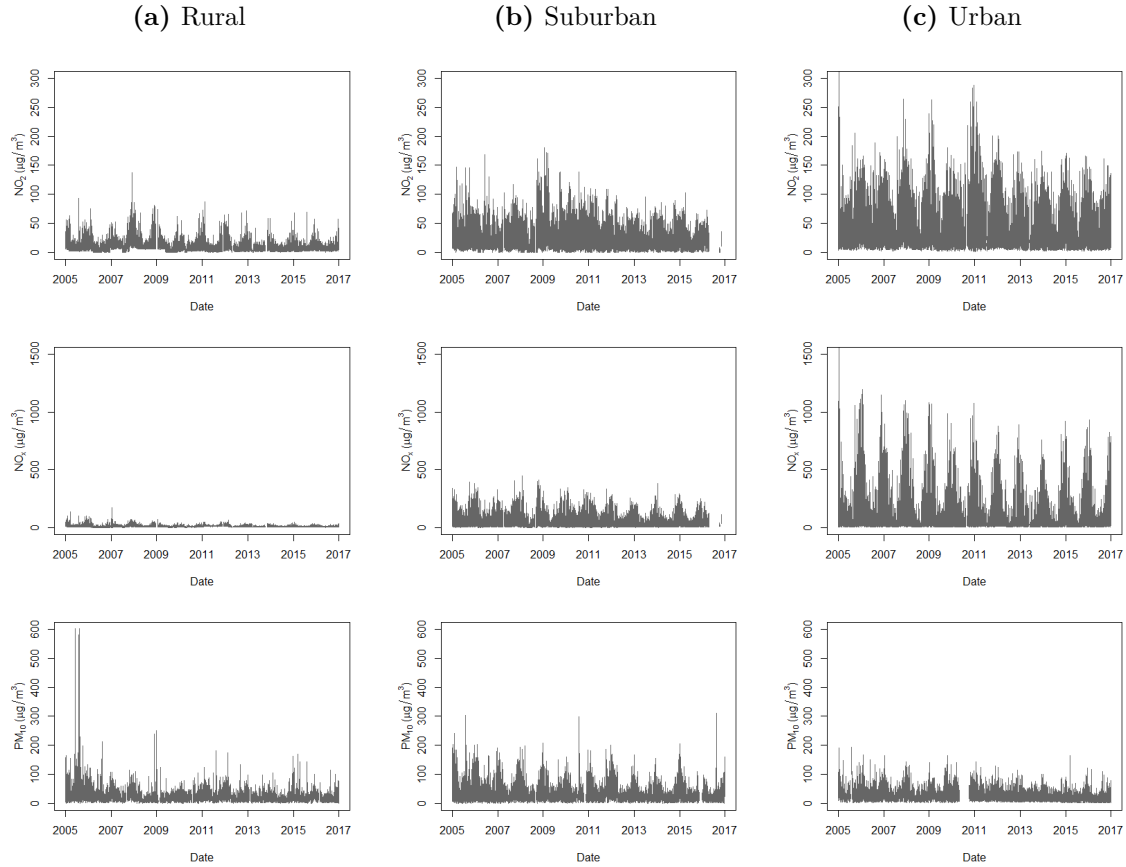


Figure 3.1: Time series of air pollutants with background influence. (a) rural , (b) suburban and (c) urban. (To be continued)

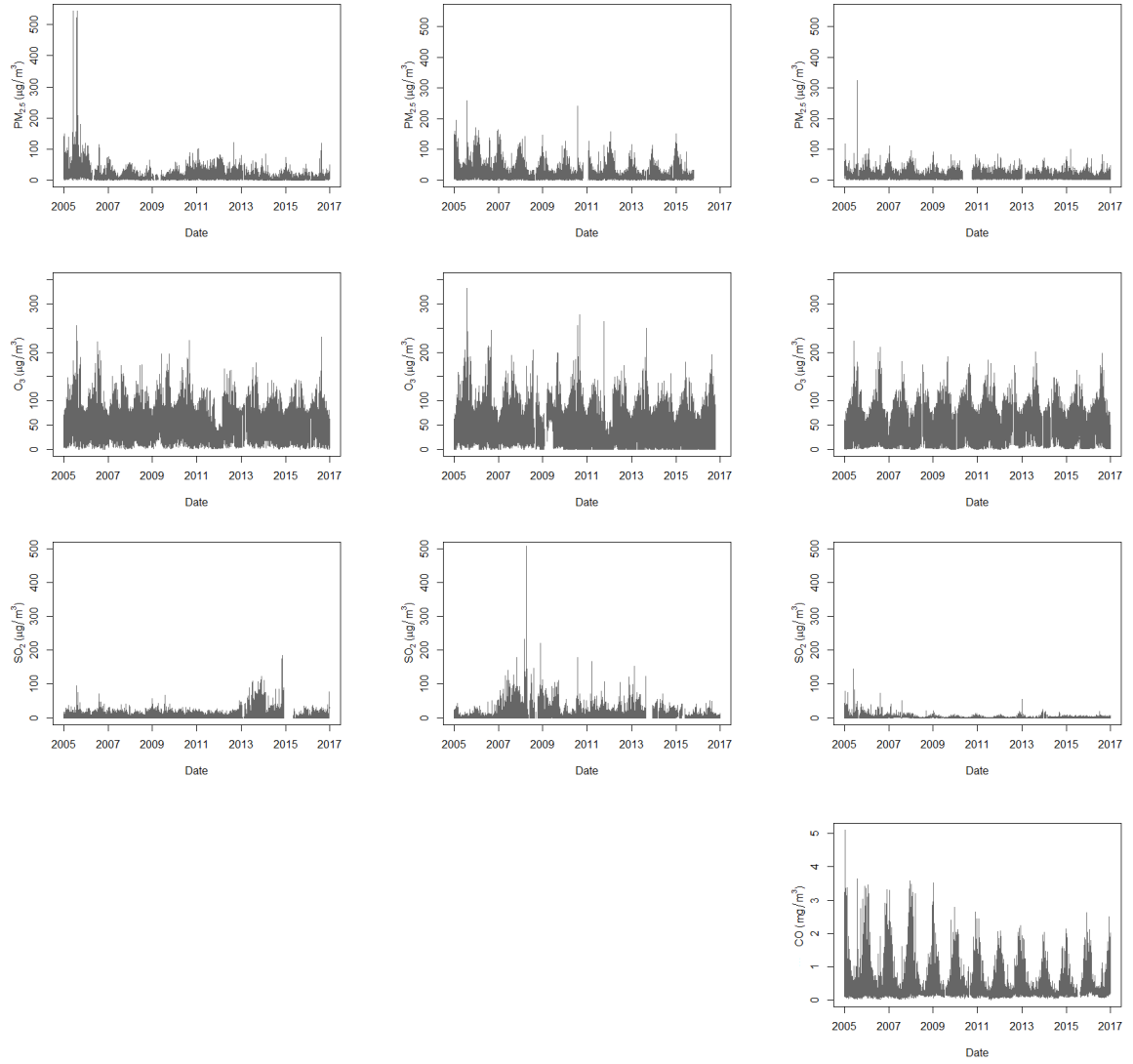


Figure 3.1: Time series of air pollutants with background influence. (a) rural, (b) suburban and (c) urban stations. (Continued)

For the displayed stations, NO_2 and NO_x have similar patterns. The urban station presents higher values than the rural and suburban areas for both pollutants. Nevertheless, this pattern does not seem to persist for all included stations. Figure 3.2 shows the distribution of the mean according to the type of environment and influence for each pollutant. Regarding NO_2 mean values, the median is higher for urban areas compared to rural areas, however, suburban areas have similar median values to urban areas. Also, if it were not the presence of an observation identified as an outlier by the box plot, the suburban environment would have higher mean values than the urban environment. Of course, it is relevant to point out that the majority of the stations are from an urban environment and only 6 and 5 stations are from suburban and rural environments, respectively. When considering the influence of each station, it is clear that areas with traffic influence have higher NO_2 mean values than industrial and background areas, which are fairly similar among themselves [Figure 3.2(a)]. Most stations considered are from a background influence (16), while 4 have an industrial influence and 9 have a traffic influence. Results for NO_x are identical to those obtained for NO_2 as depicted in Figure 3.2(b). It is not surprising that NO_2 and NO_x values are higher in stations influenced by traffic as their main emission sources are road transport [62]. Overall NO_x levels are higher than NO_2 , which is expected since NO_x levels also include NO_2 besides nitrogen oxides [Figures 3.1 and 3.2(a)].

Regarding PM_{10} , overall time series values are smaller for Ervedeira, whereas the suburban station, Estarreja, seems to have higher values than the urban station (Figure 3.1). Nonetheless, Figure 3.2(c) suggests that the median of PM_{10} mean values are similar between urban and suburban environments. Again, the number of urban stations is the largest, with only 5 stations being from suburban and rural environments. When assessing the distribution of mean PM_{10} values according to the type of influence, traffic-influenced stations have the highest median values followed by background stations [Figure 3.2(c)]. Only 2 stations are from an industrial influence, hence this may not be representative. The main source of particulate matter was thought to be traffic exhaust emissions [41], but more recent studies showed that non-exhaust emissions from road traffic [5], biomass burning and cooking [42] are also considerable sources of pollution, which might explain the high values found at background areas.

With regards to $\text{PM}_{2.5}$, a similar trend to PM_{10} is observed (Figure 3.1). It is clear that the urban station has smaller $\text{PM}_{2.5}$ values compared to the suburban station. Figure 3.2(d) shows the distribution of mean $\text{PM}_{2.5}$ values according to the type of environment and influence. Urban and suburban areas have clearly higher $\text{PM}_{2.5}$ mean values than rural areas. Furthermore, there is one station only with traffic influence, and this has the highest mean value compared to most background stations. The outlier found in the background group is

the suburban station.

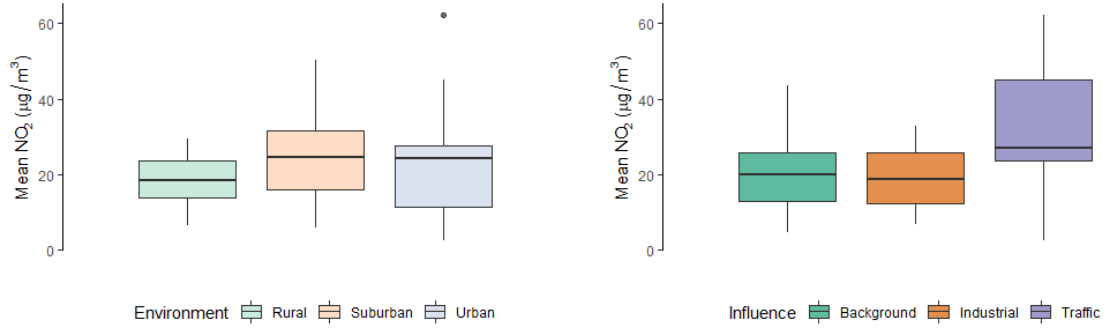
Unlike the previous pollutants, O_3 levels are higher in Ervedeira, the rural station, than in the remaining (Figure 3.1). This may be due to the consumption of O_3 in the reaction that originates NO_2 (Equation 3.1.1).



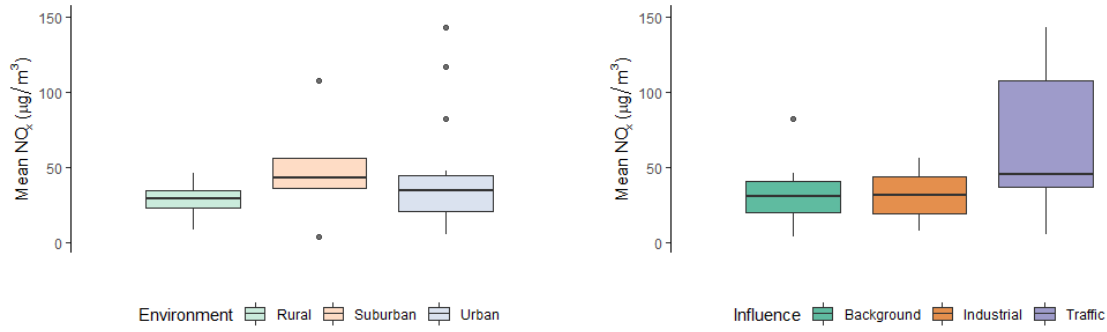
Since NO levels are higher in urban and suburban areas, it reacts with O_3 to originate NO_2 . On the contrary, in rural areas, NO levels are decreased and this reaction occurs at a smaller extension, which results in a larger concentration of O_3 in the atmosphere. Figure 3.2(e) supports that O_3 mean levels are higher in stations in rural environments than the remaining. When analysing the mean distribution by type of influence, its median value is similar for rural and industrial settings. For stations influenced by traffic, the median value is smaller, which makes sense since these areas have higher NO_2 values. Nevertheless, one must be cautious in concluding regarding traffic influence since only two stations are included.

As illustrated in Figure 3.1 overall SO_2 values are higher in Estarreja, the suburban location, while for the rural and urban location values are lower. However, when observing Figure 3.2(f), it is clear that the median of SO_2 mean values are smaller for the suburban and urban settings. Also, the values among all environments are superimposed. Hence, it seems that SO_2 are not differentiated according to the type of environment. In contrast, when analysing the type of influence, it is clear that stations with industrial influence have higher mean SO_2 values than the remaining. It is noteworthy that only one station with traffic influence is included.

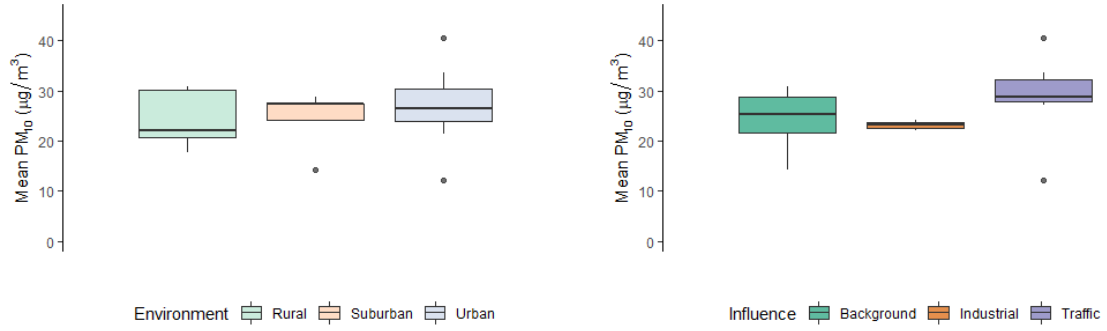
Finally, the CO time series is shown at Olivais. Carbon monoxide levels do not surpass $5mg/m^3$ (Figure 3.1), but it is noteworthy that this pollutant is measured in mg/m^3 , therefore, its fraction in the atmosphere is considerably larger than the remaining pollutants. Even though this pollutant is only measured at urban sites, it is measured at urban locations with different influence, background and traffic, as shown in Figure 3.2(g). Overall mean values are higher for stations with a traffic influence than for stations of background influence. Detailed information regarding the mean and the standard deviation of each time series can be found in Table C.1.



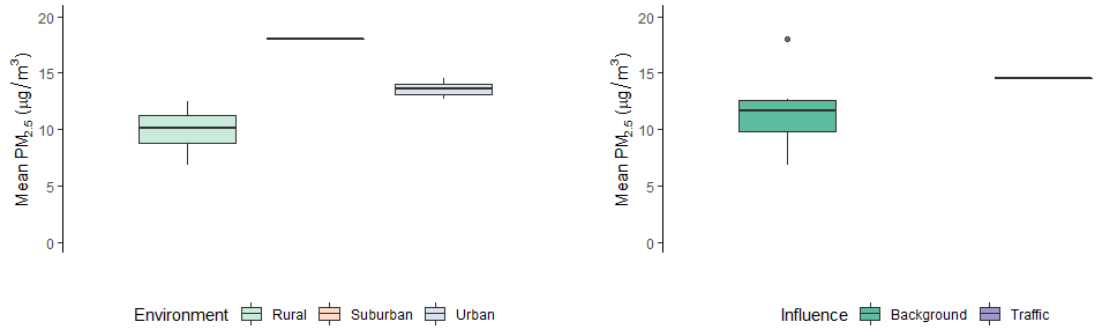
(a) NO_2



(b) NO_x



(c) PM_{10}



(d) $\text{PM}_{2.5}$

Figure 3.2: Distribution of time series mean according to type of background and environment of the 134 stations. (a) NO_2 , (b) NO_x , (c) PM_{10} , (d) $\text{PM}_{2.5}$, (e) O_3 , (f) SO_2 , (g) CO . (To be continued)

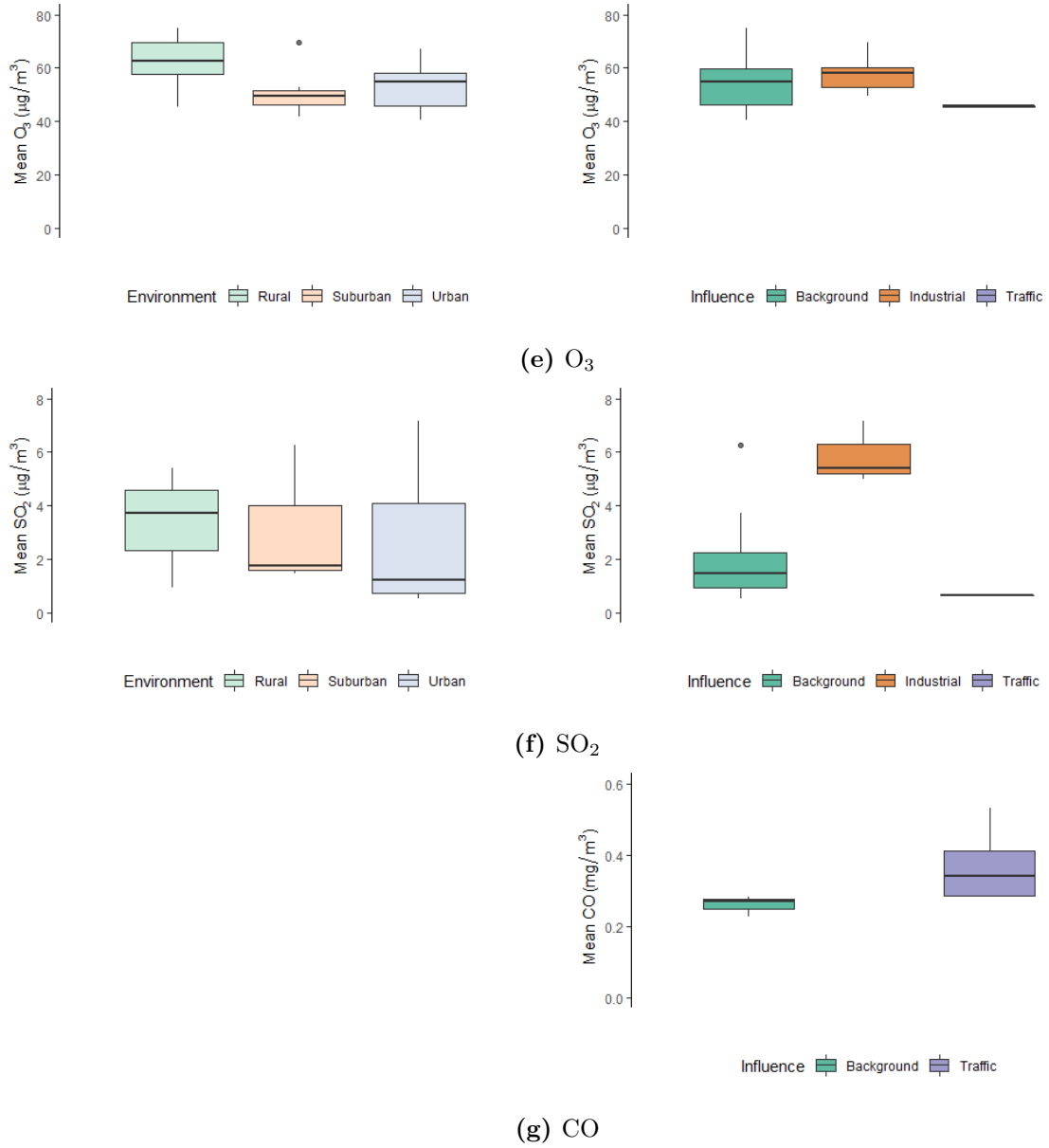


Figure 3.2: Distribution of time series mean according to type of background and environment of the 134 stations. (a) NO_2 , (b) NO_x , (c) PM_{10} , (d) $PM_{2.5}$, (e) O_3 , (f) SO_2 , (g) CO. (Continued)

3.1.2 Building a Modelling Framework

SARIMA

Figure 3.3(a) depicts the time series of CO at Aveiro station from 2005 up to 2015. It is clear that there is a pattern of annual seasonality and a decreasing trend from 2005 up to 2015. Furthermore, it is also possible to observe that there is daily seasonality in the series [Figure 3.3(b)]. Therefore, initially, it was decided to explore SARIMA models considering a 24-hour seasonality.

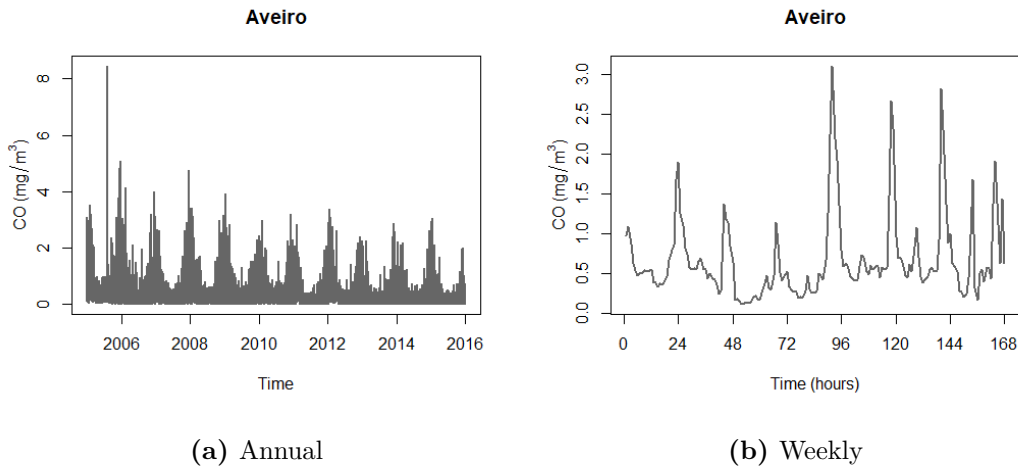


Figure 3.3: Time Series of CO at Aveiro station. (a) Complete series, (b) Weekly data from 1 January 2005 to 7 January 2005.

One of the assumptions for SARIMA analysis is that the time series must be stationary. Therefore, the first step is to observe the ACF and the PACF. From Figure 3.4(a) it is clear that the time series is not stationary; the ACF decreases slowly without reaching non-significant correlation values. Figure 3.4(b) shows the PACF and its values approach zero as the lags increase, but the partial correlation between observations persists even after a week. Therefore, it is necessary to transform the series in order to make it stationary.

Since the ACF of Aveiro station has a correlation peak at 24 hours, we conducted a seasonal differencing at lag 24, followed by a first differencing at lag = 1 to stabilise the mean. The ACF and PACF of the transformed time series can be found in the upper panel of Figure 3.5. Also, an additional transformation was performed to stabilise the variance. We performed a Box-Cox transformation ($\lambda = 0.0977$), followed by the seasonal differencing and the first order differencing (Figure 3.5 lower panel). Both transformations result in similar ACF and PACF, significant lags in the autocorrelation function are found only up to lag 24, whereas for PACF, despite the decreasing trend, significant correlations are found throughout the week.

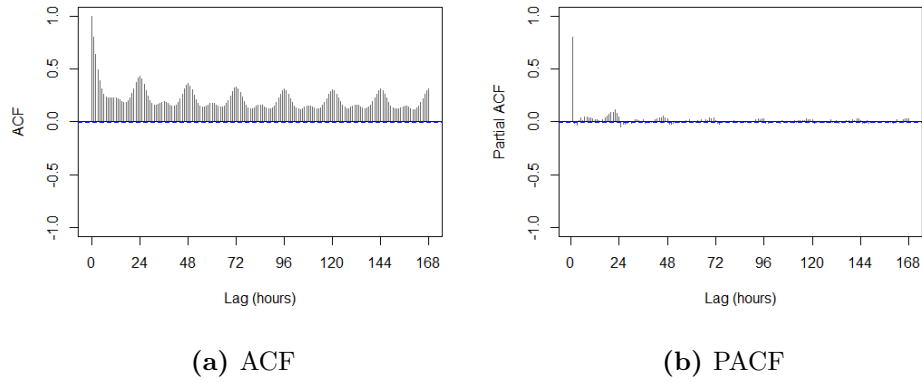


Figure 3.4: Characteristics of CO time series at Aveiro station. (a) ACF, (b) PACF.

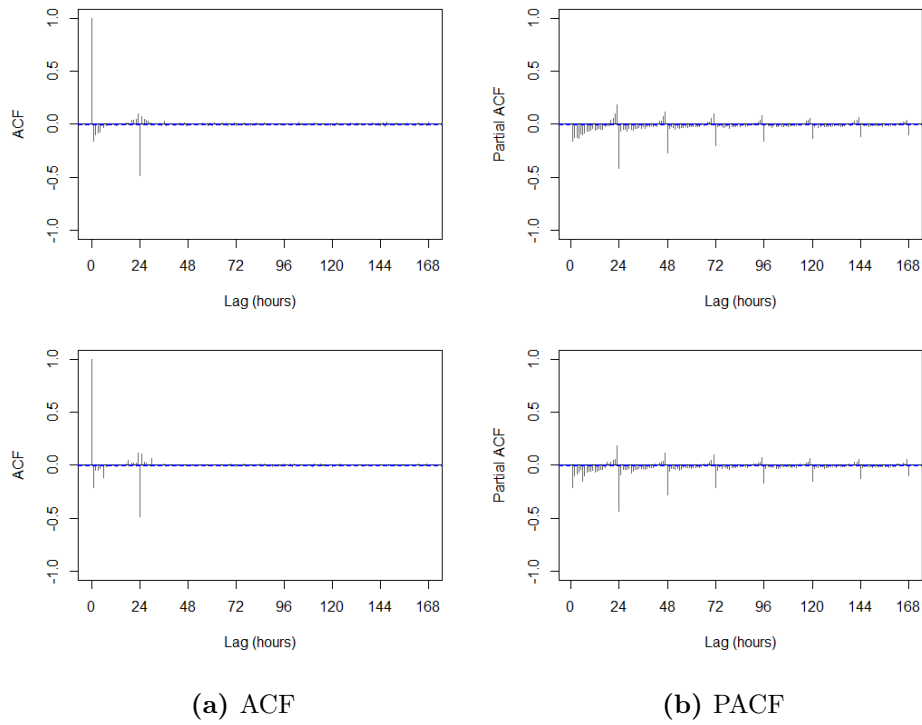


Figure 3.5: CO Aveiro station transformed time series. (a) ACF, (b) PACF. *Upper Panel* - Differenced time series, *Lower Panel* - Box-Cox transformation and differenced time series.

Since the transformations achieve a similar effect on the ACF and PACF we present results only for the differencing transformations, as the use of a Box-Cox transformation implies loss of result interpretability.

According to the algorithm described in Section 2.1.2, the optimal model selected for CO in Aveiro station is a SARIMA(0, 1, 2)(2, 1, 0)₂₄. The estimated parameters can be found in Table 3.1. The model obtained can be written as

$$(1 - \phi_1 B^{24} - \phi_2 B^{48})(1 - B)(1 - B^{24})y_t = (1 + \theta_1 B + \theta_2 B^2)\varepsilon_t,$$

where $(1 - B)(1 - B^{24}) = (1 - B^{24} - B + B^{25})$ and ε_t is a white noise with $\varepsilon_t \sim \mathcal{N}(0, 0.185)$. The model can be expanded to

$$(1 - \phi_1 B^{24} - \phi_2 B^{48})(1 - B^{24} - B + B^{25})y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}. \quad (3.1)$$

By replacing (Equation 3.1) with the estimated coefficients from Table 3.1 one can obtain the model equation.

CO Aveiro	θ_1	θ_2	Φ_1	Φ_2
Coefficients	-0.255	-0.217	-0.622	-0.308
Standard Error	0.003	0.004	0.003	0.003

Table 3.1: Coefficients of SARIMA model for CO time series at Aveiro station.

Figure 3.6 allows to analyse the residuals of the estimated model in order to conclude about its adequacy to the data. It is evident that residuals are correlated [Figure 3.6(a)], which contradicts the assumptions for SARIMA models errors. In addition, the ACF and the PACF [Figures. 3.6(b,c)] present significant correlations at several lags, even though there are less significant correlated lags than in the transformed data (Figure 3.5 *Upper Panel*). This means that there is information on the data that the model is not able to explain. Such results were expected since it was not possible to achieve stationarity with the time series transformation. Similar results were achieved for the remaining six pollutants and 133 time series.

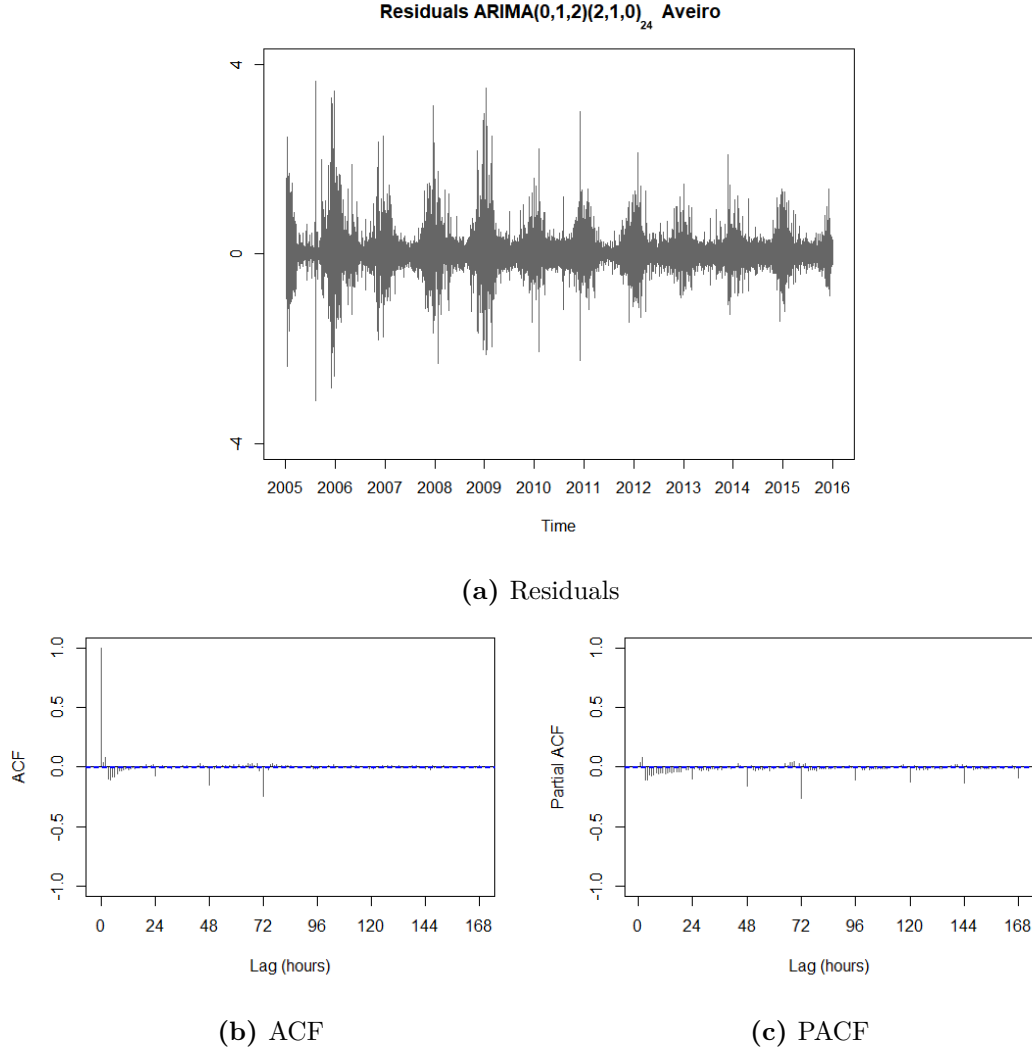


Figure 3.6: Residuals analysis of SARIMA model. (a) distribution, (b) ACF, (c) PACF.

Model estimation for one time series considering $\max.p = \max.q = \max.P = \max.Q = 2$, require a computation time of about 32 minutes, which is reasonable. However, further expanding the models search space, by increasing the parameters orders up to 5, as there is room for model improvement, exponentially increases computation time, taking about 2 days 19 hours and 17 minutes just to compute the model for only one time series. As we intend to develop a flexible framework to estimate models for the 134 time series, the use of SARIMA models is unfeasible. Also, SARIMA implementation does not allow to deal with multiple seasonality. Taking into account that the time series have, at least, annual seasonality, besides daily seasonality, it is not possible to accommodate both using SARIMA models [Figure 3.3(a)]. Furthermore, the ACF suggests that the time series may have long-

range dependence [Figure 3.4(a)], which cannot be described by SARIMA models. Therefore, it is necessary to consider other models such as dynamic harmonic regression and ARFIMA models presented below.

Dynamic Harmonic Regression Models

Unlike SARIMA, dynamic harmonic regression is flexible to include multiple seasonality. We decided to include daily, weekly and annual seasonality, as these were the most relevant seasonalities in most time series. The first step in DHR is to determine the order of the Fourier terms for each seasonality. As previously mentioned, the order of the Fourier terms is at most half of the length of the seasonality. Considering that 1 day has 24 hours, 1 week has 168 hours and 1 year has 8766 hours the maximum order of the Fourier terms is, respectively $k_d = 12$, $k_w = 84$ and $k_y = 4383$. Therefore, the search space is considerably large. In addition, the parameters p , q and d need to be determined, which further increases the search space. In order to limit the search space of the dynamic harmonic regression models, the following framework was established:

1. Find the best ARIMA model to describe the data,
2. Use the previous model to find the best value of k_d ,
3. Use the previous model and best k_d to find the best k_w ,
4. Use the previous model, best k_d and best k_w to find the best k_y ,
5. Re-estimate the ARIMA model using the best k_d , k_w and k_y .

The selection of the best ARIMA model and the order of the Fourier terms was performed based on AIC and BIC criteria. The best k_d , k_w and k_y were found when AIC and BIC stopped decreasing.

Table 3.2 shows the best ARIMA model and the best Fourier terms estimated. The computation of the best ARIMA model takes just a few minutes, while the computation of the Fourier terms is substantially slower. The selection of k_d took about an hour and a half, while the computation of k_y took nearly 19 hours. Regarding the selection of k_w , this was the longest, taking approximately 3 days, 9 hours and 20 minutes.

CO Aveiro	Parameter Values
Arima Model	(5, 1, 4)
K_d	11
K_w	37
K_y	1

Table 3.2: Best ARIMA model and Fourier terms obtained by the established framework.

Lastly, the ARIMA model was re-estimated using the order of the Fourier terms identified in Table 3.2. The best ARIMA model was the (0,1,4) with the Fourier terms $K_d = 11$, $K_w = 37$ and $K_y = 1$. The estimated coefficients of the model can be found in Table 3.3. The model is quite complex since it has many coefficients. Nevertheless, it is noteworthy that the magnitude of the θ_1 and θ_2 is similar to the magnitude of the moving average coefficients estimated by the SARIMA model (Table 3.1).

Similarly to the results found for SARIMA, the residuals are correlated, since its distribution clearly does not resembles the distribution of a white noise [Figure 3.7(a)]. Furthermore, ACF and PACF show significantly correlated lags with a cyclic pattern [Figure 3.7(b,c)]. Nevertheless, this model seems to properly describe the time series of carbon monoxide in Aveiro, since the periodogram of the fitted values is quite similar to the periodogram of the original time series (Figure 3.8). However, the model complexity (over 60 coefficients) leads to a great computational effort, requiring about one day and a half to re-estimate the model. Therefore, even if we consider the order of Fourier terms in Table 3.2 for all time series of air pollutants, it would be necessary about 200 days to estimate all models, which deems the use of dynamic harmonic regression models unfeasible.

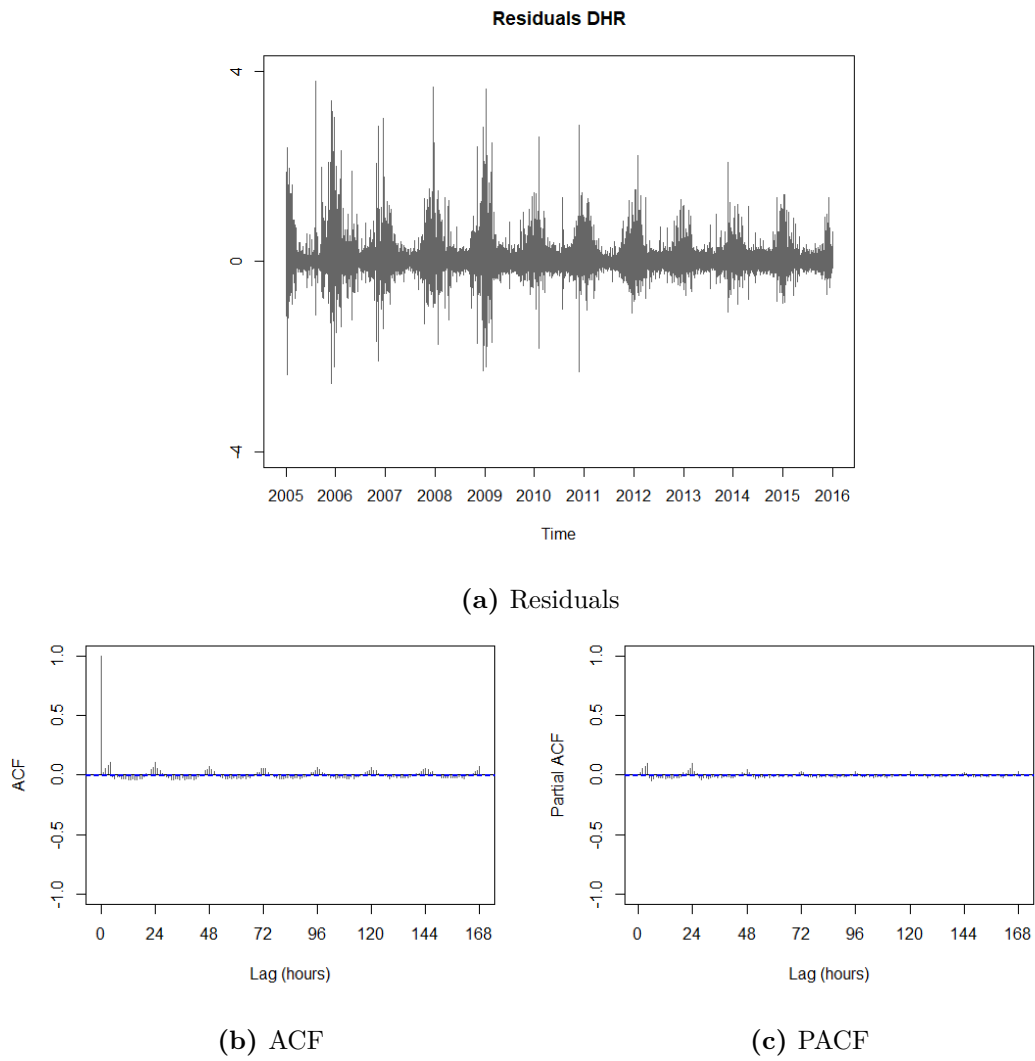


Figure 3.7: Analysis of CO DHR model residuals at Aveiro station. (a) distribution, (b) ACF, (c) PACF.

Coefficients		Coefficients		Coefficients		Coefficients		Coefficients	
θ_1	-0.245	S9-24	0.000	S9-168	0.005	S20-168	0.002	S32-168	0.000
θ_2	-0.215	C9-24	0.000	C9-168	-0.003	C20-168	-0.003	C32-168	0.000
θ_3	-0.232	S10-24	-0.003	S10-168	0.001	S22-168	-0.005	S33-168	0.001
θ_4	-0.209	C10-24	0.002	C10-168	-0.003	C22-168	-0.001	C33-168	0.002
S1-24	-0.026	S11-24	0.000	S11-168	-0.002	S23-168	0.001	S34-168	0.000
C1-24	0.036	C11-24	0.000	C11-168	0.000	C23-168	0.002	C34-168	0.002
S2-24	-0.057	S1-168	-0.006	S12-168	-0.004	S24-168	0.001	S36-168	0.000
C2-24	0.028	C1-168	0.003	C12-168	0.003	C24-168	0.001	C36-168	0.000
S3-24	0.006	S2-168	-0.007	S13-168	0.003	S25-168	0.000	S37-168	-0.001
C3-24	0.006	C2-168	0.007	C13-168	0.008	C25-168	-0.002	C37-68	0.000
S4-24	0.001	S3-168	0.003	S15-168	0.004	S26-168	-0.001	S1-8766	-0.003
C4-24	-0.012	C3-168	0.005	C15-168	0.001	C26-168	-0.002	C1-8766	0.119
S5-24	-0.007	S4-168	-0.001	S16-168	0.004	S27-168	-0.002		
C5-24	0.006	C4-168	-0.004	C16-168	-0.004	C27-168	0.000		
S6-24	0.000	S5-168	-0.009	S17-168	-0.001	S29-168	0.002		
C6-24	0.003	C5-168	0.003	C17-168	-0.001	C29-168	0.002		
S7-24	-0.001	S6-168	-0.002	S18-168	0.000	S30-168	0.002		
C7-24	-0.003	C6-168	0.010	C18-168	0.001	C30-168	0.002		
S8-24	-0.004	S8-168	0.006	S19-168	0.002	S31-168	0.002		
C8-24	0.004	C8-168	0.006	C19-168	-0.003	C31-168	-0.001		

Table 3.3: DHR coefficients for CO time series at Aveiro station.

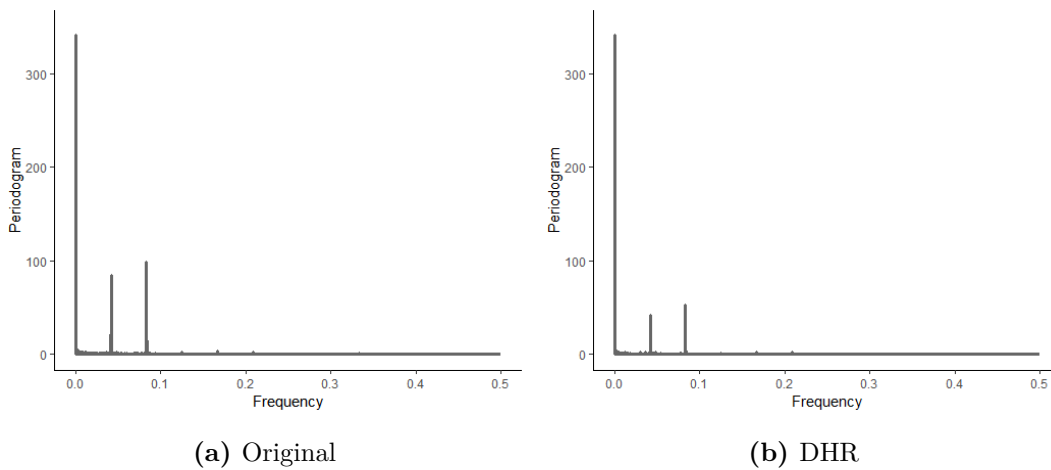


Figure 3.8: Periodogram of CO time series at Aveiro station. (a) original time series, (b) DHR fitted time series.

ARFIMA models

An alternative to Dynamic Harmonic Regression are the ARFIMA models, which allow to describe the time series long memory. The ACF of Aveiro station for CO [Figure 3.4(a)] suggests the existence of long-range dependence since there are significant correlations up to lag 168. However, in order to verify if the time series has long-range dependence (i.e., long memory) we can compute the Hurst Exponent.

Table 3.4 shows the Hurst Exponent computed through the R/S method and the adjusted R/S method, for Aveiro station for CO pollutant. For all methods, the Hurst Exponent is higher than 0.5, which indicates that the time series has a long-term positive autocorrelation, i.e., long-range dependence. Table C.2 shows the Hurst Exponent for NO₂, NO_x, PM₁₀, PM_{2.5}, O₃, SO₂ and CO for all time series. Values of the Hurst Exponent are similar to the presented, thus one can conclude that all time series have long term memory.

Station	R/S Method	Corrected R/S method
Aveiro	0.81	0.84

Table 3.4: Hurst Exponent for Aveiro station, CO air pollutant.

Table 3.5 shows the ARFIMA coefficients and the respective standard error for Aveiro station CO pollutant. The ARFIMA obtained is a (3, 0.499, 2) model. As mentioned in the methods section, the implementation of function *arfima* in R sometimes does not allow for the computation of the standard error, unless changes to the h parameter are done (finite-difference interval for approximating partial derivatives with respect to the d parameter). The standard error of the fractional difference parameter (d) is clearly smaller than the remaining, which raises questions regarding its reliability in computing the standard error of d parameter, as mentioned previously. From (Equation 2.46) we can write the model as

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3) \nabla^{0.499} y_t = (1 + \theta_1 B + \theta_2 B^2) \varepsilon_t, \quad (3.2)$$

CO Aveiro	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2
Coefficients	0.499	1.297	-0.363	-0.055	1.015	-0.095
Standard Error*	1.04×10^{-6}	0.029	0.032	0.039	0.032	0.011

*Standard error is computed for $h \times 0.01 = 1.221 \times 10^{-4}$.

Table 3.5: Coefficients of ARFIMA model for CO time series at Aveiro station.

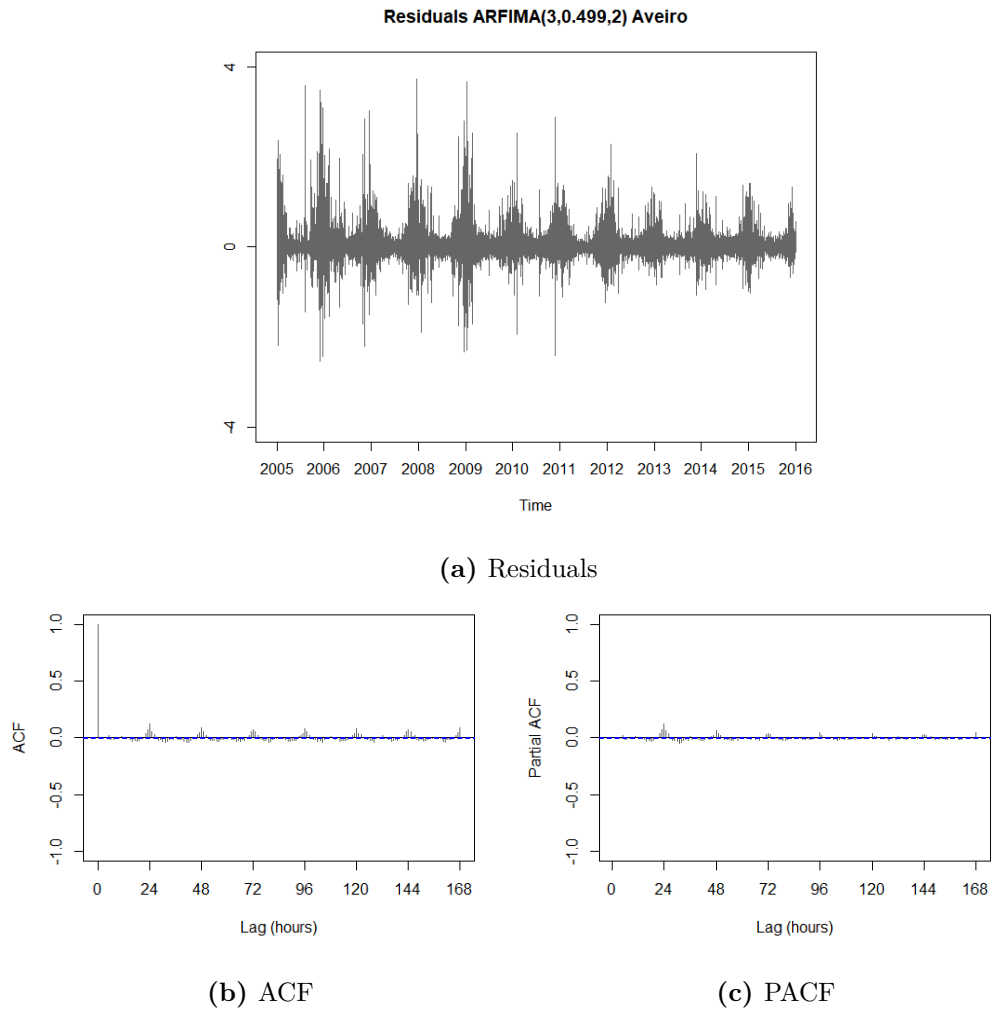


Figure 3.9: Analysis of CO ARFIMA model residuals at Aveiro station.(a) distribution, (b) ACF, (c) PACF.

Figure 3.9 shows the analysis of the residuals of the Aveiro time series for CO pollutant. The residuals do not follow a random pattern, thus, these do not approach the distribution of a white noise. Regarding the ACF, when compared to Figure 3.4 (ACF of the original time series), the significantly correlated lags decrease considerable, but a significant seasonal pattern remains. PACF also seems to improve slightly compared to the original time series, but significant lags persist.

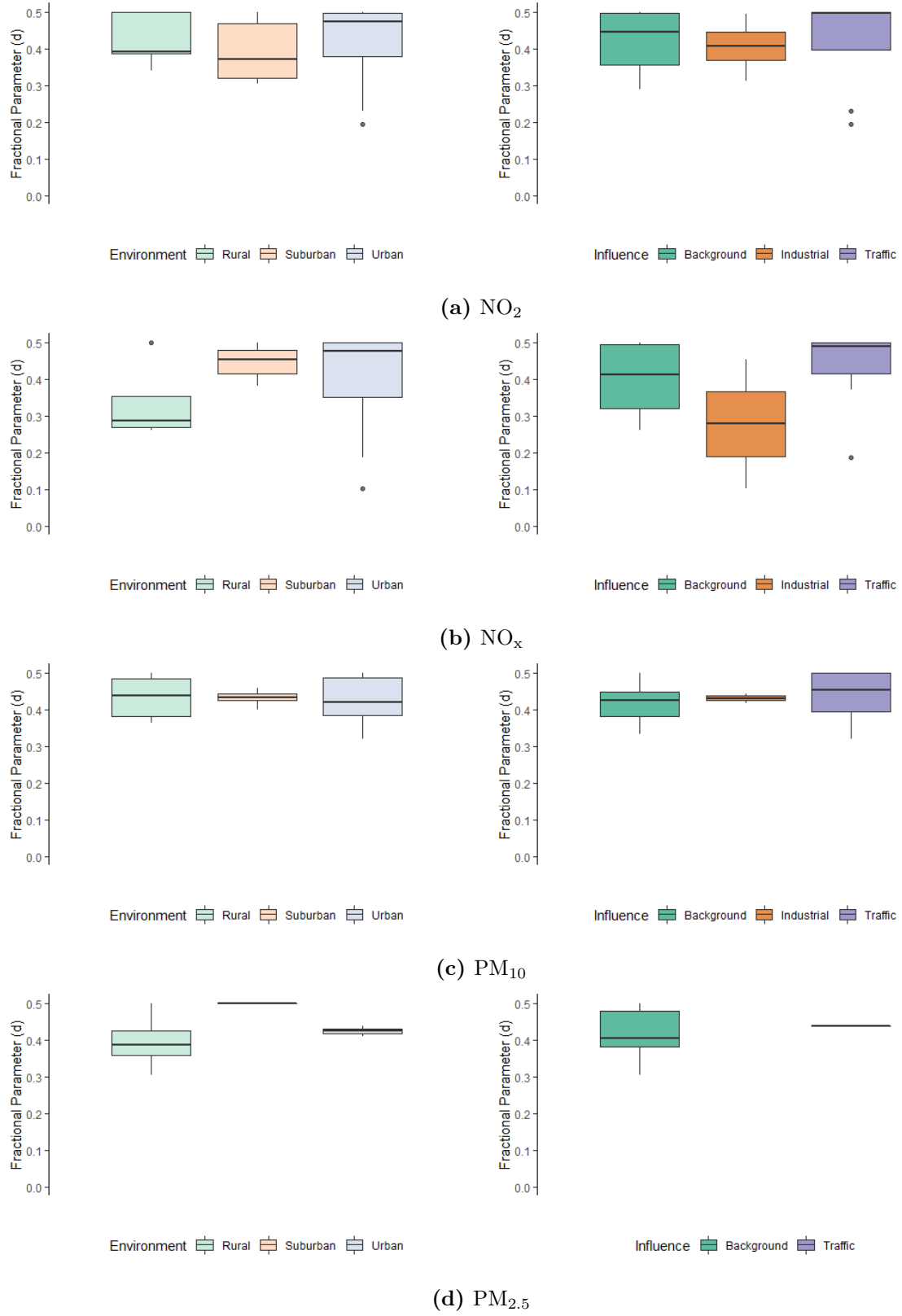


Figure 3.10: Distribution of d parameter for all time series according to type of environment and influence. (a) NO₂, (b) NO_x, (c) PM₁₀, (d) PM_{2.5}, (e) O₃, (f) SO₂, (g) CO. (To be continued)

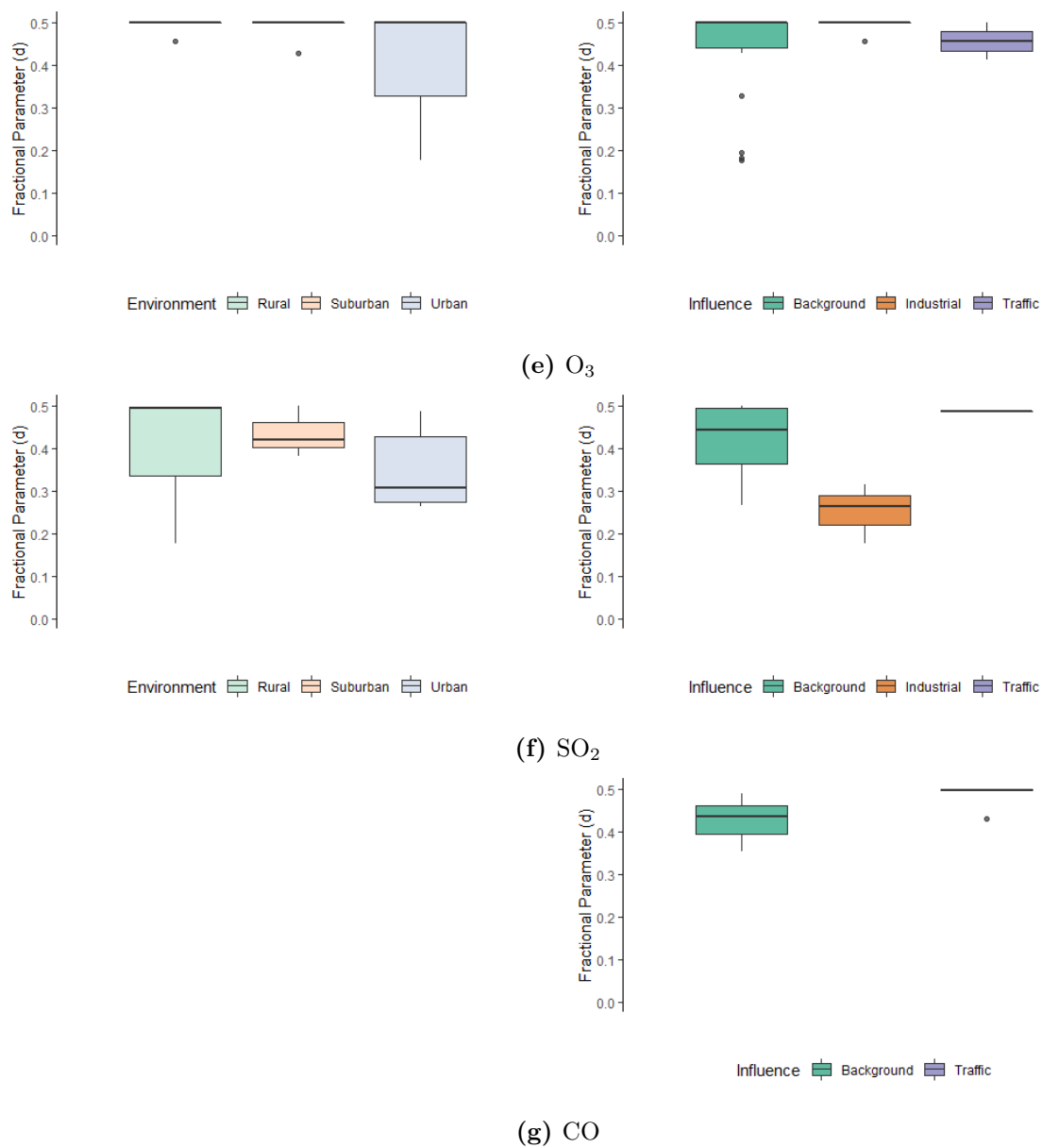


Figure 3.10: Distribution of d parameter for all time series according to type of environment and influence. (a) NO₂, (b) NO_x, (c) PM₁₀, (d) PM_{2.5}, (e) O₃, (f) SO₂, (g) CO. (Continued)

Figure 3.10 shows the distribution of the d parameter for all time series according to their type of environment and influence. Median d is similar for rural and suburban NO_2 time series and slightly higher for urban areas. Regarding the type of influence, median d values are similar for all influences of NO_2 time series [Figure 3.10(a)]. With respect to NO_x , median d values are smaller for rural areas, and similar for suburban and urban stations. When analysing the distribution of d according to type of influence for NO_x stations, its median values are higher for background stations and traffic-influenced stations [Figure 3.10(b)]. For particulate matter air pollutants and O_3 , the median d values are nearly 0.5 and are similar regardless of type of environment and influence [Figure 3.10(c,d,e)]. SO_2 presents higher median d values for rural and suburban stations, while stations from a background and traffic influence have higher median d than industrial-influenced time series. Finally, CO has slightly higher d median value for traffic-influenced stations than background stations. Nonetheless, no common pattern is found for all pollutants.

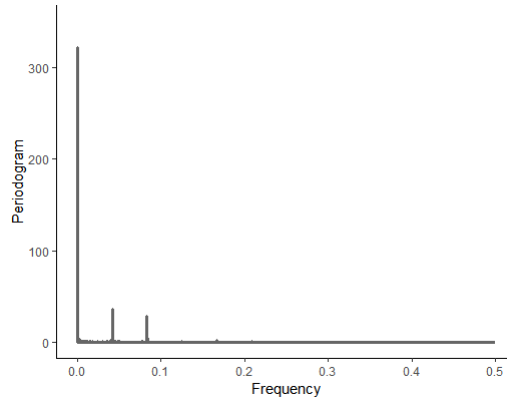


Figure 3.11: Periodogram of ARFIMA fitted time series of CO at Aveiro station.

The periodogram of the fitted results of ARFIMA models seems to describe a bit less information for the frequencies corresponding to the 24h (0.04) and the 12h (0.08), than the DHR models (Figure 3.11). However, its computation time is quite faster, requiring only about 20 minutes. Thus, these models are preferred to the DHR model. Since the autocorrelation function, the partial autocorrelation function and the periodogram suggest that there is daily seasonality left to explain in the residuals, we decided to combine SARIMA with ARFIMA models - SARFIMA models - to further improve the models' ability to describe the data.

SARFIMA models

Since ARFIMA models have long-range dependence, and the autocorrelation function of ARFIMA residuals' presents some significant lags resembling daily seasonality, SARIMA and ARFIMA models were combined in order to include the short term memory in an attempt to further improve the models. Since the goal is to describe the short term memory, that is the daily seasonality, the seasonal part of the model can be restricted. The autoregressive and the moving average orders of the model can be at most one ($P.max = Q.max = 1$) and $D = 0$. Further increasing the order of P and Q or considering $D \neq 0$ may remove important information of the long term memory handled by ARFIMA. Therefore, the SARFIMA models considered were:

- SARFIMA(p, d, q)(0, 0, 0)₂₄
- SARFIMA(p, d, q)(1, 0, 0)₂₄
- SARFIMA(p, d, q)(0, 0, 1)₂₄
- SARFIMA(p, d, q)(1, 0, 1)₂₄

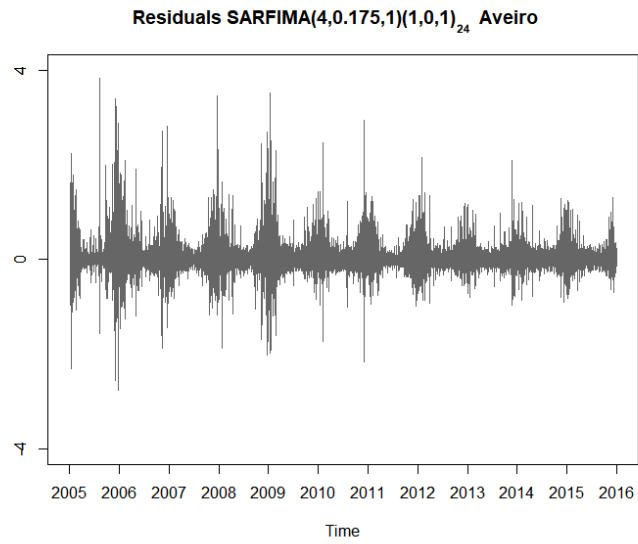
For all stations of each pollutant, the best SARFIMA model selected, based on the AIC and BIC criteria, was the SARFIMA (p, d, q)(1, 0, 1)₂₄ model, which is not surprising since, of all models considered, this is the one that allows retaining the most information regarding the daily variability.

Table 3.6 shows the coefficients of the SARFIMA model for CO pollutant at Aveiro station. The time series is described by a SARFIMA(4, 0.175, 1)(1, 0, 1)₂₄ model. Comparatively to the results obtained to the ARFIMA model (Table 3.5), the value of the parameter d is smaller (0.175 vs. 0.499), which is expected since the daily seasonality is described by the (1, 0, 1)₂₄ seasonal component of the model. The non-seasonal autoregressive terms and moving average terms are also smaller in the SARFIMA model than in the ARFIMA, as a result of including the daily seasonal component. The model can be written as

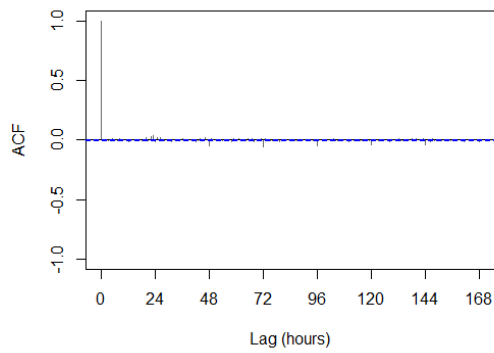
$$(1 - \Phi_1 B^0)y_t = c + (1 + \Theta_1 B^0) + z_t, \quad (3.3)$$

where, $c = \bar{x} \times (1 - \Phi_1) = 0.289 \times (1 - 0.971) = 0.008$ and z_t is a correlated error with an ARFIMA(4, 0.175, 1) distribution

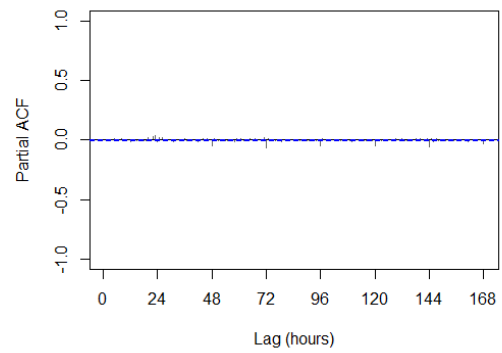
$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4) \nabla^{0.175} z_t = (1 + \theta_1 B) \varepsilon_t. \quad (3.4)$$



(a) Residuals



(b) ACF



(c) PACF

Figure 3.12: Analysis of CO SARFIMA model residuals at Aveiro station. (a) distribution, (b) ACF, (c) PACF.

CO Aveiro	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	θ_1	Φ_1	Θ_1	\bar{x}
Coefficients	0.175	0.062	0.302	-0.002	-0.029	-0.485	0.971	-0.797	0.289
Standard Error*	2.15×10^{-7}	0.004	0.004	0.036	0.066	0.066	0.002	0.005	0.004

* Standard error is computed for $h \times 0.01 = 1.221 \times 10^{-4}$ for coefficients d, ϕ_1 , ϕ_2 , ϕ_3 , ϕ_4 and θ_2 .

Table 3.6: Coefficients of SARFIMA model CO time series at Aveiro station.

Residuals of the SARFIMA model are clearly not similar to a white noise since a distinct pattern can be observed [Figure 3.12(a)]. Therefore, residuals are correlated, which means that the model does not account for all variability of the data. Nevertheless, the ACF and PACF of the SARFIMA residuals [Figure 3.12(b,c)] improved comparatively to the ACF and PACF of the ARFIMA residuals [Figure 3.9(b,c)], and significant lags are found only at lags multiples of 24. Even though the correlations found at these lags are still significant, its absolute value is considerably less than the correlation of the original ACF and PACF (Figure 3.4). Furthermore, if we compare the periodogram of SARFIMA fitted values (Figure 3.13) to the observed time series [Figure 3.8(a)], one can see that these are quite similar. Hence, of all studied models, this is the one that better describes the data. Tables C.3 to C.9 display the models' coefficients for all pollutants at each station. It is noteworthy that some models have $d = 0.000$. This does not mean that d is in fact zero, but that its value is pretty small so that its effect is almost negligible and the error distribution (z_t) of the model approaches an ARMA(p, q).

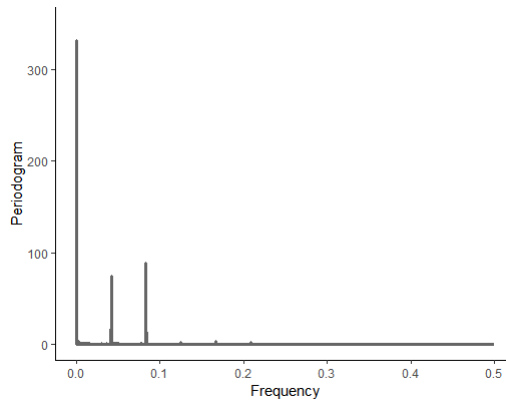


Figure 3.13: Periodogram of SARFIMA fitted values of CO at Aveiro station.

SARFIMA models allow a flexible approach to describe the air pollutants time series. Furthermore, the computation time of the model is about 25 minutes, which is considerably less than the computation time of SARIMA and DHR models.

3.1.3 Model Forecast

As the approach to obtain the models has been established, one can perform forecast of the time series. The forecast of the Aveiro time series for the CO pollutant is presented in Figure 3.14. The forecasts were computed for the year 2016 using bootstrap since the model residuals are correlated. Forecasts start to converge to the mean of the process around April, which means that in the remaining of the year forecasts are fairly constant.

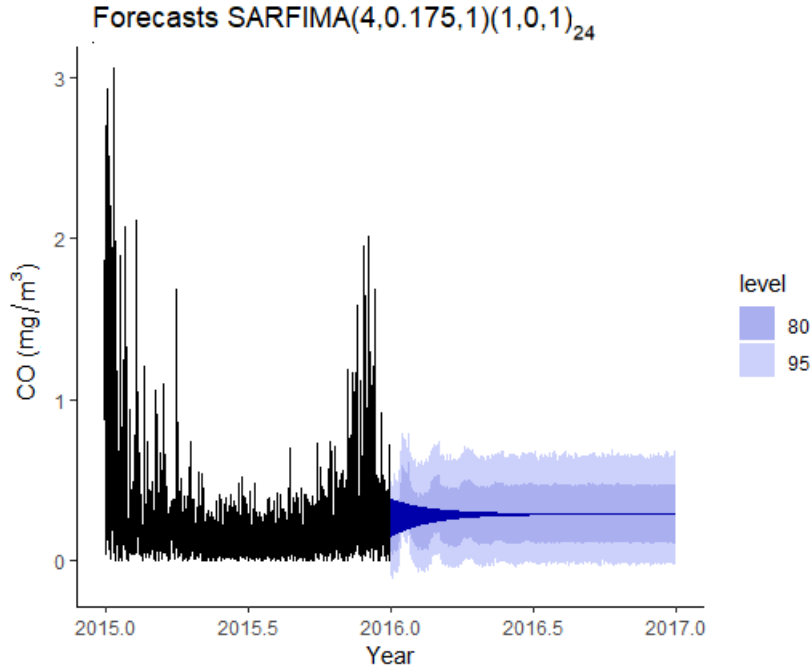


Figure 3.14: Forecasts of SARFIMA model for Aveiro station, CO pollutant.

Table 3.7 shows the performance measures of the model. The ME, RMSE and MAE are scale dependent. Hence, these measures can be compared to the mean and the standard deviation of the time series of 2016 (test set) so that the magnitude of the error is assessed. The mean and the standard deviation are 0.205 ± 0.193 . Therefore, these errors, are at most, approximately one standard deviation. The MPE and the MAPE cannot be computed for this station. These measures are infinite or undefined if $x_t = 0$ for any t in the time series, and having an extremely skewed distribution when any value of x_t is close to zero. Thus, these measures are not useful to asses the accuracy of these time series, since there are observations $x_t = 0$, which can lead to infinite or undefined measures or even substantially large MPE or MAPE values. Lastly, since the MASE is 0.542, which is less than one, the forecast is better than the average one-step naive forecast computed in-sample. Overall, it seems that the model has good performance. Furthermore, the prediction interval at 95% includes 96.6%

of the true observations. One can note that the inferior 95% prediction interval at some points is negative, however, negative values must be disregarded since the air pollution values are non-negative.

	ME	RMSE	MAE	MAP	MAPE	MASE	PI 80%	PI 95%
Aveiro	-0.082	0.208	0.158			0.542	59.1	96.6

Table 3.7: Performance measures and percentage of observations within the prediction intervals for CO at Aveiro station.

Figure 3.15 shows the distribution of MASE for each pollutant according to the type of environment and influence. O_3 has the smallest median MASE values [Figure 3.15(e)], whereas $PM_{2.5}$ has the highest [Figure 3.15(d)]. Hence, for O_3 stations the forecast are better than the one-step naive forecast. On the contrary, for $PM_{2.5}$ the forecasts are worse than the one-step naive forecast. The median MASE values are fairly similar for NO_x , NO_2 and PM_{10} (≈ 0.5) [Figure 3.15(a,b,c)]. SO_2 has median MASE values quite close to one, which suggests that its forecasts are as good as the one-step naive forecast. No clear pattern of performance is observed when considering MASE values stratified by environment or influence, with exception of CO. All CO stations are from an urban environment, but its influence can be either background or traffic. Median MASE value is higher for traffic influence than for background influence, which suggests that the forecast might perform better for CO background stations.

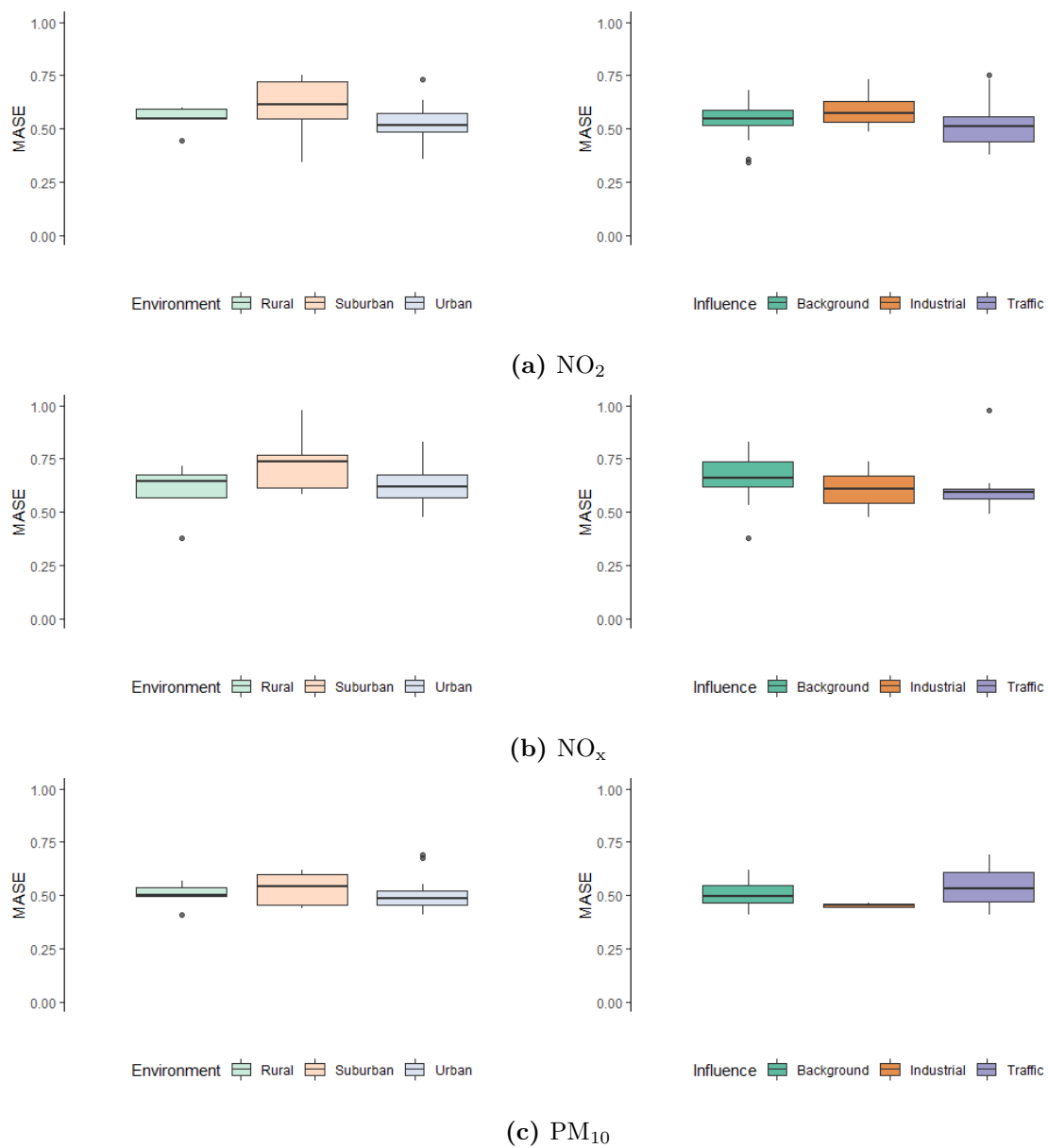


Figure 3.15: MASE Distribution according to type of environment and influence. (a) NO_2 , (b) NO_x , (c) PM_{10} , (d) $\text{PM}_{2.5}$, (e) O_3 , (f) SO_2 , (g) CO . (To be continued)

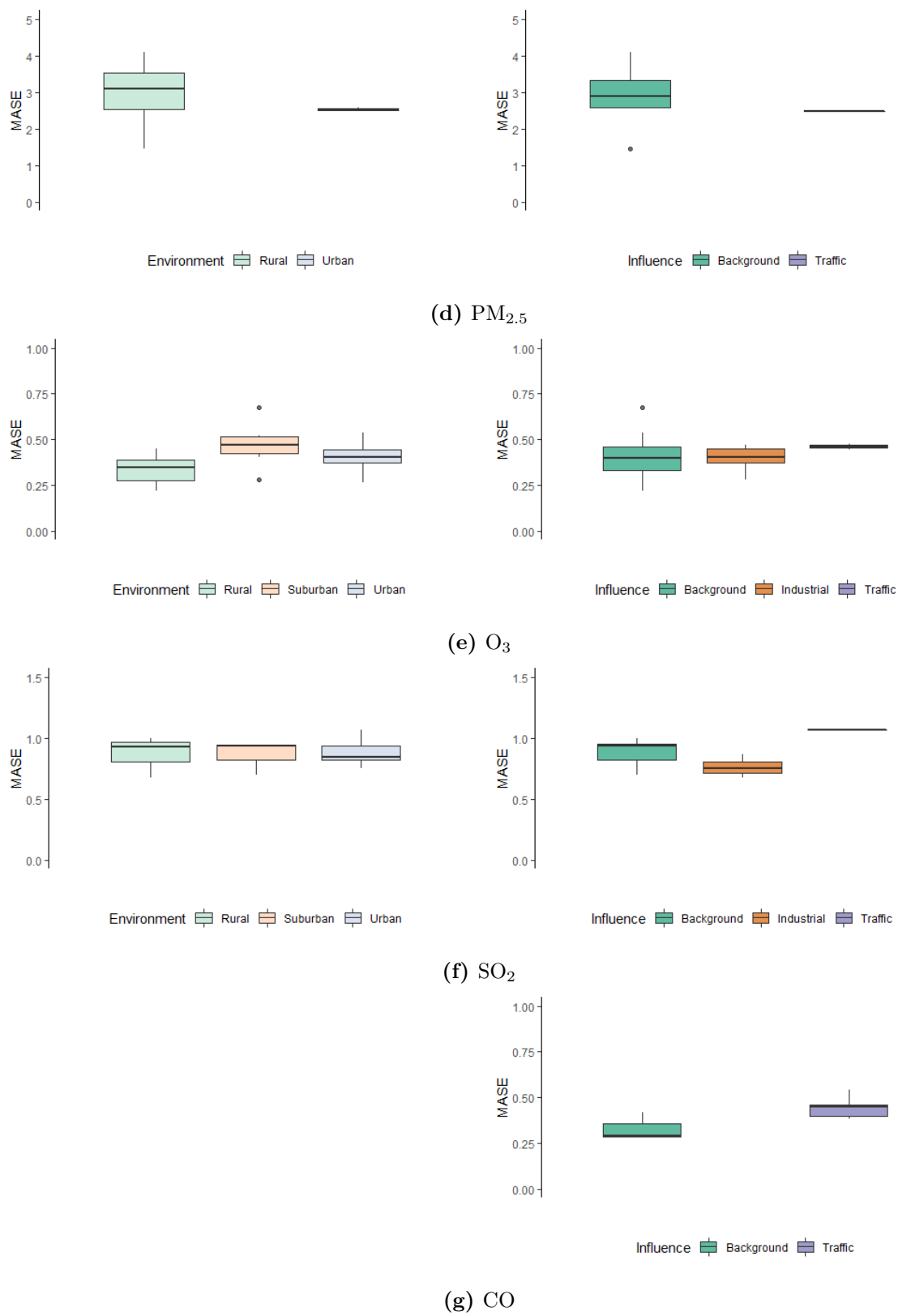


Figure 3.15: MASE Distribution according to type of environment and influence. (a) NO_2 , (b) NO_x , (c) PM_{10} , (d) $\text{PM}_{2.5}$, (e) O_3 , (f) SO_2 , (g) CO . (Continued)

In Appendix C, detailed results for all pollutants considering all performance measures can be found. Tables C.10 up to C.16 present the performance measures for all time series for each air pollutant as well as the percentage of true observations within the prediction intervals. In general, all pollutants present fair to good performance measures, in accordance to MASE results. Mean error, root mean squared error and mean absolute error are less than the time series standard deviation. Exceptions are the Francisco Sá Carneiro Station (NO_2 and NO_x), Sonega Station (NO_x and O_3), Fornelo do Monte Station (PM_{10}), Monte Chãos Station (O_3) and Custóias Station (NO_x), where the RMSE is slightly higher than the time series standard deviation. For SO_2 , CO and NO_x the prediction intervals of most time series at 95% includes, includes nearly 95% of true observations. NO_2 has 60% of the 95%PIs containing at least 90% of true observations. Regarding PM_{10} few 95%PIs include 95% of true observations. Nevertheless, half of 95%PIs include at least 75% of true observations. For $\text{PM}_{2.5}$, the 95%PI of Chamusca is the only that does not contain 95% of true observations. As for ozone, the 95% PIs, with some exceptions, contain at most 75% of observations. It is also worth-note that the 80% PIs perform poorly than the 95% PI in all cases, i.e., while the 95% intervals can contain 95% or more of the true observations, the 80% intervals, in general, contain less than 80% of the true observations. In conclusion, the SARFIMA models describe and predict the time series of the studied air pollutants fairly well, without depending on their surrounding environment and influence.

3.2 Hospital Admissions in Aveiro

3.2.1 Descriptive Results

Hospital Admissions

Figure 3.16 shows the distribution of daily hospital admissions per ICD-9 code (see Table 1.4 for full description of the codes). The median number of hospital admissions and the associated variability is fairly similar throughout the years. The code 460 (acute nasopharyngitis), is responsible for the highest number of hospital admissions, while the remaining have a median number of daily hospital admissions below 5.

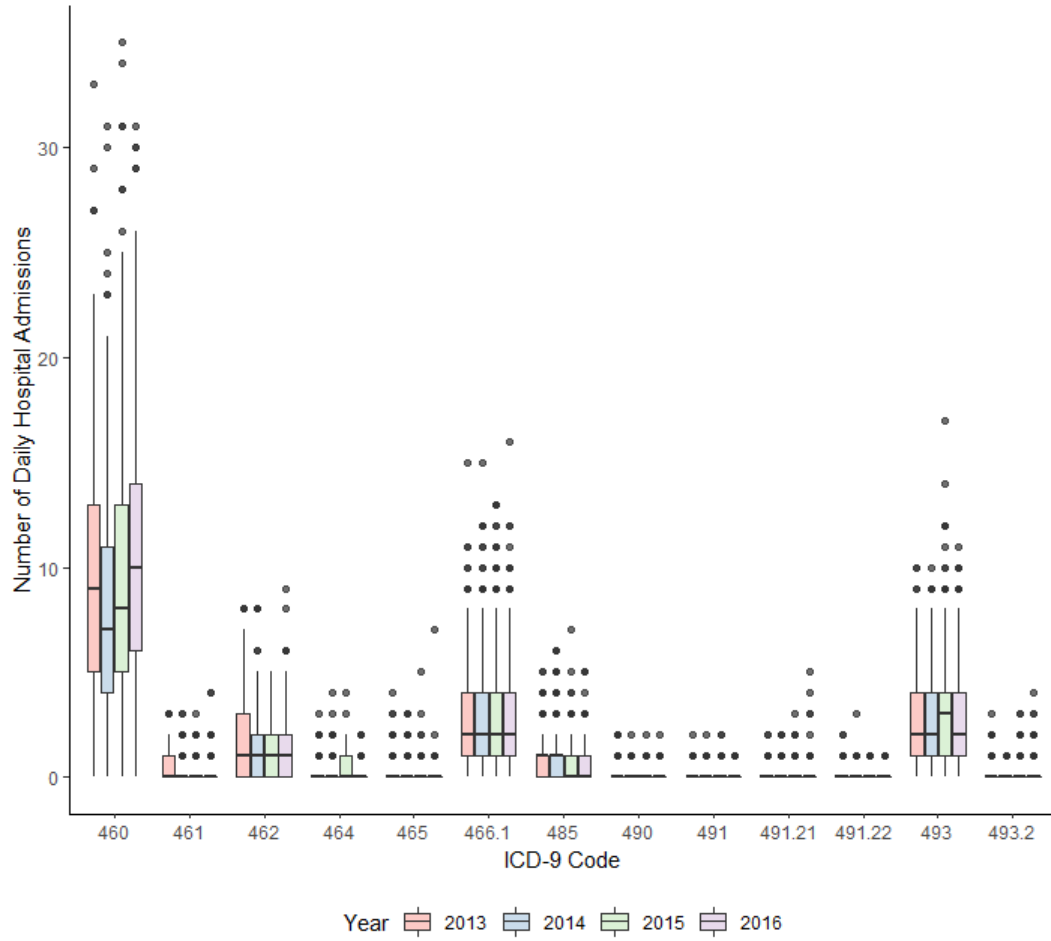


Figure 3.16: Annual distributions of the daily number of hospital admissions (2013-2016) for each ICD-9 code (Table 1.4)

Therefore, we decided to combine all codes to study respiratory hospital admissions, otherwise, most codes would have too few observations. The resulting time series is presented in Figure 3.17, in which an annual seasonality of the data is clearly depicted.

Figure 3.18(a) displays the ACF of daily hospital admission up to lag 30, showing significant correlations at all lags. Additionally, a peak at lags multiples of seven is observed, which suggests that there is weekly seasonality in this count time series. Furthermore, the PACF also supports the presence of weekly seasonality since, after lag 7 there are significant correlations at multiples 14, 21, 28, which indicates that there is a significant correlation between these lags even if the correlations of the lags in-between are removed.

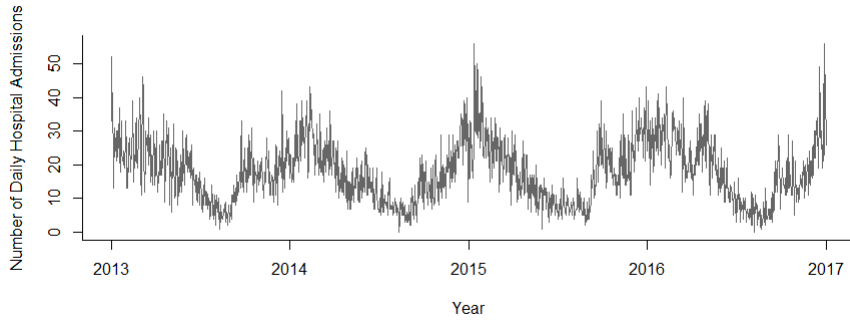


Figure 3.17: Time series of daily hospital admissions (all ICD-9 codes combined).

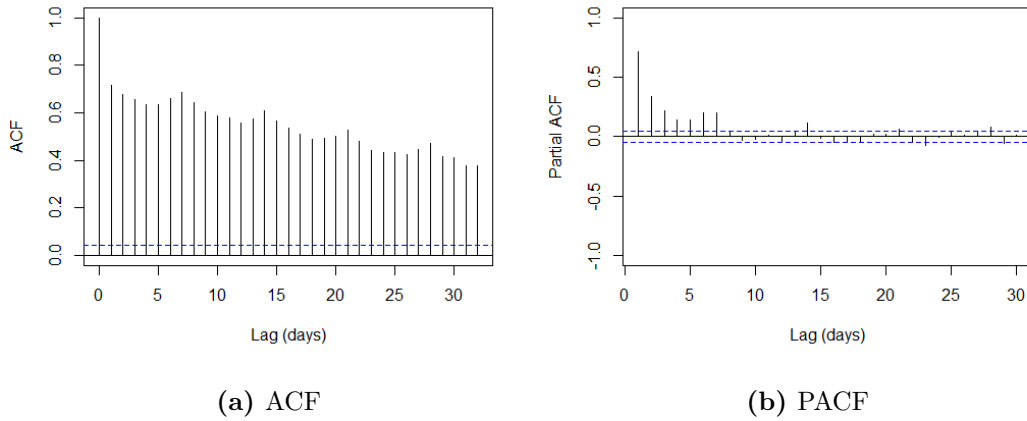
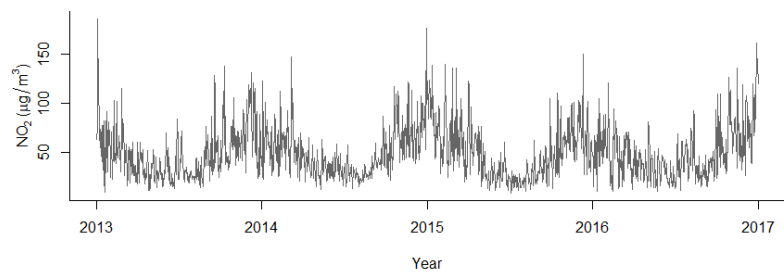


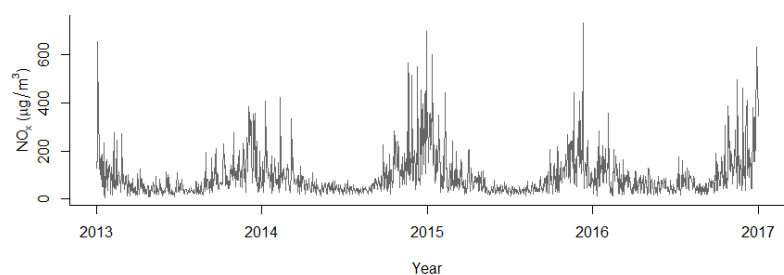
Figure 3.18: Daily hospital admission. (a) ACF, (b) PACF.

Hospital Admissions and Air Pollutants

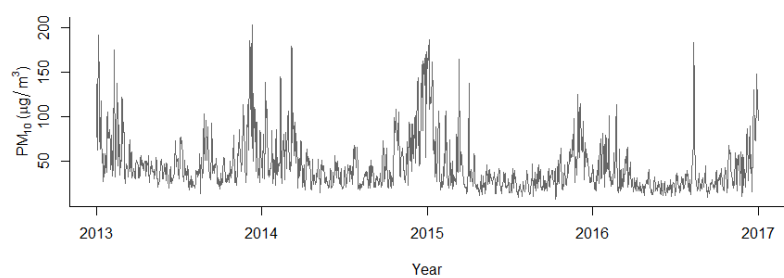
Air pollutants time series are hourly sampled, hence these time series were transformed into daily time series. To obtain a daily time series, the maximum daily value of pollutants were considered as the daily observation. It is noteworthy, that not all pollutants have information for Aveiro. Nevertheless, they have information for Estarreja and Ílhavo, which are the closest air quality stations to Aveiro. Estarreja time series were selected when information for Aveiro was not available since an overall higher cross-correlation was found between this station and daily hospital admissions than for Ílhavo, adding to the fact that both stations exhibit series with similar mean and standard deviation (Table C.1). Figure 3.19 presents time series of daily maximum air pollutants, showing an annual seasonal pattern for all pollutants, except for SO_2 , for which seasonality is less clear [Figure 3.19(f)].



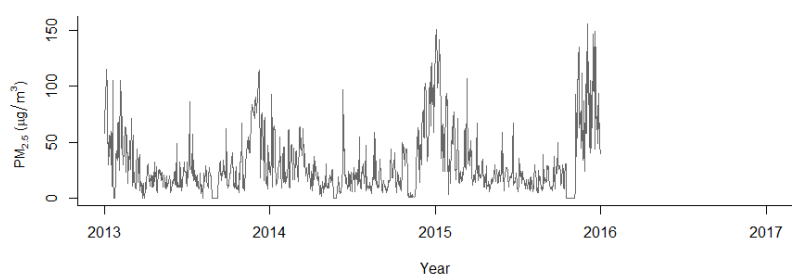
(a) NO₂



(b) NO_x

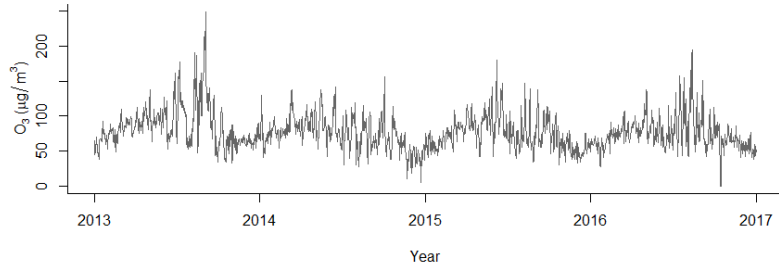


(c) PM₁₀

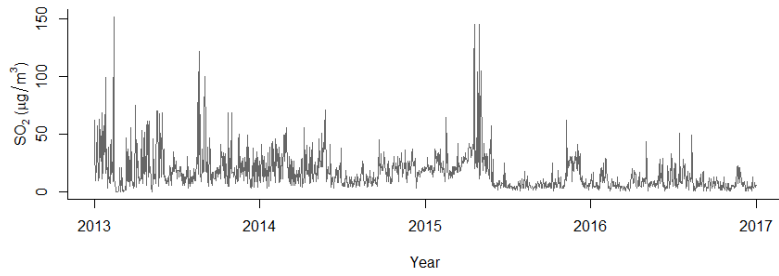


(d) PM_{2.5}

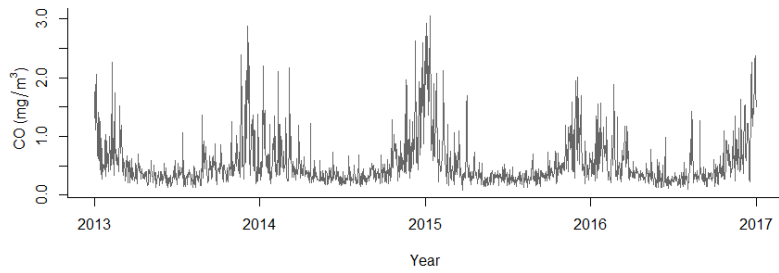
Figure 3.19: Daily maximum value air pollutants time series. (a) NO₂, (b) NO_x, (c) PM₁₀, (d) PM_{2.5}, (e) O₃, (f) SO₂, (g) CO. (To be continued)



(e) O₃



(f) SO₂



(g) CO

Figure 3.19: Daily maximum value air pollutants time series. (a) NO₂, (b) NO_x, (c) PM₁₀, (d) PM_{2.5}, (e) O₃, (f) SO₂, (g) CO. (Continued)

The cross-correlation function between daily hospital admissions and air pollutants was analysed, which allowed to establish at which lag the correlation between daily hospital admissions and each air pollutant was the highest. These results can be found in Figure 3.20. A remark on the lags studied must be made. After discussing with the air quality expert, we decided that in order to quantify the short-term effect of air pollution on health it would be sufficient to consider lags up to 7 days. Furthermore, the cross-correlation function is always estimated for positive and negative lags. However, we only considered positive lags, since air pollution leads to hospital admissions, and not the opposite.

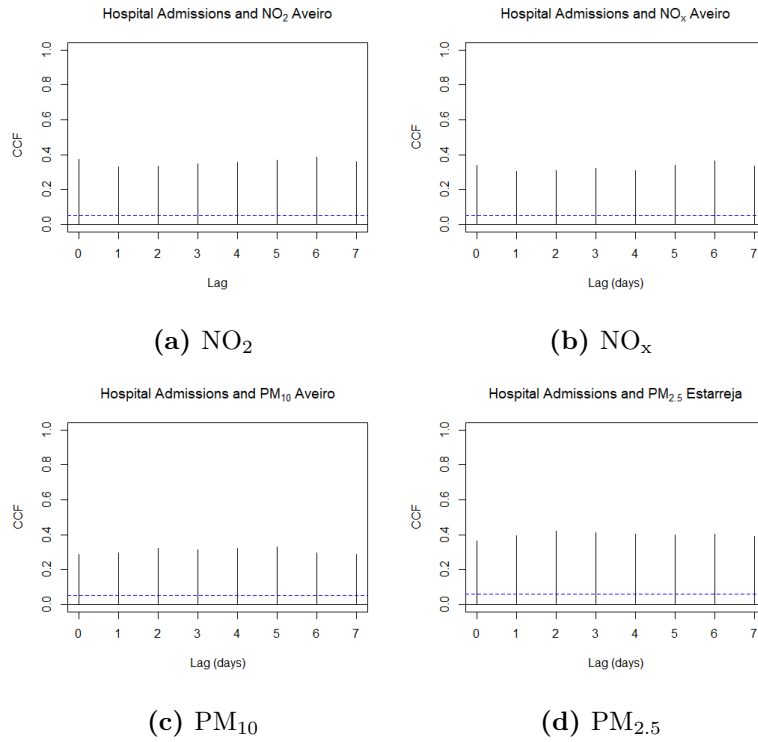


Figure 3.20: CCF between Aveiro hospital admissions and air pollutants in the nearest stations. (a) NO₂, (b) NO_x, (c) PM₁₀, (d) PM_{2.5}, (e) O₃, (f) SO₂, (g) CO.

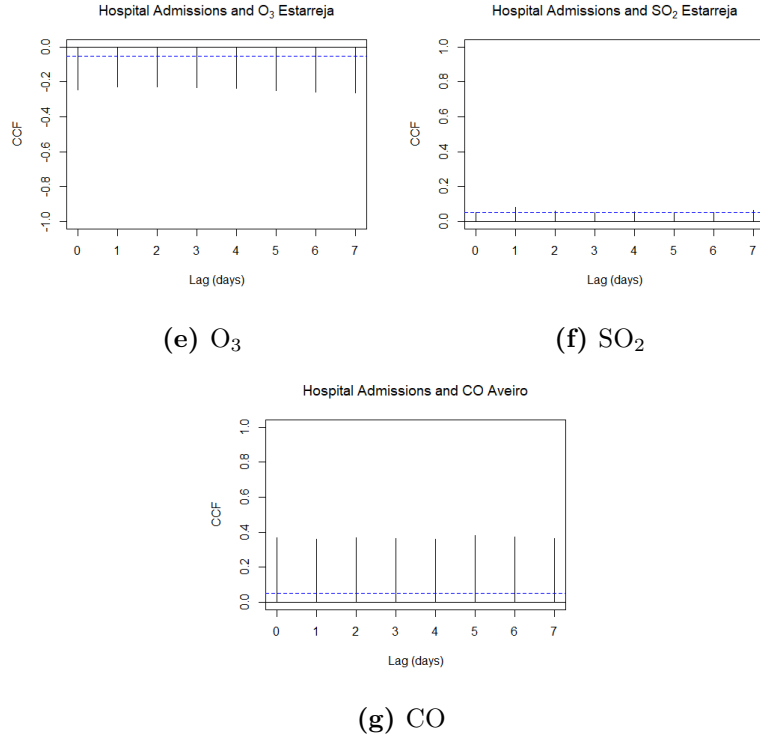


Figure 3.20: CCF between hospital admissions and air pollutants in the nearest stations. (a) NO₂, (b) NO_x, (c) PM₁₀, (d) PM_{2.5}, (e) O₃, (f) SO₂, (g) CO.

Regarding NO₂ and NO_x the highest correlation with daily hospital admissions occurs at lag 6, which suggests that the effect of these pollutants on hospital admissions might occur six days later [Figure 3.20(a,b)]. Since NO_x measurements contain about 90% of NO₂, it is no surprise that the same lag describes the highest correlation between both pollutants and daily hospital admissions.

As for particulate matter, the lag with the highest correlation found was 5 and 2, for PM₁₀ and PM_{2.5}, respectively [Figure 3.20(c,d)]. Hence, it may be the case that the impact of PM₁₀ and PM_{2.5} levels on hospital admissions are observed after 5 and 2 days, respectively. Particulate matter with a diameter inferior to $2.5\mu m$ has more damaging effects on health than PM₁₀, since PM_{2.5} are smaller in size and can get attached to the alveoli, causing respiratory symptoms [62].

As for O₃, it presents a negative correlation at all lags evaluated [Figure 3.20(e)]. However, it is highly unlikely that ozone has a protective effect on hospital admissions [62]. The negative correlations most certainly result from the fact that O₃ is rapidly consumed in the reaction with nitrogen oxide (Equation 3.1.1), which leaves little O₃ in the atmosphere so that any direct effect of this pollutant on hospital admissions can be quantified.

SO₂ shows a significant, but small correlation at lag 1 only [Figure 3.20(f)]. This pollutant levels' have been within the legislation limits for several years now.

Lastly, CO highest lag correlated with daily hospital admission is lag 5 [Figure 3.20(g)]. Similarly to PM₁₀, the effect of this pollutant on hospital admissions seems to be reflected after 5 days of the emissions.

After establishing which lags are to be considered for each pollutant, it is necessary to study the correlation between pollutants, so that collinearity is not introduced in the model, which may lead to poor performance [16]. Table 3.8 shows the correlation matrix between pollutants at the previously selected lags. NO₂ and NO_x values are highly correlated ($R = 0.86$), which was expected since NO_x composition is about 90% of NO₂. Curiously, PM₁₀ presents considerable correlation with both NO₂, NO_x and CO, thus, including these variables all together in the model will induce colinearity issues. Carbon monoxide presents some correlation with NO₂ and NO_x, but this is probably small enough so that no colinearity is introduced in the model. As for PM_{2.5}, SO₂ and O₃ no high correlation are found with other air pollutants. Besides, it is noteworthy, that O₃ is negatively correlated with all air pollutants and SO₂ has very low correlations with the remaining air pollutants.

	NO ₂ (Lag ₆)	NO _x (Lag ₆)	PM ₁₀ (Lag ₅)	PM _{2.5} (Lag ₂)	O ₃ (Lag ₇)	SO ₂ (Lag ₁)	CO (Lag ₅)
NO ₂ (Lag ₆)	1	0.86	0.61	0.43	-0.14	0.08	0.51
NO _x (Lag ₆)		1	0.62	0.49	-0.25	0.04	0.53
PM ₁₀ (Lag ₅)			1	0.49	-0.11	0.14	0.73
PM _{2.5} (Lag ₂)				1	-0.34	0.02	0.52
O ₃ (Lag ₇)					1	0.08	-0.25
SO ₂ (Lag ₁)						1	0.08
CO (Lag ₅)							1

Table 3.8: Pearson correlation between air pollutants lagged time series. Bold indicates correlations higher than 0.50.

Since the optimal lag to describe the relationship between hospital admissions and air pollution as well as the colinearity were studied, there are conditions to establish which air pollutants must be included in the model. It was decided that SO₂ would not be included since its values are well below the established legal limits. Also, O₃ was not included in the model since it is mainly consumed in the reaction with NO_x, hence its impact on hospital admissions is difficult to quantify. PM₁₀ has a strong correlation with CO, which results from the fact that both are originated from some common sources, such as combustion [62]. As such, we decided to include CO instead of PM₁₀, since the later also has moderate correlations with other pollutants. Finally, as NO_x is composed mainly by NO₂ (up to 90%), we decided to include NO_x in the model, to contemplate the effect of NO_x as well as that of NO₂ in

the model. Therefore, the model included as covariables CO and NO_x. This decision was discussed with and supported by the air quality expert. PM_{2.5} was also a good candidate to be included in the model as covariate, however in 2016 there were no measurements performed at this station, most likely as a result of malfunctioning of the probe, and, thus, PM_{2.5} was excluded from the model.

3.2.2 Model Fit

As previously mentioned, generalised linear models can be estimated with the Poisson Distribution or the Negative Binomial Distribution. The Poisson Distribution assumes that the conditional mean of the process is equal to its conditional variance. However, in hospital admissions time series it is not common for such to happen. Therefore, it was decided to use the Negative Binomial distribution to model the distribution of the conditional mean of the process.

Furthermore, to fit a generalised linear model it is necessary to define the order of the sets P and Q (Section 2.2.1), which is made by analysing the empirical autocorrelation function (Figure 3.18). The set P allows for regression on past observations, meaning that it accounts for serial dependence, which can be accounted for considering $P = 1$. As for set Q , it allows for regression on lagged conditional means, that is accounts for seasonality, which is accounted by considering $Q = 7$.

Lastly, it is also necessary to define the link function, which can be the identity or the logarithmic function. The later has the advantage of allowing for negative correlations and easier accommodation of covariates. A negative correlation is not a problem since the selected covariates are positively correlated with the response variable [Figure 3.20(a, d, g)]; however, covariates accommodation may be a problem with the identity link function. Therefore, both models were fitted and compared. Table 3.9 shows the coefficients for each model and the respective AIC and BIC. The model with the identity link function (Model I) has lower AIC and BIC, which indicates that this model is preferred to describe the relationship between hospital admissions and the air pollutants.

The overdispersion coefficient (σ^2), is related to the dispersion parameter (ϕ) of the Negative Binomial Distribution by the following relationship, $\phi = \frac{1}{\sigma^2}$. Therefore, the fitted model for the number of daily hospital admissions y_t is given by $y_t | \mathcal{F}_{(t-1)} \sim \text{NegBin}(\lambda_t, 17.857)$ with,

$$\lambda_t = 1.123 + 0.398y_{t-1} + 0.439\lambda_{t-7} + 1.298x_{t-5} + 0.0013z_{t-6}, \quad (3.5)$$

where x_t is the time series of CO and z_t is the time series of NO_x. It is noteworthy that the overdispersion coefficient is significantly different from zero, which supports the choice of

using the Negative Binomial Distribution instead of the Poisson. Otherwise, the appropriate distribution would be the Poisson Distribution, since when $\phi \rightarrow \infty$ the Poisson distribution is a limiting case of the Negative Binomial [39].

The 95% confidence interval of all coefficients do not contain the zero value and, therefore, the corresponding parameters of the model are significantly different from zero, which supports that NO_x and CO levels respectively, 6 and 5 days earlier, have significant impact on hospitals admissions (Table 3.9).

	Model I			Model L		
	Coef.	Std. Error	95% CI*	Coef.	Std. Error	95% CI*
β_0	1.123	0.311	[0.635, 1.870]	0.371	0.051	[0.270, 0.469]
β_1	0.398	0.026	[0.336, 0.441]	0.421	0.028	[0.357, 0.476]
α_7	0.439	0.032	[0.383, 0.502]	0.425	0.035	[0.353, 0.487]
CO	1.298	0.523	[0.342, 2.329]	0.050	0.023	[0.009, 0.101]
NO_x	0.013	0.002	[0.007, 0.017]	0.001	0.0002	[0.0002, 0.0007]
σ^2	0.056	0.005	[0.046, 0.066]	0.058	0.005	[0.048, 0.068]
AIC	6842.8			6877.7		
BIC	6872.8			6907.7		

*Confidence intervals computed with bootstrap ($B = 500$).

Table 3.9: Coefficients and adequacy measures for the generalised linear models. Model I - Identity Link Function, Model L - Logarithmic Link Function.

Finally, the distribution of the residuals of both models are identical, as well as the ACF and PACF of the residuals (Figure 3.21), thus there is no advantage in using the logarithmic function as the link function. Moreover, because the use of the logarithmic function requires the transformation of the response variable, and increases the difficulty in the interpretation of the results.

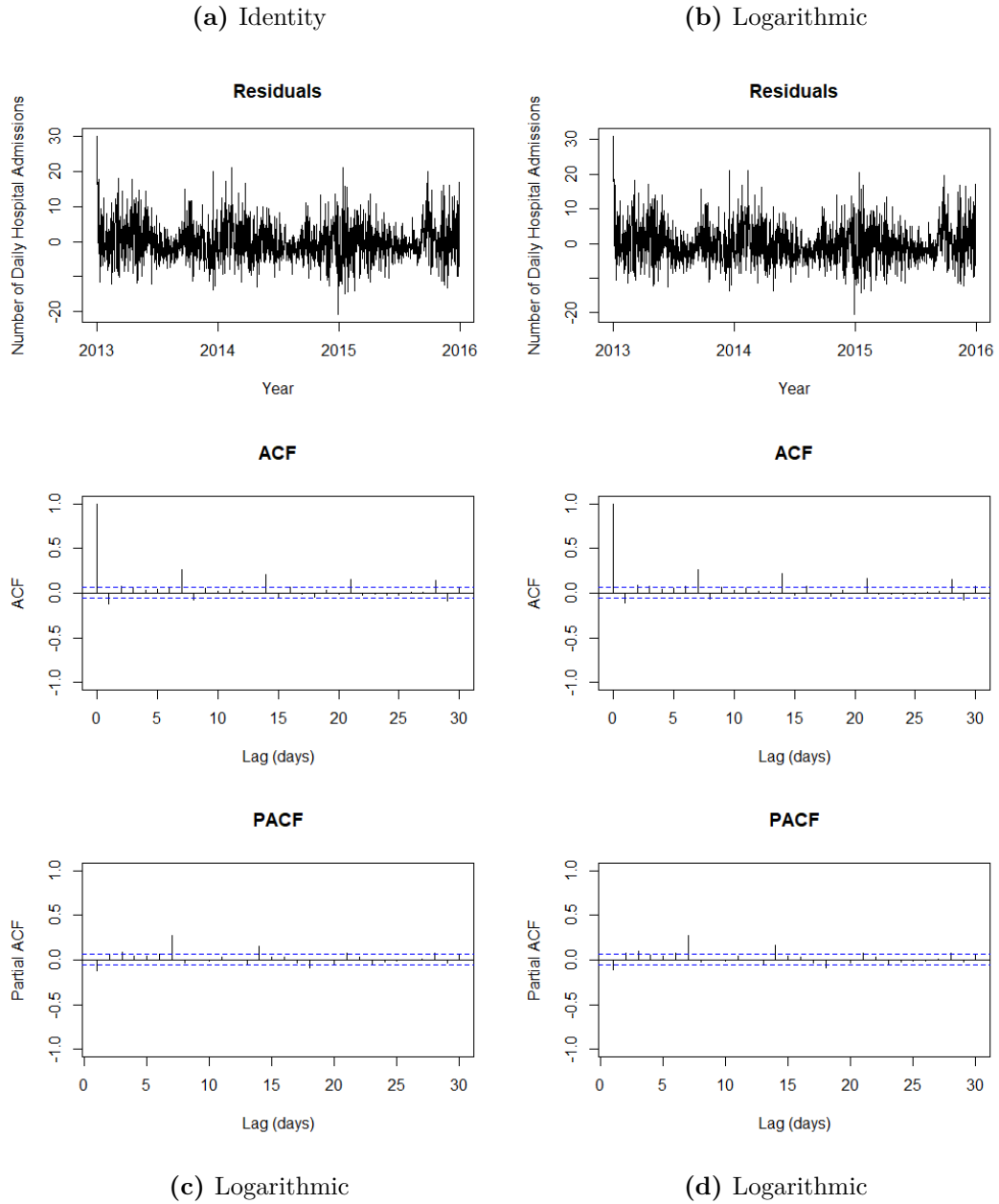


Figure 3.21: Analysis of residuals of hospital admissions. (a) Link Function - Identity, (b) Link Function - Logarithmic.

Even though we have built our model under theoretical and empirical decisions, we now show that this model is optimal to describe the data. Alternative models to be considered are presented in Table 3.10 and consider only NO_x , NO_2 , PM_{10} and CO , since O_3 , SO_2 , $\text{PM}_{2.5}$ were definitively excluded. The remaining pollutants NO_x , NO_2 , PM_{10} and CO , should not be considered all under the same model, because i) NO_x and NO_2 represent pretty much the

same information and ii) CO and PM_{10} are highly correlated. Therefore, leaving the options considered in Table 3.10. It is clear that the model considering $NO_x + CO$, with 6 and 5 lagged days, respectively, has the lowest AIC and BIC, which supports our choice.

Model	AIC	BIC
CO + NO_x	6842.8	6872.8
CO + NO_2	6852.9	6882.9
$NO_2 + PM_{10}$	6865.4	6895.4
$NO_x + PM_{10}$	6847.9	6877.9

Table 3.10: AIC and BIC information criteria for the optimal model and 3 alternative models.

3.2.3 Model Forecast Strategies

Since the model is established, it is possible to compute the forecasts. In order to comply with the defined objective, forecasts for January 2016 using the observed air pollutants data and the forecasts of SARFIMA models were performed.

Figure 3.22 shows the mean predicted value and the 95% confidence interval of the forecasts using the observed covariables data and forecasts from SARFIMA models. Overall, mean predicted values are somewhat similar up to mid-January. The 95% confidence intervals seem a bit narrower than the predicted values using the forecasts from SARFIMA, nevertheless, both confidence intervals contain all of the observed data (Figure 3.22). Performance measures for each forecast can be found in Table 3.11. All performance measures are better for the forecasts using the observed covariables data, as expected. However, the forecasts using the air pollutants forecasts from SARFIMA model also seem to behave fairly well, since its mean MASE is smaller than one, which indicates that the forecasts are better than the average 1-step naive forecast. Considering that the time series of hospital admissions has mean 24 and standard deviation 10, the scale-dependent measures show good performance of both forecast strategies. Regarding the MAP and the MAPE, they show that the bias within each forecast ranges from 10% up to 26%, which is fairly reasonable.

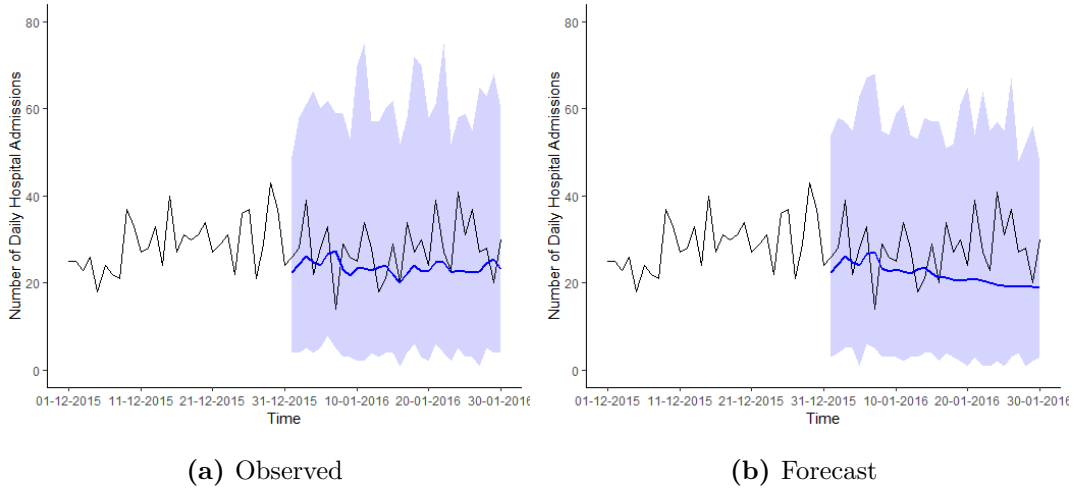


Figure 3.22: Prediction intervals at 95%. (a) using observed NO_x and CO, (b) using forecasts from NO_x and CO.

Figure 3.23 shows the observed time series of daily hospital admissions (green) and the mean forecast of daily hospital admissions using the observed air pollutants data (purple) and the forecasts of the SARFIMA models (orange). Since CO lag is 5 days, the forecasts start only on January 6. The forecast from the observed air pollutants is able to follow the trend

of the series, except when there is a substantial increase or decrease of daily admissions in the original time series. Nevertheless, the predicted mean values using the SARFIMA forecasts are similar to the forecasts using the observed air pollutant data up to 16th of January, decaying to the time series mean afterwards. This indicates that air pollution forecasts from SARFIMA models allow to obtain reliable values for a 10-day period, which can be more than helpful to establish weekly management plans in a hospital to deal with more critical periods, such as forest fires.

	ME	RMSE	MAE	MAP	MAPE	MASE	PI 95%
Observed Covariables	−4.342	7.821	6.355	10.408	22.431	0.743	100.00
Forecast Covariables	−5.946	9.215	7.600	16.139	26.327	0.889	100.00

Table 3.11: Performance measures and percentage of observations within the prediction intervals of one month forecasts for each model.

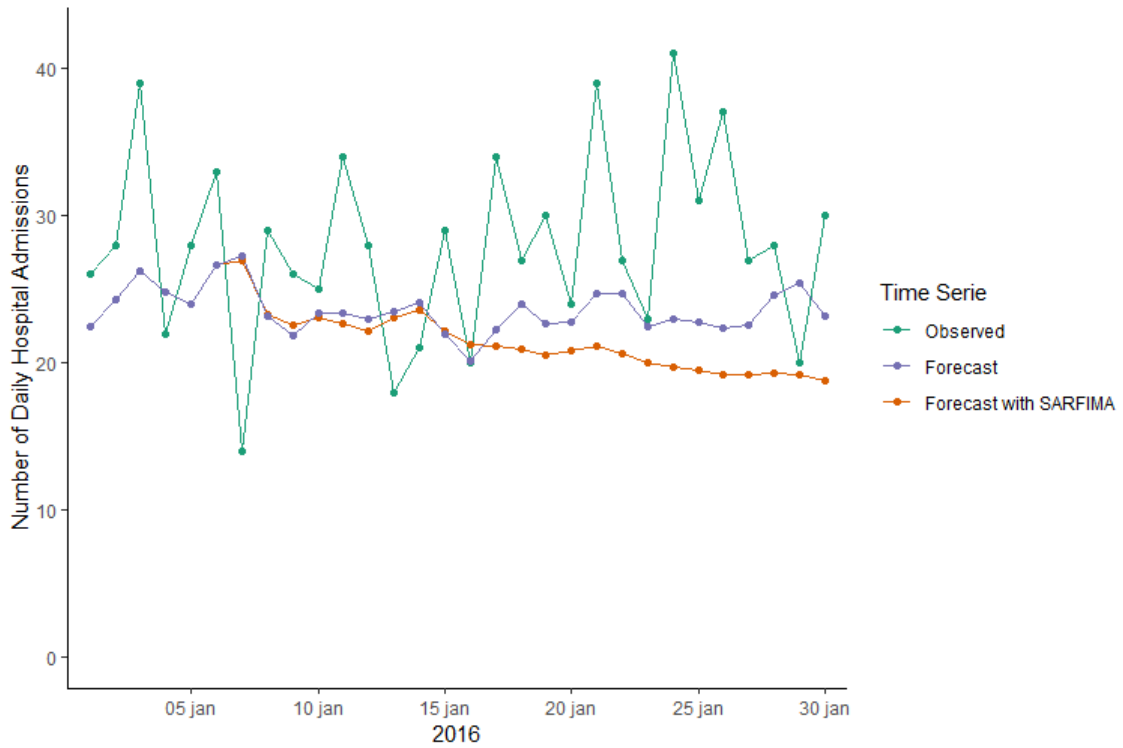


Figure 3.23: Time series of daily hospital admissions and forecasts by air pollutants covariates with observed covariates and predicted SARFIMA covariates.

Chapter 4

Conclusions

This section presents the main conclusions of this study. A framework to model and forecast hourly air pollutants data using ARIMA-based models was successfully implemented, as expected. A positive and significant association between respiratory hospital admissions and air pollutants was found, which contributes to the body of knowledge of air pollutants impact on health, particularly in Portugal, where no studies outside of Lisbon area were found. Future perspectives are also discussed.

This research work establishes a flexible framework to model and forecast hourly time series of environmental pollutants. The idea is to use a pure SARMA model to describe the short memory of the data and to fit an ARFIMA model to the SARMA residuals in order to model the persistent memory component. In general, this approach works well for all pollutants evaluated at the different stations and is able to accurately describe the data. Furthermore, it is possible to obtain forecasts for about 4 months before convergence to the mean is observed. This approach might be interesting to use on other real data applications with hourly sample data or higher.

Regarding daily hospital admissions due to respiratory causes, it was possible to establish a model in which carbon monoxide (lag 5) and nitrogen dioxide (lag 6) were covariables. In fact, these explanatory variables presented statistically significant coefficients, which also indicates that these variables are significantly associated with daily hospital admissions due to respiratory causes. Therefore, it is shown that air pollutants, namely CO and NO_x have a short-term impact on daily hospital admissions due to respiratory outcomes in Aveiro hospital. Furthermore, it is important to highlight that CO and NO_x are generally within the legal limits. Therefore, this study stresses the importance to continue to study the impact of air quality on health, despite the mandatory limits being complied with.

Another interesting finding in this study, is that the forecasts of hospital admissions using SARFIMA air pollutants forecasts are quite reliable when compared to the forecasts of hospital admissions using air pollutants observed data. Hence, the air pollutants models developed may be of great use for hospital admissions predictions, and, perhaps, to increase the forecast period on daily hospital admissions. This can be of added value for hospital planning and management of resources, particularly at times of increased air pollution such as in forest fires seasons.

Nevertheless, there are additional questions that are quite relevant to explore. Therefore, *future perspectives* include to carry out an intervention analysis (i.e., sudden changes in the time series), considering forest fires as interventions, and understand how these impact hospital admissions. Furthermore, it is of the utmost importance to include temperature in these models, since it is known that temperature is possibly associated with air pollutants, which may have a direct impact on the magnitude and significance of air pollutants coefficients. Additionally, it is also of interest to reproduce this study to other locations in Portugal, as there are a limited number of studies on the association of air pollutants and hospital admissions in Portugal. Since these types of studies are lacking in Portugal, to quantify the attributable proportion of disease and mortality due to air pollution would be imperative to establish good air quality guidelines. Note that it is also necessary to expand the hospital admissions cause beyond respiratory outcomes, including cardiovascular, neurological, among other causes.

Finally, there are several contributions to time series analysis that require further development, particularly on space-time count models. Integer ARMA extensions to space-time (STARMA) have been largely overlooked in the literature, but these are of potential use in this context, as they can ensure the discreteness of the process by replacing the multiplication in the conventional models by suitable thinning operators [48]. The INGARCH models were first formulated as analogous to GARCH where the conditional distribution of the observed counts given past outcomes was assumed to be Poisson [17]. In this approach, the time-dependent conditional mean depends on the past values of the series and on its own past values. Further INGARCH developments included the Poisson replacement by other well-known discrete distributions including negative binomial and generalized Poisson. More recently, the integer-valued process with general infinitely divisible discrete probability laws was proposed to unify and enlarge the class of INGARCH models [24]. Lastly, it was just last year that a class of models extending the INGARCH class to account for small-scale spatial variation (SPINGARCH) was proposed [13]. Therefore, the development of space-time INARMA models is still in its infancy and scientific contributions to this topic are highly desirable given the large applicability of these models beyond the analysis of daily hospital admissions time series.

Bibliography

- [1] QualAr - Base de Dados Online sobre a Qualidade do Ar, howpublished = <https://qualar1.apambiente.pt/qualar/>, note = Accessed: 2019-05-30.
- [2] Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*, 2008.
- [3] Decreto-Lei 102/2010 de 23 de Setembro. *Diário da República Portuguesa*, 2010.
- [4] C. A. Alves, M. G. Scotto, and M. C. Freitas. Air pollution and emergency admissions for cardiorespiratory diseases in Lisbon (Portugal). *Química Nova*, 33(2):337–344, 2010.
- [5] F. Amato, F. R. Cassee, H. A. D. van der Gon, R. Gehrig, M. Gustafsson, W. Hafner, R. M. Harrison, M. Jozwicka, F. J. Kelly, T. Moreno, et al. Urban air quality: the challenge of traffic non-exhaust emissions. *Journal of Hazardous Materials*, 275:31–36, 2014.
- [6] H. R. Anderson. Air pollution and mortality: A history. *Atmospheric Environment*, 43(1):142–152, 2009.
- [7] M. L. Bell and D. L. Davis. Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environmental Health Perspectives*, 109(suppl 3):389–394, 2001.
- [8] L. Beretta and A. Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, 16(3):74, 2016.
- [9] C. Bisognin and S. R. C. Lopes. Properties of seasonal long memory processes. *Mathematical and Computer Modelling*, 49(9-10):1837–1851, 2009.
- [10] P. J. Brockwell, R. A. Davis, and M. V. Calder. *Introduction to time series and forecasting*. Springer, 2 edition, 2002.

- [11] J. Carré, N. Gatimel, J. Moreau, J. Parinaud, and R. Léandri. Does air pollution play a role in infertility? A systematic review. *Environmental Health*, 16(1):82, 2017.
- [12] V. Christou and K. Fokianos. Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis*, 35(1):55–78, 2014.
- [13] N. J. Clark, M. S. Kaiser, and P. M. Dixon. A spatially correlated auto-regressive model for count data. *arXiv preprint arXiv:1805.08323*, 2018.
- [14] A. M. J. Cruz, C. Alves, S. Gouveia, M. G. Scotto, M. d. C. Freitas, and H. T. Wolterbeek. A wavelet-based approach applied to suspended particulate matter time series in Portugal. *Air Quality, Atmosphere & Health*, 9(8):847–859, Dec 2016.
- [15] D. Dias, O. Tchepel, and C. Borrego. Health impact assessment of exposure to inhalable particles in Lisbon metropolitan area. *WIT Transactions on Biomedicine and Health*, 14:91–101, 2009.
- [16] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.
- [17] R. Ferland, A. Latour, and D. Oraichi. Integer-valued GARCH process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.
- [18] K. Fokianos and D. Tjøstheim. Log-linear Poisson autoregression. *Journal of Multivariate Analysis*, 102(3):563–578, 2011.
- [19] C. Fraley, F. Leisch, M. Maechler, V. Reisen, and A. Lemonte. Package ‘fracdiff’ manual. *CRAN*, 2015.
- [20] P. Fu, X. Guo, F. M. H. Cheung, and K. K. L. Yung. The association between PM2.5 exposure and neurological disorders: A systematic review and meta-analysis. *Science of The Total Environment*, 2018.
- [21] G. Gardner, A. C. Harvey, and G. D. Phillips. Algorithm as 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):311–322, 1980.
- [22] P. Garrett and E. Casimiro. Short-term effect of fine particulate matter (PM2.5) and ozone on daily mortality in Lisbon, Portugal. *Environmental Science and Pollution Research*, 18(9):1585–1592, 2011.

- [23] J. Geweke and S. Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4):221–238, 1983.
- [24] E. Gonçalves, N. Mendes-Lopes, and F. Silva. Infinitely divisible distributions in integer-valued GARCH models. *Journal of Time Series Analysis*, 36(4):503–527, 2015.
- [25] S. Gouveia, M. G. Scotto, A. Monteiro, and A. M. Alonso. Wavelets-based clustering of air quality monitoring sites. *Environmental Monitoring and Assessment*, 187(11):694, 2015.
- [26] C. W. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis*, 1(1):15–29, 1980.
- [27] J. Haslett and A. E. Raftery. Space-time modelling with long-memory dependence: Assessing Ireland’s wind power resource. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 38(1):1–21, 1989.
- [28] J. Hosking. Fractional differencing. *Biometrika*, 69:165–176, 1981.
- [29] J. Hosking. Fractional differencing modeling in hydrology. *JAWRA Journal of the American Water Resources Association*, 21(4):677–682, 1985.
- [30] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
- [31] R. J. Hyndman, Y. Khandakar, et al. Automatic time series for forecasting: the forecast package for R. *Journal of Statistical Software, Articles*, (6/07), 2007.
- [32] R. J. Hyndman, Y. Khandakar, et al. Package ‘forecast’ manual. *CRAN*, 2018.
- [33] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
- [34] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.
- [35] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin. 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, 1992.
- [36] P. J. Landrigan, R. Fuller, N. J. Acosta, O. Adeyi, R. Arnold, A. B. Baldé, R. Bertollini, S. Bose-O’Reilly, J. I. Boufford, P. N. Breyse, et al. The Lancet Commission on pollution and health. *The Lancet*, 391(10119):462–512, 2018.

- [37] J. Lelieveld, K. Klingmüller, A. Pozzer, U. Pöschl, M. Fnais, A. Daiber, and T. Münzel. Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions. *European Heart Journal*, 03 2019.
- [38] M.-H. Li, L.-C. Fan, B. Mao, J.-W. Yang, A. M. Choi, W.-J. Cao, and J.-F. Xu. Short-term exposure to ambient fine particulate matter increases hospitalizations and mortality in COPD: a systematic review and meta-analysis. *Chest*, 149(2):447–458, 2016.
- [39] T. Liboschik, K. Fokianos, and R. Fried. tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software, Articles*, 82(5):1–51, 2017.
- [40] D. Loomis, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, R. Baan, H. Mattock, and K. Straif. The carcinogenicity of outdoor air pollution. *The Lancet Oncology*, 14(13):1262–1263, 2013.
- [41] F. Mirante, P. Salvador, C. Pio, C. Alves, B. Artiñano, A. Caseiro, and M. A. Revuelta. Size fractionated aerosol composition at roadside and background environments in the madrid urban atmosphere. *Atmospheric Research*, 138:278–292, 2014.
- [42] C. Mohr, P. DeCarlo, M. Heringa, R. Chirico, J. Slowik, R. Richter, C. Reche, A. Alastuey, X. Querol, R. Seco, et al. Identification and quantification of organic aerosol from cooking and other sources in Barcelona using aerosol mass spectrometer data. *Atmospheric Chemistry and Physics*, 12(4):1649–1665, 2012.
- [43] A. Montanari, R. Rosso, and M. S. Taquq. Fractionally differenced arima models applied to hydrologic time series: Identification, estimation, and simulation. *Water resources research*, 33(5):1035–1044, 1997.
- [44] E. Moore, L. Chatzidiakou, M.-O. Kuku, R. L. Jones, L. Smeeth, S. Beevers, F. J. Kelly, B. Barratt, and J. K. Quint. Global associations between air pollutants and chronic obstructive pulmonary disease hospitalizations. a systematic review. *Annals of the American Thoracic Society*, 13(10):1814–1827, 2016.
- [45] B. Murteira, D. Müller, and K. F. Turkman. *Análise de sucessões cronológicas*. 2000.
- [46] M. G. Porpora, I. Piacenti, S. Scaramuzzino, L. Masciullo, F. Rech, and P. Benedetti Panici. Environmental contaminants exposure and preterm birth: A systematic review. *Toxics*, 7(1):11, 2019.
- [47] S. Robertson and M. R. Miller. Ambient air pollution and thrombosis. *Particle and Fibre Toxicology*, 15(1):1, 2018.

- [48] M. G. Scotto, C. H. Weiß, and S. Gouveia. Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling*, 15(6):590–618, 2015.
- [49] A. S. Shah, J. P. Langrish, H. Nair, D. A. McAllister, A. L. Hunter, K. Donaldson, D. E. Newby, and N. L. Mills. Global association of air pollution and heart failure: a systematic review and meta-analysis. *The Lancet*, 382(9897):1039–1048, 2013.
- [50] R. H. Shumway and D. S. Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2017.
- [51] Q. Song, D. Christiani, J. Ren, et al. The global contribution of outdoor air pollution to the incidence, prevalence, mortality and hospital admission for chronic obstructive pulmonary disease: a systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, 11(11):11822–11832, 2014.
- [52] D. M. Stieb, L. Chen, M. Eshoul, and S. Judek. Ambient air pollution, birth weight and preterm birth: a systematic review and meta-analysis. *Environmental Research*, 117:100–111, 2012.
- [53] G. Sun, G. Hazlewood, S. Bernatsky, G. G. Kaplan, B. Eksteen, and C. Barnabe. Association between air pollution and the development of rheumatic disease: a systematic review. *International Journal of Rheumatology*, 2016, 2016.
- [54] R.-C. Team. The R stats package.
- [55] D. E. Warburton, S. S. Bredin, E. M. Shellington, C. Cole, A. de Faye, J. Harris, D. D. Kim, and A. Abelsohn. A systematic review of the short-term health effects of air pollution in persons living with coronary heart disease. *Journal of Clinical Medicine*, 8(2):274, 2019.
- [56] C. H. Weiß. Serial dependence and regression of Poisson INARMA models. *Journal of Statistical Planning and Inference*, 138(10):2975–2990, 2008.
- [57] R. Weron. Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications*, 312(1-2):285–299, 2002.
- [58] WHO. *Air quality guidelines for Europe. Copenhagen: WHO Regional Office for Europe*. World Health Organization, 1987.
- [59] WHO. International Classification of Diseases-Ninth Revision (ICD-9). *Weekly Epidemiological Record*, 63(45):343–344, 1988.

- [60] WHO. *Air quality guidelines for Europe*. WHO, second edition edition, 2000.
- [61] WHO. Health aspects of air pollution: results from the WHO project” Systematic review of health aspects of air pollution in Europe”. Technical report, Copenhagen: WHO Regional Office for Europe, 2004.
- [62] WHO. *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. World Health Organization, 2006.
- [63] WHO. Available evidence for the future update of the who global air quality guidelines (AQGs), 2018.
- [64] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.
- [65] X. Xu, S. Nie, H. Ding, and F. F. Hou. Environmental pollution and kidney diseases. *Nature Reviews Nephrology*, 2018.
- [66] B.-Y. Yang, Z. Qian, S. W. Howard, M. G. Vaughn, S.-J. Fan, K.-K. Liu, and G.-H. Dong. Global association between ambient air pollution and blood pressure: a systematic review and meta-analysis. *Environmental Pollution*, 235:576–588, 2018.
- [67] P. C. Young, D. J. Pedregal, and W. Tych. Dynamic harmonic regression. *Journal of Forecasting*, 18(6):369–394, 1999.
- [68] L. Yuan, Y. Zhang, Y. Gao, and Y. Tian. Maternal fine particulate matter (PM_{2.5}) exposure and adverse birth outcomes: an updated systematic review based on cohort studies. *Environmental Science and Pollution Research*, pages 1–21, 2019.
- [69] T. Zhao, I. Markevych, M. Romanos, D. Nowak, and J. Heinrich. Ambient ozone exposure and mental health: A systematic review of epidemiological studies. *Environmental Research*, 165:459–472, 2018.
- [70] X.-y. Zheng, H. Ding, L.-n. Jiang, S.-w. Chen, J.-p. Zheng, M. Qiu, Y.-x. Zhou, Q. Chen, and W.-j. Guan. Association between air pollutants and asthma emergency room visits and hospital admissions in time series studies: a systematic review and meta-analysis. *PloS One*, 10(9):e0138146, 2015.

Appendices

Appendix A

Introduction

A.1 Air Pollutants Legal Limits

Station	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Arcos	0	0	0	0	0	0	0	0	0	0	0	0
Aveiro	8	0	2	5	0	0	0	0	0	0	0	0
Av. Liberdade	35	81	39	20	91	21	37	13	15	20	20	14
Beato	0	0	2	0	0	0	0	0	0	0	0	0
Chamusca	0	0	0	0	0	0	0	0	0	0	0	0
Custóias	0	1	0	1	16	1	15	16	0	0	0	0
Entrecampos	2	4	9	3	20	15	0	8	0	2	0	0
Ermesinde	0	0	2	1	33	0	4	0	0	0	0	0
Ervedeira	0	0	0	0	0	0	0	0	0	0	0	0
Escavadeira	0	0	12	0	0	0	0	0	0	0	0	0
Estarreja	0	0	0	0	0	0	0	0	0	0	0	0
Fornelo Monte	0	0	0	0	0	0	0	0	0	0	0	0
Fr. Sá Carneiro	4	2	17	5	2	7	11	8	0	0	10	6
Fr. Bartolomeu	0	3	4	0	0	0	12	1	8	0	0	0
Frossos	0	0	0	0	0	0	0	0	0	0	0	0
Fundão	0	0	0	0	0	0	0	0	0	0	0	0
Ílhavo	0	0	0	0	0	0	0	0	0	0	0	0
Instituto Geofísico	0	0	0	0	0	0	0	0	0	0	0	0
Laranjeiro	3	0	7	4	4	4	2	0	0	0	1	0
Loures	0	0	0	0	0	0	0	0	0	0	0	0
Mem Martins	0	0	0	0	0	0	0	0	0	0	0	0
Monte Chãos	0	0	0	0	0	0	0	0	0	0	0	0
Olivais	26	0	3	1	9	17	18	2	0	0	0	0
Paio Pires	0	0	0	0	0	0	0	0	0	0	0	0
Pe Moreira Neves	0	0	0	0	0	0	0	0	0	0	0	0
Quebedo	0	0	0	0	0	0	0	0	0	0	0	0
Restelo	0	0	0	0	0	0	2	0	0	0	0	0
Sonega	0	0	0	0	0	0	0	0	0	0	0	0
Vermoim	0	0	1	0	0	0	2	0	0	0	0	0

Table A.1: Number of one hour limit value exceedances of NO₂ for all stations (2005-2016). Bold numbers surpass the annual maximum number of exceedances (18).

Station	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Arcos	20.0	18.8	21.3	20.6	19.2	16.1	17.6	15.6	14.7	13.1	14.2	12.6
Aveiro	27.5	29.4	30.4	30.2	28.4	27.6	25.4	24.6	23.0	25.0	23.1	22.7
Av. Liberdade	63.6	71.2	74.2	64.6	69.5	64.8	60.8	58.3	52.4	52.9	58.6	57.1
Beato	27.8	29.0	29.9	27.1	27.6	25.2	23.3	22.8	19.3	19.9	20.8	19.8
Chamusca	6.6	6.9	7.8	7.3	7.7	6.9	6.4	5.8	6.3	5.5	5.7	4.8
Custóias	28.8	25.8	43.0	32.9	27.0	9.4	21.5	33.1	26.4	27.4	27.8	29.8
Entrecampos	51.3	50.4	48.7	42.8	52.4	46.1	41.3	41.5	38.7	36.7	39.5	36.9
Ermesinde	29.9	29.6	37.0	33.2	35.5	30.2	30.0	28.5	24.1	21.4	28.6	26.0
Ervedeira	8.7	5.9	12.0	11.5	6.2	6.0	5.6	6.3	5.2	4.6	5.9	5.3
Escavadeira	27.1	28.3	28.8	25.1	24.3	20.7	19.6	19.0	15.8	14.7	18.1	15.8
Estarreja	24.8	19.8	21.1	19.5	21.2	20.8	19.0	15.7	14.7	14.8	13.7	13.0
Fornelo Monte	3.2	2.9	2.7	1.9	3.5	5.9	3.9	1.5	2.5	1.6	1.3	1.5
Fr. Sá Carneiro	47.8	46.9	48.4	47.2	46.9	50.6	47.6	43.9	36.1	46.4	64.4	67.3
Fr. Bartolomeu	40.2	34.6	38.3	45.9	47.7	47.8	54.4	49.1	49.4	44.0	46.5	48.5
Frossos	16.5	16.2	14.9	14.2	16.5	15.2	16.9	15.1	12.0	4.3	12.7	13.1
Fundão	4.9	6.3	7.0	6.0	5.4	3.8	6.5	5.2	5.1	6.6	6.9	6.0
Ílhavo	15.7	10.4	11.2	11.4	8.6	9.6	12.1	5.3	5.3	8.5	13.5	12.1
Instituto Geofísico	20.3	19.6	21.5	18.4	15.3	14.6	7.5	10.0	14.8	14.8	14.7	14.8
Laranjeiro	31.3	30.4	35.1	28.7	30.4	27.7	29.8	27.1	24.0	22.4	26.4	23.0
Loures	22.1	23.5	26.6	23.2	25.1	18.9	20.7	18.2	20.0	18.2	18.1	17.1
Mem Martins	16.0	16.5	17.7	13.3	14.4	14.5	15.4	14.1	12.7	11.2	12.0	11.9
Monte Chãos	4.4	2.6	2.2	4.2	4.8	3.8	4.3	3.6	7.6	6.9	5.7	4.0
Olivais	36.8	36.3	37.7	34.5	33.9	35.6	36.5	30.3	28.9	26.1	29.6	27.8
Paio Pires	25.6	26.4	30.0	24.5	26.4	25.5	24.7	21.8	20.0	17.9	20.7	19.8
Pe Moreira Neves	27.3	29.4	29.0	27.4	26.5	27.4	31.6	28.3	23.3	21.6	24.5	25.8
Quebedo	31.9	32.4	34.9	31.2	24.5	22.7	25.9	21.1	19.1	19.0	18.8	17.2
Restelo	25.1	24.6	27.3	23.5	25.0	23.7	24.9	22.8	22.4	21.5	21.5	18.6
Sonega	5.8	3.1	3.8	4.0	2.4	2.7	2.3	2.5	4.2	4.3	5.0	4.2
Vermoim	32.6	29.9	33.1	30.2	31.4	30.9	27.1	26.9	22.4	23.3	27.0	13.6

Table A.2: Mean annual NO₂ value for all stations (2005-2016). Bold numbers exceed the calendar year value ($40\mu\text{g}/\text{m}^3$).

Station	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Aveiro	74	47	93	53	58	18	96	63	35	42	23	12
Av. Liberdade	176	140	144	77	93	91	110	76	40	31	63	17
Chamusca	20	16	2	0	0	3	3	1	2	3	2	7
Custóias	113	53	64	13	30	33	45	16	17	1	2	0
Entrecampos	79	82	83	21	33	28	54	21	18	10	3	2
Ermesinde	75	84	80	31	36	32	58	21	19	8	0	5
Ervedeira	68	20	18	4	1	6	19	15	2	3	16	2
Escavadeira	75	73	87	47	25	6	27	11	4	6	6	8
Estarreja	104	74	77	40	38	31	62	49	26	31	23	19
Fornelo Monte	0	5	1	3	1	6	4	3	4	2	2	7
Fr. Sá Carneiro	73	67	58	16	20	26	41	31	16	0	2	0
Fr. Bartolomeu	91	95	102	21	39	36	5	11		0	0	0
Frossos	26	23	26	16	18	8	14	14	3	1	2	0
Fundão	12	13	1	1	0	7	0	6	1	2	2	2
Ílhavo	26	38	33	18	1	7	39	51	17	18	34	6
Instituto Geofísico	26	23	28	4	2	2	12	4	2	6	5	4
Laranjeiro	38	34	27	10	33	17	28	6	5	6	10	8
Loures	41	23	25	8	12	5	9	1	0	2	5	6
Meco	73	55	64	36	44	36	49	49	29	12	9	0
Mem Martins	19	22	15	5	9	2	6	0	0	3	6	4
Monte Velho	5	27	20	2	1	4	3	2	0	1	3	4
Olivais	34	45	22	11	14	9	26	6	4	2	6	8
Quebedo	0	0	0	0	0	0	0	0	0	0	0	0
Vermoim	79	82	83	21	33	28	54	21	18	10	3	2

Table A.3: Number of one hour limit value exceedances of PM₁₀ for all stations (2005-2016). Bold numbers surpass the annual maximum number of exceedances (35).

Station	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Aveiro	38.4	33.6	41.0	37.3	35.8	34.2	41.1	35.1	31.9	31.2	23.7	20.1
Av. Liberdade	53.8	49.0	48.8	40.8	39.4	41.5	43.8	38.1	34.1	30.0	36.0	29.5
Chamusca	26.8	22.6	20.4	16.0	16.7	17.2	17.6	15.4	16.0	15.2	17.3	14.7
Custóias	47.8	37.6	34.4	22.9	24.6	27.6	30.2	20.0	20.7	14.7	17.9	12.5
Entrecampos	45.2	41.8	37.1	30.4	30.5	32.8	30.6	25.2	22.4	23.3	25.4	23.4
Ermesinde	40.9	39.8	38.8	28.8	30.0	29.0	32.7	25.2	24.6	22.0	14.1	20.5
Ervedeira	37.3	25.5	25.5	15.4	15.7	20.4	25.7	21.8	19.7	16.5	22.3	17.7
Escavadeira	37.1	38.2	40.0	34.0	25.6	21.7	26.1	20.9	22.6	19.6	20.8	19.8
Estarreja	40.8	35.7	38.1	32.2	30.0	30.2	33.9	29.9	26.1	24.1	25.5	23.7
Fornelo Monte	8.8	10.8	9.4	10.4	12.4	15.1	15.2	12.6	12.7	11.8	12.7	11.6
Fr. Sá Carneiro	47.9	46.6	45.2	30.2	36.5	32.0	36.2	19.9	14.3	11.9	15.5	15.5
Fr. Bartolomeu	39.9	36.4	35.3	25.6	25.3	27.9	34.1	26.6	24.5	18.0	17.1	14.9
Frossos	26.8	28.7	26.0	23.1	20.5	17.8	20.3	19.5	14.8	13.5	13.6	16.6
Fundão	21.1	21.8	14.8	11.6	12.4	14.6	11.0	12.5	11.4	10.1	13.8	13.0
Ílhavo	27.5	27.9	27.7	26.7	20.8	25.5	28.9	29.3	24.2	23.1	26.7	19.1
Instituto Geofísico	27.3	28.5	27.4	17.1	20.0	19.5	22.3	17.1	19.6	18.9	20.6	15.0
Laranjeiro	31.3	29.3	29.6	23.6	30.8	26.7	27.3	21.5	23.0	20.4	22.6	20.0
Loures	32.7	31.4	28.4	23.3	27.8	24.9	24.7	19.6	17.9	17.5	21.3	17.7
Meco	38.2	35.3	36.5	30.5	31.8	30.1	33.6	31.3	28.5	23.8	22.2	14.7
Mem Martins	27.2	26.0	28.6	21.2	22.6	21.9	22.5	17.4	19.2	19.7	20.6	16.6
Monte Velho	25.6	31.0	30.1	21.7	23.5	22.6	21.8	19.9	21.4	21.2	21.9	19.6
Olivais	30.4	31.2	28.7	24.1	26.2	28.0	31.2	23.5	23.5	21.4	20.9	18.1
Quebedo	39.6	33.9	34.9	28.7	28.6	28.1	28.4	32.4	20.6	20.6	23.7	22.0
Vermoim	39.9	37.2	38.3	24.5	26.0	25.3	31.3	25.5	20.9	14.4	15.0	11.5

Table A.4: Mean annual PM₁₀ value for all stations (2005 and 2016). Bold numbers exceed the calendar year value ($40\mu\text{g}/\text{m}^3$).

Station	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Arcos	23	22	8	6	24	25	16	10	31	12	7	15
Beato	16	12	4	8	2	7	4	6	18	0	5	7
Chamusca	56	43	26	23	54	56	38	24	48	14	24	38
Custóias	1	21	2	1	3	5	3	2	13	0	2	0
Entrecampos	0	7	0	6	3	2	1	2	12	1	2	3
Ermesinde	22	27	5	8	9	9	7	2	9	0	2	4
Ervedeira	30	30	17	8	20	29	0	7	23	3	3	9
Escavadeira	21	16	15	7	6	23	13	10	36	13	9	14
Estarreja	41	27	24	7	8	32	4	11	23	3	6	4
Fornelo Monte	5	63	51	24	47	65	2	22	44	16	20	33
Frossos	13	25	26	1	1	10	17	13	18	2	8	13
Fundão	50	52	15	19	30	31	14	11	20	8	13	9
Ílhavo	27	29	27	10	16	21	13	7	11	1	2	6
Instituto Geofísico	29	15	3	3	21	17	0	1	13	3	6	11
Laranjeiro	11	16	6	6	5	8	17	11	26	8	6	5
Loures	22	19	12	9	11	11	15	13	36	8	6	12
Meco	13	26	13	0	16	4	0	3	4	0	0	3
Mem Martins	37	21	12	14	32	22	13	10	40	8	7	18
Monte Chãos	11	13	0	9	29	7	24	9	30	7	11	35
Olivais	11	20	2	3	6	13	13	6	16	5	3	7
Paços Ferreira	36	33	27	5	13	5	10	15	0	0	0	0
Paio Pires	11	21	7	1	5	19	14	8	20	4	2	6
Restelo	20	19	7	8	8	18	13	7	29	10	5	9
Sonega	17	9	0	1	0	0	0	5	52	10	1	49
VNTelha	17	21	7	4	3	6	2	1	8	0	2	2

Table A.5: Number of daily maximum 8h-mean O₃ exceedances for all stations (2005-2016). Bold numbers surpass the annual maximum value allowed (25).

Station	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Arcos	4	5	1	3	3	1	1	2	8	0	0	1
Beato	4	3	0	1	0	0	0	1	4	0	0	3
Chamusca	15	22	0	3	3	16	4	4	2	0	0	6
Custóias	2	6	0	0	1	1	2	0	6	0	0	0
Entrecampos	0	0	0	0	0	0	0	0	0	0	0	1
Ermesinde	17	17	8	1	7	6	4	9	6	0	0	0
Ervedeira	17	20	0	0	9	8	0	0	0	0	0	3
Escavadeira	12	12	11	6	3	4	0	3	13	2	0	8
Estarreja	32	31	3	8	4	30	2	0	11	0	0	4
Fornelo Monte	8	45	6	29	11	36	1	5	8	0	0	32
Frossos	10	12	8	0	0	3	2	3	9	0	0	0
Fundão	2	2	0	0	0	3	0	0	0	0	0	1
Ílhavo	14	57	7	1	1	13	2	0	2	0	0	2
Instituto Geofísico	4	14	0	0	4	7	0	0	3	0	0	0
Laranjeiro	3	18	2	2	1	1	1	3	10	0	0	5
Loures	5	15	0	0	2	0	3	0	6	0	0	0
Meco	1	13	1	0	9	0	0	0	0	0	0	2
Mem Martins	12	17	0	0	4	3	0	3	13	0	0	0
Monte Chãos	1	2	0	0	0	0	0	0	0	0	0	7
Olivais	3	16	1	0	2	0	1	0	1	0	0	2
Paços Ferreira	30	31	10	1	7	0	7	0	0	0	0	0
Paio Pires	4	20	2	0	2	1	0	1	0	0	0	3
Restelo	3	23	4	0	0	6	0	4	8	0	0	3
Sonega	5	0	0	0	0	0	0	0	0	2	0	6
VNTelha	12	13	5	0	0	1	0	0	1	0	0	0

Table A.6: Number of one hour exceedances of the Information Alert of O₃ ($180\mu\text{g}/\text{m}^3$), for all stations (2005-2016).

A.2 Air Pollutants Data Preprocessing

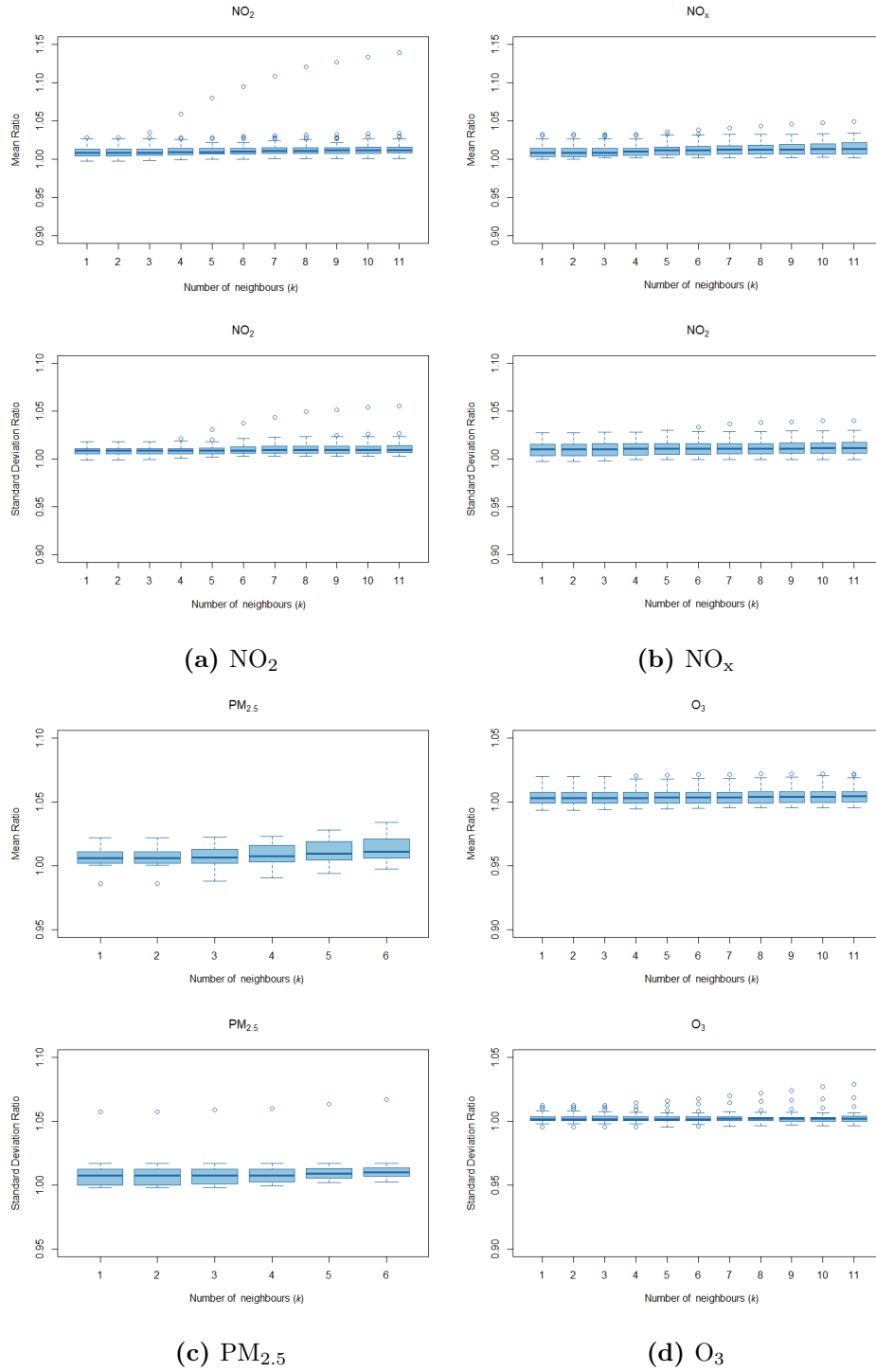
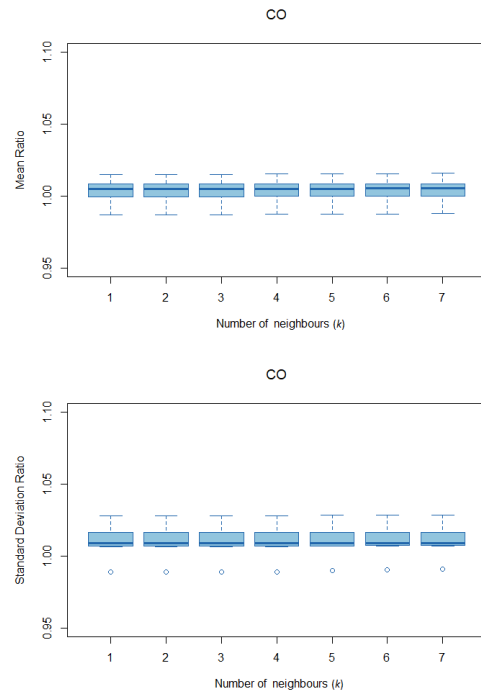


Figure A.1: Boxplot of the ratio between the mean of the imputed series and the mean of the series with missing data (upper panel). Same representation for the standard deviation (lower panel). (a) NO_2 , (b) NO_x , (c) PM_{10} , (d) O_3 , (e) CO . (To be continued)



(e) CO

Figure A.1: Boxplot of the ratio between the mean of the imputed series and the mean of the series with missing data (upper panel). Same representation for the standard deviation (lower panel). (a) NO₂, (b) NO_x, (c) PM₁₀, (d) O₃, (e) CO. (Continued)

Appendix B

Methods

B.1 ARFIMA estimation

Here is presented the algorithm to determine the h' value closest to h , for which the Hessian matrix is stable and is possible to compute the standard error of the coefficients. The idea is to search a neighbourhood of the finite difference interval (h) and try to find a value for which it is possible to estimate the Hessian. The algorithm is as follows:

1. For each time series check whether the h values allows the estimation of the standard error or not;
2. If the standard error is not computed, define a vector $k = (\frac{1}{n}, \dots, \frac{1}{1})$, $n \in \mathbb{N}$;
3. Re-compute the Hessian for all $j = h \times k$;
4. Select $h' = \min\{dist(h - j)\}$.

When the algorithm stops, it may occur that no j value allowed the computation of the Hessian, and as a consequence, the standard error is not computed. Nevertheless, the user may try to increase k , that is, decreasing the value of the h' parameter. Nevertheless, for the several k tested (up to 100 000), for quite a few time series it was not possible to compute any value which allowed the computation of the standard error.

Table B.1 shows the coefficients of an ARFIMA(4, 0.478, 1). In particular, this model characterises NO_x Arcos' Station. The values of the coefficients are not affected by the h value, as this only impacts the computation of the Hessian, and as a consequence the standard error, the z -value and the p -value. The h parameter of the model is $h = 4.282 \times 10^{-3}$, however, this value does not allow for the computation of the standard error. Using the algorithm with $k = 10$, it is possible to obtain the standard error and the remaining parameters for $h' = 4.282 \times 10^{-4}$. Therefore, decreasing h by a factor of 10 allows the standard error computation, in this case. Nevertheless, as one can see in the Table B.1, the standard error values are substantially smaller, particularly for the AR1 and AR4 coefficients. As a consequence, z -values are incredibly high, which results in all coefficients being significant. As the standard errors are so small compared to the respective coefficient, we do not believe that these values are reliable.

NO _x Arcos	Coefficient	Standard Error	<i>z-value</i>	<i>p-value</i>
d	0.478	3.296×10^{-4}	1.450×10^3	<0.001
ϕ_1	1.051	1.533×10^{-8}	6.855×10^7	<0.001
ϕ_2	-0.282	4.710×10^{-3}	-5.987×10^1	<0.001
ϕ_3	-0.026	3.225×10^{-3}	-8.243	<0.001
ϕ_4	-0.047	1.165×10^{-7}	-4.055×10^5	<0.001
θ_1	0.750	3.334×10^{-3}	-2.250×10^2	<0.001

Original value $h = 4.282 \times 10^{-3}$ and $h' = 4.282 \times 10^{-4}$.

Table B.1: Model coefficients, standard error, *z-value* and *p-value* for NO_x at Arcos Station.

Additionally, we further suspected on the reliability of the standard error, when assessing SARFIMA models. Tables B.2 and B.3 shows the coefficients, standard error, *z-value* and *p-value* of the ARFIMA part of the SARFIMA models. For Custóias Station the default h value does not allow to compute the standard error. On the contrary, for the Chamusca Station, the default h can achieve the standard error. The models of these stations have the same value for the fractional differencing parameter and similar magnitude for the remaining coefficients. But, while the Custóias Station the coefficient of the d parameter is non significant ($p - value > 0.05$), for the Chamusca Station this parameter is significant ($p - value < 0.05$). Furthermore, the standard error of d at Chamusca is quite smaller than the standard error of this parameter at Custóias Station. Taking everything into account, the computation of the standard error seems to be subject to an undetermined uncertainty. Therefore, we decided not to make any conclusions regarding the significance of the coefficients. The models are written taking all coefficients into consideration, and the fitted values will be the result of all coefficients. Thus, even if we cannot establish if a coefficient is non-significant, and this is included in the model, its effect on the estimates will be negligible.

NO ₂	Custóias	Coefficient	Standard Error	<i>z-value</i>	<i>p-value</i>
	d	4.583×10^{-5}	1.872×10^{-4}	0.245	0.807
	ϕ_1	1.460	2.960×10^{-8}	4.933×10^7	<0.001
	ϕ_2	-0.380	3.241×10^{-3}	-1.172×10^2	<0.001
	ϕ_3	-0.070	5.807×10^{-3}	-1.211×10^1	<0.001
	ϕ_4	-0.026	3.583×10^{-7}	-7.225×10^4	<0.001
	θ_1	0.902	3.401×10^{-3}	2.651×10^2	<0.001

The original value $h = 4.127 \times 10^{-3}$ and $h' = 4.127 \times 10^{-6}$.

Table B.2: Model coefficients, standard error, *z-value* and *p-value* for NO₂ at Custóias Station.

NO ₂	Chamusca	Coefficient	Standard Error	<i>z-value</i>	<i>p-value</i>
	d	4.583×10^{-5}	1.497×10^{-9}	3.061×10^4	<0.001
	ϕ_1	1.590	1.607×10^{-3}	9.891×10^2	<0.001
	ϕ_2	-0.655	8.598	-7.620×10^2	<0.001
	ϕ_3	-0.043	1.914×10^{-3}	-2.245×10^1	<0.001
	θ_1	0.837	1.400×10^{-3}	5.982×10^2	<0.001

The original value $h = 2.179 \times 10^{-3}$, for which is possible to compute the standard error.

Table B.3: Model coefficients, standard error, *z-value* and *p-value* for NO₂ at Chamusca Station.

Appendix C

Results

C.1 Air Pollutants Descriptive Statistics

Station	NO ₂ ($\bar{x} \pm \sigma$) $\mu g/m^3$	NO _x ($\bar{x} \pm \sigma$) $\mu g/m^3$	PM ₁₀ ($\bar{x} \pm \sigma$) $\mu g/m^3$	PM _{2.5} ($\bar{x} \pm \sigma$) $\mu g/m^3$	O ₃ ($\bar{x} \pm \sigma$) $\mu g/m^3$	SO ₂ ($\bar{x} \pm \sigma$) $\mu g/m^3$	CO ($\bar{x} \pm \sigma$) g/m^3
Arcos	17.07 \pm 13.55	22.79 \pm 25.59			60.94 \pm 28.70		0.23 \pm 0.13
Aveiro	26.42 \pm 18.69	38.37 \pm 46.78	33.55 \pm 24.03				0.28 \pm 0.26
Av. Liberdade	62.37 \pm 37.01	142.60 \pm 126.01	40.47 \pm 23.59				0.40 \pm 0.29
Beato	24.69 \pm 20.73	33.59 \pm 44.10			54.75 \pm 27.60		
Chamusca	6.47 \pm 3.75	7.74 \pm 4.36	17.72 \pm 13.75	9.49 \pm 8.55	71.13 \pm 25.58		
Custóias	27.88 \pm 25.10	42.54 \pm 49.09			45.99 \pm 27.36		
Entrecampos	43.71 \pm 30.62	81.62 \pm 95.45	30.18 \pm 20.13	14.53 \pm 11.54	45.44 \pm 27.90	1.49 \pm 3.17	0.34 \pm 0.26
Ermesinde	29.64 \pm 24.64	45.89 \pm 56.15	30.17 \pm 22.74		44.11 \pm 29.31		
Ervedeira	6.91 \pm 6.42	6.99 \pm 6.10	22.02 \pm 18.28	12.49 \pm 15.88	56.69 \pm 28.16	3.71 \pm 6.61	
Escavadeira	21.48 \pm 17.50	35.73 \pm 43.48	27.37 \pm 20.43		57.96 \pm 29.10	4.97 \pm 13.36	
Estarreja	18.49 \pm 15.74	32.77 \pm 36.76	30.85 \pm 25.11	18.00 \pm 19.41	45.97 \pm 32.57	6.28 \pm 11.13	
Fornelo Monte	2.60 \pm 3.31		12.23 \pm 13.17		74.82 \pm 23.92		
Fr. Bartolomeu	45.18 \pm 27.72	116.57 \pm 115.07	30.74 \pm 31.86				
Fr. Sá Carneiro	50.28 \pm 33.21	107.42 \pm 100.50	28.72 \pm 21.70				0.53 \pm 0.36
Frossos	14.00 \pm 12.34	27.20 \pm 39.98	20.70 \pm 20.03		41.47 \pm 33.00		
Fundão	5.85 \pm 3.57	3.10 \pm 4.07	14.09 \pm 14.06	6.83 \pm 7.15	64.96 \pm 28.29	0.93 \pm 1.78	
Ílhavo	10.11 \pm 9.27	10.96 \pm 13.68	25.53 \pm 18.80		50.10 \pm 30.35	1.46 \pm 3.24	
Instituto Geofísico	15.66 \pm 13.73	19.58 \pm 22.48	21.24 \pm 18.38		49.51 \pm 27.42		
Laranjeiro	27.97 \pm 24.21	43.31 \pm 67.36	25.25 μm 17.20		56.02 \pm 28.58		0.27 \pm 0.22
Lavradio						7.14 \pm 33.56	
Loures	21.37 \pm 17.53	33.92 \pm 49.77	24.21 \pm 16.15		55.43 \pm 30.92		
Meco			30.41 \pm 25.59		49.47 \pm 28.06		
Mem Martins	14.06 \pm 14.82	18.47 \pm 24.82	21.97 \pm 14.45		67.07 \pm 25.91	0.53 \pm 1.33	
Monte Chãos	4.56 \pm 4.80				69.57 \pm 22.50		
Monte Velho			23.44 \pm 16.71				
Olivais	32.83 \pm 26.85	55.99 \pm 84.09	25.11 \pm 16.55	12.63 \pm 10.32	51.14 \pm 29.30	0.92 \pm 2.16	0.28 \pm 0.24
Paços Ferreira					40.61 \pm 30.22		
Paio Pires	23.68 \pm 17.18	36.33 \pm 38.68			52.91 \pm 28.50	1.73 \pm 6.48	
Pe Moreira Neves	26.98 \pm 17.10	45.18 \pm 45.20					
Quebedo	24.94 \pm 17.58	40.61 \pm 46.37	28.24 \pm 17.76			0.63 \pm 2.83	0.28 \pm 0.20
Quinta Marquês		21.10 \pm 29.37					
Restelo	23.52 \pm 19.45	30.70 \pm 34.46			58.43 \pm 27.78		
Sonega	3.69 \pm 3.32	4.87 \pm 3.94			59.95 \pm 24.34	5.42 \pm 11.72	
Terena				10.81 \pm 12.88	45.15 \pm 22.50		
Vermoim	28.81 \pm 22.48	47.20 \pm 59.16	27.19 \pm 30.37				
VNTelha					45.84 \pm 27.99		

Table C.1: Mean (\bar{x}) and standard deviation (σ) of each pollutant time series.

C.2 ARFIMA Models

Station	NO ₂		NO _x		PM ₁₀		PM _{2.5}		O ₃		SO ₂		CO	
	R/S	Corrected	R/S	Corrected	R/S	Corrected	R/S	Corrected	R/S	Corrected	R/S	Corrected	R/S	Corrected
	Method	R/S Method	Method	R/S Method	Method	R/S Method	Method	R/S Method	Method	R/S Method	Method	R/S Method	Method	R/S Method
Arcos	0.79	0.84	0.78	0.83					0.75	0.82			0.76	0.82
Aveiro	0.77	0.82	0.78	0.82	0.78	0.83							0.81	0.84
Av. Liberdade	0.78	0.79	0.78	0.82	0.81	0.84							0.82	0.86
Beato	0.79	0.83	0.76	0.81					0.76	0.83				
Chamusca	0.80	0.84	0.81	0.86	0.80	0.81	0.79	0.82	0.74	0.81				
Custóias	0.81	0.89	0.80	0.86	0.81	0.87			0.74	0.80				
Entrecampos	0.78	0.81	0.76	0.80	0.83	0.84	0.82	0.84	0.77	0.85	0.78	0.87	0.80	0.84
Ermesinde	0.77	0.81	0.75	0.80	0.82	0.85			0.72	0.79				
Ervedeira	0.84	0.95	0.79	0.92	0.81	0.85	0.82	0.89	0.76	0.83	0.74	0.84		
Escavadeira	0.82	0.83	0.80	0.81	0.85	0.85			0.76	0.83	0.80	0.88		
Estarreja	0.79	0.86	0.77	0.83	0.79	0.84	0.80	0.86	0.78	0.85	0.81	0.96		
Fornelo Monte	0.82	0.95			0.77	0.78			0.78	0.84				
Fr. Bartolomeu	0.77	0.84	0.72	0.82	0.83	0.90								
Fr. Sá Carneiro	0.76	0.82	0.76	0.82	0.81	0.85							0.81	0.88
Frossos	0.81	0.88	0.79	0.86	0.82	0.88			0.76	0.83				
Fundão	0.79	0.91	0.82	0.94	0.81	0.84	0.80	0.81	0.72	0.80	0.84	0.91		
Ílhavo	0.82	0.95	0.82	0.93	0.73	0.80			0.76	0.80	0.84	1.03		
Instituto Geofísico	0.82	0.89	0.80	0.86	0.80	0.83			0.76	0.84				
Laranjeiro	0.76	0.80	0.74	0.78	0.79	0.81			0.76	0.82			0.76	0.80
Lavradio											0.80	0.86		
Loures	0.78	0.85	0.79	0.85	0.82	0.84			0.76	0.83				
Meco					0.77	0.82			0.76	0.82				
Mem Martins	0.76	0.83	0.73	0.80	0.80	0.81			0.72	0.80	0.84	0.94		
Monte Chãos	0.83	0.96							0.83	0.89				
Monte Velho					0.79	0.82								
Olivais	0.77	0.83	0.74	0.80	0.79	0.82	0.77	0.81	0.75	0.83	0.81	0.92	0.77	0.82
Paços Ferreira									0.80	0.86				
Paio Pires	0.80	0.84	0.78	0.83							0.80	0.83		
Pe Moreira Neves	0.76	0.81	0.74	0.82										
Quebedo	0.84	0.88	0.84	0.87	0.82	0.85					0.82	1.01	0.80	0.85
Quinta Marquês			0.81	0.85										
Restelo	0.74	0.80	0.75	0.80					0.76	0.84				
Sonega	0.83	0.97	0.83	0.95					0.87	0.94	0.76	0.90		
Terena							0.74	0.80	0.73	0.80				
Vermoim	0.78	0.81	0.75	0.80	0.80	0.86								
VNTelha									0.75	0.83				

Table C.2: Hurst Exponent for all pollutants time series at each station.

C.3 SARFIMA Models

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_1	θ_1	θ_1	θ_1	Φ_1	Θ_1	\bar{x}	Model
Arcos	0.000	1.356	-0.251	-0.126			0.618	0.234				0.964	-0.786	17.379	(3, 0.000, 2)(1, 0, 1)[24]
Aveiro	0.291	0.176	0.081				-0.324	-0.077	-0.010			0.971	-0.812	26.637	(2, 0.291, 3)(1, 0, 1)[24]
Av. Liberdade	0.009	0.847					0.028	-0.014	-0.008			0.784	-0.392	62.607	(1, 0.009, 3)(1, 0, 1)[24]
Beato	0.159	0.607	0.058	0.029	0.005	0.010						0.950	-0.756	29.977	(5, 0.159, 0)(1, 0, 1)[24]
Chamusca	0.000	1.590	-0.655	-0.043			0.837					0.956	-0.810	6.607	(3, 0.000, 1)(1, 0, 1)[24]
Custóias	0.000	1.460	-0.380	-0.070	-0.026		0.902					0.948	-0.714	27.572	(5, 0.000, 0)(1, 0, 1)[24]
Entrecampos	0.000	0.913	-0.074	0.012	-0.016	0.032						0.896	-0.622	44.220	(5, 0.000, 0)(1, 0, 1)[24]
Ermesinde	0.190	0.358	0.235				-0.207	0.068	0.019			0.930	-0.686	29.018	(2, 0.190, 3)(1, 0, 1)[24]
Ervedeira	0.000	1.692	-0.701	-0.010	-0.009		0.928					0.982	-0.828	7.034	(5, 0.000, 0)(1, 0, 1)[24]
Escavadeira	0.000	1.610	-0.630	-0.032	0.006		0.883					0.952	-0.772	21.801	(4, 0.000, 1)(1, 0, 1)[24]
Estarreja	0.267	0.238	0.144	0.007			-0.295					0.935	-0.651	18.534	(3, 0.267, 1)(1, 0, 1)[24]
Fornelo Monte	0.000	1.672	-0.685				1.228	-0.271	-0.027			0.952	-0.781	2.784	(2, 0.000, 3)(1, 0, 1)[24]
Fr. Bartolomeu	0.342	0.143	0.040				-0.118	-0.053	-0.041			0.958	-0.732	43.664	(2, 0.342, 3)(1, 0, 1)[24]
Fr. Sá Carneiro	0.272	0.927	-0.125	-0.047	-0.020		0.535					0.975	-0.816	47.868	(4, 0.272, 1)(1, 0, 1)[24]
Frossos	0.221	0.348	0.170	0.002	-0.003		-0.244					0.933	-0.602	13.927	(4, 0.221, 1)(1, 0, 1)[24]
Fundão	0.000	1.503	-0.430	0.053	-0.034		0.909					0.940	-0.604	5.810	(4, 0.000, 1)(1, 0, 1)[24]
Ílhavo	0.000	1.606	-0.625				0.814	0.048				0.951	-0.678	9.979	(2, 0.000, 2)(1, 0, 1)[24]
Instituto Geofísico	0.219	0.546					0.031	-0.040	-0.036	-0.022		0.974	-0.798	15.520	(1, 0.219, 4)(1, 0, 1)[24]
Laranjeiro	0.111	0.776	0.003	0.005	-0.010	0.008						0.927	-0.699	28.126	(5, 0.111, 0)(1, 0, 1)[24]
Loures	0.000	1.560	-0.541	-0.040	-0.004		0.880					0.976	-0.810	21.604	(4, 0.000, 1)(1, 0, 1)[24]
Mem Martins	0.283	0.200	0.250				-0.354	0.027	0.010			0.960	-0.783	14.199	(2, 0.238, 3)(1, 0, 1)[24]
Monte Chãos	0.000	1.483	-0.537	0.045	-0.012		0.889					0.979	-0.877	4.523	(4, 0.000, 1)(1, 0, 1)[24]
Olivais	0.156	0.732	-0.022	0.025	-0.012	0.010						0.952	-0.740	32.982	(5, 0.156, 0)(1, 0, 1)[24]
Paio Pires	0.093	0.861					0.171	0.057	0.050	0.055		0.922	-0.649	23.796	(1, 0.093, 4)(1, 0, 1)[24]
Pe Moreria Neves	0.152	0.730					0.227	0.011	0.012	0.010		0.954	-0.692	26.448	(1, 0.152, 4)(1, 0, 1)[24]
Quebedo	0.000	1.644	-0.0692	0.022	0.009		0.869					-0.976	-0.788	25.595	(4, 0.000, 1)(1, 0, 1)[24]
Restelo	0.000	1.704	-0.790	0.053	0.014		0.847					0.917	-0.668	23.682	(4, 0.000, 1)(1, 0, 1)[24]
Sonoga	0.000	1.613	-0.626				0.995	-0.175	-0.051			0.957	-0.781	3.655	(2, 0.000, 3)(1, 0, 1)[24]
Vermoim	0.151	0.652	0.052	-0.001	-0.001	0.008						0.918	-0.635	28.504	(5, 0.151, 0)(1, 0, 1)[24]

Table C.3: SARFIMA models coefficients for each station - NO₂.

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_2	θ_3	θ_4	θ_5	Φ_1	Θ_1	\bar{x}	Model
Arcos	0.311	0.423	0.360	-0.180			0.038	0.384				0.983	-0.866	23.230	(3, 0.311, 2)(1, 0, 1)[24]
Aveiro	0.000	1.417	-0.444				0.674	0.160	0.046			0.979	-0.854	38.651	(2, 0.000, 3)(1, 0, 1)[24]
Av. Liberdade	0.143	0.679	0.040	0.003	0.004	-0.008						0.923	-0.697	142.850	(5, 0.143, 0)(1, 0, 1)[24]
Beato	0.160	0.381	0.207	0.004	0.009		-0.268					0.978	-0.878	33.880	(5, 0.160, 0)(1, 0, 1)[24]
Chamusca	0.000	1.543	-0.624	0.047	0.008		0.829					0.970	-0.848	7.872	(5, 0.000, 0)(1, 0, 1)[24]
Custóias	0.259	0.172	0.112				-0.111	-0.021	-0.033			0.956	-0.775	41.708	(2, 0.259, 3)(1, 0, 1)[24]
Entrecampos	0.329	0.286	0.158	-0.023	-0.016		-0.316					0.979	-0.869	82.423	(5, 0.329, 0)(1, 0, 1)[24]
Ermesinde	0.196	0.461					-0.085	-0.069	-0.050	-0.029		0.963	-0.814	45.069	(1, 0.196, 4)(1, 0, 1)[24]
Ervedeira	0.000	1.664	-0.658	-0.032	0.015		0.925					0.981	-0.827	7.167	(5, 0.000, 0)(1, 0, 1)[24]
Escavadeira	0.195	0.294	0.164				-0.166	0.059	0.003			0.981	-0.892	35.587	(2, 0.0195, 3)(1, 0, 1)[24]
Estarreja	0.402	0.944	-0.224	-0.003	-0.038		0.560					0.948	-0.670	32.856	(4, 0.402, 1)(1, 0, 1)[24]
Fr. Bartolomeu	0.342	0.869	-0.140	-0.050	-0.025		0.596					0.967	-0.772	113.215	(4, 0.342, 1)(1, 0, 1)[24]
Fr. Sá Carneiro	0.190	0.205					-0.266	-0.176	-0.087	-0.021		0.989	-0.876	102.056	(1, 0.190, 4)(1, 0, 1)[24]
Frossos	0.071	0.829					0.121	0.006	0.006	0.020		0.872	-0.388	27.305	(1, 0.071, 4)(1, 0, 1)[24]
Fundão	0.000	1.476	-0.402	0.057	-0.034		0.909					0.934	-0.620	3.632	(4, 0.000, 1)(1, 0, 1)[24]
Ílhavo	0.000	1.588	-0.608				0.776	0.070				0.941	-0.634	10.992	(2, 0.000, 2)(1, 0, 1)[24]
Instituto Geofísico	0.204	0.173	0.070	0.065			-0.370	-0.124				0.979	-0.845	19.600	(3, 0.204, 2)(1, 0, 1)[24]
Laranjeiro	0.105	0.768	-0.015	-0.006	0.004	-0.008						0.952	-0.813	43.610	(5, 0.105, 0)(1, 0, 1)[24]
Loures	0.278	0.188	0.108	-0.002	-0.020		-0.215					0.983	-0.850	34.394	(4, 0.278, 1)(1, 0, 1)[24]
Mem Martins	0.296	0.238	0.108	-0.016	-0.006		-0.282					0.981	-0.875	18.731	(4, 0.296, 1)(1, 0, 1)[24]
Olivais	0.157	0.329	0.216	-0.020	0.010		-0.413					0.970	-0.834	56.111	(4, 0.157, 1)(1, 0, 1)[24]
Paio Pires	0.253	0.329	0.166	-0.005	-0.014		-0.231					0.962	-0.782	36.451	(4, 0.253, 1)(1, 0, 1)[24]
Pe Moreria Neves	0.195	0.280	0.136	0.019	-0.006		-0.200					0.975	-0.778	43.636	(4, 0.195, 1)(1, 0, 1)[24]
Quebedo	0.309	0.440	0.396	-0.214			-0.024	0.420				0.989	-0.877	42.829	(3, 0.309, 2)(1, 0, 1)[24]
Quinta Marquês	0.306	0.218	0.141	0.007	-0.014		-0.157					0.970	-0.810	21.030	(4, 0.306, 1)(1, 0, 1)[24]
Restelo	0.000	1.744	-0.893	0.132			0.870					0.962	-0.814	30.952	(3, 0.000, 1)(1, 0, 1)[24]
Sonega	0.247	0.201	0.184				-0.040	0.144	0.008			0.970	-0.846	4.718	(2, 0.247, 3)(1, 0, 1)[24]
Vermoim	0.199	0.623	0.019	-0.011	-0.019	-0.012						0.945	-0.733	46.634	(5, 0.199, 0)(1, 0, 1)[24]

Table C.4: SARFIMA models coefficients for each station - NO_x.

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_2	θ_3	θ_4	θ_5	sAR1	Θ_1	\bar{x}	Model
Aveiro	0.000	0.487	0.412	-0.001	-0.050		-0.382					0.873	-0.512	34.465	(4, 0.000, 1)(1, 0, 1)[24]
Av. Liberdade	0.012	0.334	0.517				-0.612	-0.037	-0.027			0.743	-0.331	41.358	(2, 0.012, 3)(1, 0, 1)[24]
Chamusca	0.189	0.791	0.004	-0.014	0.072	0.018						0.746	-0.309	17.914	(5, 0.189, 0)(1, 0, 1)[24]
Custóias	0.000	1.285	-0.236	-0.001	-0.071		0.850					0.871	-0.553	27.406	(4, 0.000, 1)(1, 0, 1)[24]
Entrecampos	0.013	0.898					0.003	-0.012	-0.024	-0.027		0.867	-0.468	27.851	(1, 0.013, 4)(1, 0, 1)[24]
Ermesinde	0.000	0.909					0.244	-0.035	-0.018	-0.028		0.802	-0.422	30.575	(1, 0.000, 4)(1, 0, 1)[24]
Ervedeira	0.000	0.096	0.727	0.014			-0.652	0.099				0.801	0.356	30.018	(3, 0.000, 2)(1, 0, 1)[24]
Escavadeira	0.000	0.900					0.062	-0.031	-0.043			0.780	-0.394	22.307	(1, 0.000, 3)(1, 0, 1)[24]
Estarreja	0.000	0.885					0.094	-0.026	-0.029	-0.016		0.833	-0.494	27.991	(1, 0.000, 4)(1, 0, 1)[24]
Fornelo Monte	0.000	0.690	0.040	0.078	0.039	0.014						0.864	-0.387	31.212	(5, 0.000, 0)(1, 0, 1)[24]
Fr. Bartolomeu	0.000	1.381	-0.327	-0.078			0.917	-0.049				0.683	-0.233	12.133	(3, 0.000, 2)(1, 0, 1)[24]
Fr. Sá Carneiro	0.000	1.341	-0.390				0.643					0.942	-0.695	32.404	(2, 0.000, 1)(1, 0, 1)[24]
Frossos	0.122	0.944	-0.287	0.140			0.429	-0.213				0.835	-0.474	28.776	(3, 0.122, 2)(1, 0, 1)[24]
Fundão	0.001	0.661	0.242	-0.039	0.020		-0.308					0.938	-0.661	20.828	(4, 0.001, 1)(1, 0, 1)[24]
Ílhavo	0.000	0.908					0.191	-0.042	-0.026	-0.028		0.717	-0.261	14.131	(1, 0.000, 4)(1, 0, 1)[24]
Instituto Geofísico	0.008	0.902					0.171	0.166	0.086	-0.008		0.849	-0.424	25.898	(1, 0.008, 4)(1, 0, 1)[24]
Laranjeiro	0.028	0.321	0.478	0.050			-0.535	-0.016				0.814	-0.475	21.541	(3, 0.028, 2)(1, 0, 1)[24]
Loures	0.000	0.187	0.584	0.070			-0.506	-0.068				0.757	-0.376	25.537	(3, 0.000, 2)(1, 0, 1)[24]
Meco	0.000	1.534	-0.562				0.697	0.228	-0.103			0.816	-0.428	24.197	(2, 0.000, 3)(1, 0, 1)[24]
Mem Martins	0.000	0.905					0.187	-0.019	-0.008	-0.022		0.842	-0.565	31.198	(1, 0.000, 4)(1, 0, 1)[24]
Monte Velho	0.036	1.144	-0.200	0.022	-0.018		0.790					0.798	-0.409	22.054	(4, 0.036, 1)(1, 0, 1)[24]
Olivais	0.000	0.906					0.092	0.021				0.796	-0.544	23.711	(1, 0.000, 2)(1, 0, 1)[24]
Quebedo	0.000	0.204	0.911	-0.236			-0.573	0.418				0.768	-0.384	25.477	(3, 0.000, 2)(1, 0, 1)[24]
Vermoim	0.000	1.667	-0.809	0.195	-0.075		0.845					0.810	-0.359	28.335	(5, 0.000, 0)(1, 0, 1)[24]

Table C.5: SARFIMA models coefficients for each station - PM₁₀.

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_2	θ_3	θ_4	θ_5	Φ_1	Θ_1	\bar{x}	Model
Chamusca	0.002	0.916					0.066	-0.008	0.089	-0.032		0.746	-0.328	8.925	(1, 0.002, 4)(1, 0, 1)[24]
Entrecampos	0.014	0.463	0.286	0.061	0.036		-0.311					0.819	-0.456	13.866	(4, 0.014, 1)(1, 0, 1)[24]
Ervedeira	0.000	1.358	-0.255	-0.647	-0.061		0.822					0.879	-0.528	12.522	(4, 0.000, 1)(1, 0, 1)[24]
Estarreja	0.048	0.521	0.198	0.084	0.019	-0.013						0.896	-0.456	17.178	(5, 0.048, 4)(1, 0, 1)[24]
Fundão	0.011	1.33	-0.303	-0.035	-0.028		0.742					0.760	-0.392	6.436	(4, 0.011, 1)(1, 0, 1)[24]
Olivais	0.000	1.432	-0.442	0.002	-0.026		0.723					0.812	-0.453	11.896	(4, 0.000, 1)(1, 0, 1)[24]
Terena	0.250	0.798	-0.024				0.779	-0.070				0.930	-0.750	0.109	(2, 0.250, 2)(1, 0, 1)[24]

Table C.6: SARFIMA models coefficients for each station - PM_{2.5}.

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_2	θ_3	θ_4	θ_5	Φ_1	Θ_1	\bar{x}	Model
Arcos	0.000	0.888	-0.047	0.017	0.020							0.946	-0.637	61.134	(4, 0.000, 0)(1, 0, 1)[24]
Beato	0.084	0.843					0.152	0.023	0.002	0.000		0.946	-0.661	54.367	(1, 0.084, 4)(1, 0, 1)[24]
Chamusca	0.000	0.899					-0.033	0.012	0.006	0.001		0.878	-0.364	71.127	(1, 0.000, 4)(1, 0, 1)[24]
Custóias	0.101	0.248	0.455				-0.458	0.048	-0.016			0.941	-0.641	42.249	(2, 0.101, 3)(1, 0, 1)[24]
Entrecampos	0.152	0.803					0.300	0.004	-0.001			0.964	-0.744	44.636	(1, 0.152, 3)(1, 0, 1)[24]
Ermesinde	0.147	0.402	0.296				-0.253	0.028	-0.016			0.952	-0.665	43.273	(2, 0.147, 3)(1, 0, 1)[24]
Ervedeira	0.142	1.562	-0.630				0.687	0.012	-0.054			0.952	0.623	57.055	(2, 0.142, 3)(1, 0, 1)[24]
Escavadeira	0.098	0.670	0.070	0.018	0.005	0.017						0.938	-0.625	57.434	(5, 0.098, 0)(1, 0, 1)[24]
Estarreja	0.089	0.778					0.063	-0.021	-0.029	-0.019		0.957	-0.619	42.278	(1, 0.089, 4)(1, 0, 1)[24]
Fornelo Monte	0.000	1.095	-0.124	-0.028			0.562					0.816	-0.371	75.243	(3, 0.000, 1)(1, 0, 1)[24]
Frossos	0.127	1.560	-0.631				0.792	-0.080	-0.021			0.972	-0.679	41.467	(2, 0.127, 3)(1, 0, 1)[24]
Fundão	0.095	0.293	0.444				-0.499	0.016	-0.006			0.960	-0.650	65.340	(2, 0.095, 3)(1, 0, 1)[24]
Ílhavo	0.052	1.044	-0.176	0.009	-0.021							0.952	-0.638	50.656	(4, 0.052, 0)(1, 0, 1)[24]
Instituto Geofísico	0.083	0.809					-0.040	-0.036	-0.023	-0.020		0.925	-0.531	49.480	(1, 0.083, 4)(1, 0, 1)[24]
Laranjeiro	0.075	0.837					0.095	0.018	0.008	-0.004		0.939	-0.641	55.269	(1, 0.075, 4)(1, 0, 1)[24]
Loures	0.039	0.768	0.082	0.004	-0.002							0.943	-0.627	55.240	(4, 0.039, 0)(1, 0, 1)[24]
Meco	0.092	0.808					0.043	-0.020	-0.008	-0.026		0.954	-0.680	48.790	(1, 0.092, 4)(1, 0, 1)[24]
Mem Martins	0.000	0.878	-0.011	0.016	-0.018	0.021						0.878	-0.482	67.090	(5, 0.000, 0)(1, 0, 1)[24]
Monte Chãos	0.000	0.333	0.539				-0.167	0.255				0.897	-0.509	67.637	(2, 0.000, 2)(1, 0, 1)[24]
Olivais	0.138	0.396	0.319				-0.245	0.062	-0.003			0.958	-0.695	50.617	(2, 0.138, 3)(1, 0, 1)[24]
Paços Ferreira	0.197	1.704	-0.837	0.063			1.216	-0.397				0.968	-0.684	41.649	(3, 0.197, 2)(1, 0, 1)[24]
Paio Pires	0.000	0.853	0.017	-0.001	-0.013	0.016						0.942	-0.614	52.571	(5, 0.000, 0)(1, 0, 1)[24]
Restelo	0.140	0.711	0.026	0.018	0.008	0.004						0.936	-0.633	58.771	(5, 0.140, 0)(1, 0, 1)[24]
Sonega	0.000	0.928					0.376	0.030	0.034			0.935	-0.501	57.868	(1, 0.000, 3)(1, 0, 1)[24]
Terena	0.145	0.810					0.210	-0.003	0.009			0.972	-0.734	45.824	(1, 0.145, 3)(1, 0, 1)[24]
VNTelha	0.130	0.398	0.245	0.064			-0.221	-0.023				0.954	-0.668	45.149	(3, 0.130, 2)(1, 0, 1)[24]

Table C.7: SARFIMA models coefficients for each station - O_3 .

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_2	θ_3	θ_4	θ_5	Φ_1	Θ_1	\bar{x}	Model
Entrecampos	0.173	0.340	0.123	0.039	0.024		-0.241					0.969	-0.876	1.525	(4, 0.173, 1)(1, 0, 1)[24]
Ervedeira	0.089						-0.570	-0.391	-0.255	-0.203	-0.096	0.972	-0.862	3.671	(0, 0.089, 5)(1, 0, 1)[24]
Escavadeira	0.161	0.352	-0.016	0.022	0.005	0.039						0.980	-0.912	5.300	(5, 0.161, 0)(1, 0, 1)[24]
Estarreja	0.000	1.550	-0.568				0.916					0.933	-0.700	6.446	(2, 0.000, 1)(1, 0, 1)[24]
Fundão	0.132	1.080	-0.229				0.480					0.981	-0.873	0.973	(2, 0.132, 1)(1, 0, 1)[24]
Ílhavo	0.002	0.545	0.213				-0.159	0.084	0.041			0.957	-0.754	1.623	(2, 0.002, 3)(1, 0, 1)[24]
Lavradio	0.337	0.083	-0.094				-0.102	-0.145	-0.025			0.942	-0.809	7.586	(2, 0.337, 3)(1, 0, 1)[24]
Mem Martins	0.009	1.608	-0.625			0.958	-0.036	-0.016			0.989	-0.914	0.565		
Olivais	0.142	0.281	0.237	0.014			-0.294	0.063				0.969	-0.863	0.919	(3, 0.142, 2)(1, 0, 1)[24]
Paio Pires	0.113	0.212	0.191				-0.451	-0.015	0.032			0.930	-0.780	1.817	(2, 0.113, 3)(1, 0, 1)[24]
Quebedo	0.083	0.543	-0.053	0.046	0.011							0.991	-0.934	0.672	(4, 0.083, 0)(1, 0, 1)[24]
Sonega	0.093	0.871	-0.783	0.333			0.343	-0.578				0.985	-0.902	5.406	(3, 0.093, 2)(1, 0, 1)[24]

Table C.8: SARFIMA models coefficients for each station - SO₂.

Station	d	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	θ_1	θ_1	θ_1	θ_1	Φ_1	Θ_1	\bar{x}	Model
Arcos	0.266	0.193	0.167				-0.273	0.062	0.033			0.938	-0.703	0.225	(2, 0.266, 3)(1, 0, 1)[24]
Aveiro	0.175	0.062	0.302	-0.002	-0.029		-0.484					0.980	-0.854	0.292	(4, 0.175, 1)(1, 0, 1)[24]
Av. Liberdade	0.128	0.749					0.084	0.010	-0.006	-0.030		0.973	-0.814	0.417	(1, 0.128, 4)(1, 0, 1)[24]
Entrecampos	0.195	0.700	-0.004	-0.002	0.007	-0.012						0.974	-0.833	0.342	(5, 0.195, 0)(1, 0, 1)[24]
Fr. Sá Carneiro	0.237	0.242	0.141	0.005	-0.015		-0.173					0.978	-0.811	0.542	(4, 0.237, 1)(1, 0, 1)[24]
Laranjeiro	0.093	0.768					-0.021	0.008	-0.020			0.935	-0.755	0.268	(1, 0.093, 3)(1, 0, 1)[24]
Olivais	0.000	0.248	0.432				-0.628	-0.008	0.029			0.967	-0.814	0.282	(2, 0.000, 3)(1, 0, 1)[24]
Quebedo	0.256	0.213	0.167				-0.315	0.040	0.010			0.971	-0.797	0.289	(2, 0.256, 3)(1, 0, 1)[24]

Table C.9: SARFIMA models coefficients for each station - CO.

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Arcos	-4.909	10.912	9.268	-119.94	132.724	0.533	49.8	95.4
Aveiro	-4.489	16.589	13.000			0.488	51.4	89.6
Av. Liberdade	-5.905	33.805	27.238	-82.470	102.859	0.435	46.9	72.0
Beato	-5.598	17.321	14.383	-133.038	149.977	0.576	44.7	91.2
Chamusca	-1.951	3.421	2.919	-118.827	127.728	0.442	37.0	71.8
Custóias	0.968	24.017	18.779	-207.47	239.758	0.681	54.4	89.4
Entrecampos	-7.544	27.154	22.851	-117.498	136.650	0.517	36.3	71.0
Ermesinde	-1.363	20.105	15.861	-80.326	105.285	0.547	57.3	91.7
Ervedeira	-1.778	4.338	3.412			0.485	65.1	92.8
Escavadeira	-6.367	13.983	11.989	-142.998	156.314	0.550	47.0	94.9
Estarreja	-5.483	11.683	10.111			0.546	51.6	94.4
Fornelo Monte	-1.229	2.648	2.039			0.732	56.0	98.4
Fr. Bartolomeu	2.779	24.713	19.158	-30.499	58.398	0.439	68.0	88.7
Fr. Sá Carneiro	22.751	48.217	36.065	-9.537	61.941	0.753	47.4	68.6
Frossos	-0.262	10.838	8.304	-89.896	116.687	0.596	57.1	89.3
Fundão	0.154	3.034	1.980	-13.875	33.1	0.341	79.6	93.3
Ílhavo	1.981	5.825	3.581	4.075	25.369	0.359	84.5	93.5
Instituto Geofísico	-0.607	11.22	8.284	-62.365	85.299	0.534	70.4	92.8
Laranjeiro	-5.126	20.767	16.367	-136.057	154.407	0.582	42.0	85.5
Loures	-5.002	15.750	13.375	-152.723	171.055	0.619	45.1	91.5
Mem Martins	-2.943	13.579	10.376	-195.318	214.45	0.731	39.5	89.2
Monte Chãos	-0.565	4.170	2.859			0.632	68.2	95.8
Olivais	-5.538	22.167	18.046	-103.842	122.499	0.547	45.2	86.5
Paio Pires	-4.621	14.785	12.193	-128.753	145.523	0.512	45.6	83.6
Pe Moreira Neves	-4.251	12.395	9.924	-47.121	59.575	0.375	72.4	95.8
Quebedo	-7.955	15.035	13.127	-150.867	162.278	0.513	41.8	84.6
Restelo	-5.006	16.908	14.09	-166.072	184.567	0.595	40.2	84.4
Sonega	0.547	3.164	2.041	-31.12	62.984	0.558	99.3	99.8
Vermoim	-14.802	14.927	14.802			0.519	31.9	98.6

Table C.10: SARFIMA performance measures for each station - NO₂.

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Arcos	-7.288	18.957	14.322	-135.406	146.97	0.617	78.0	96.4
Aveiro	-7.264	38.209	24.484			0.633	72.7	96.5
Av. Liberdade	-12.334	117.215	85.676	-169.937	193.601	0.600	48.8	87.7
Beato	-7.669	33.635	22.311	-138.411	153.384	0.659	57.6	94.7
Chamusca	-1.712	3.566	2.963	-57.657	67.376	0.376	44.5	84.2
Custóias	1.957	53.362	32.030	-140.57	167.607	0.768	73.0	93.0
Entrecampos	-15.164	81.455	57.166	-183.579	202.855	0.694	42.8	91.9
Ermesinde	-5.201	37.147	28.438	-118.122	140.1	0.631	72.4	94.3
Ervedeira	-2.313	3.948	3.416			0.477	51.0	94.1
Escavadeira	-13.141	29.920	22.446	-176.716	186.314	0.613	75.6	97.7
Estarreja	-10.243	23.853	19.507	-188.043	200.827	0.594	74.5	97.5
Fr. Bartolomeu	-10.368	82.402	65.145	-108.362	131.338	0.575	79.4	95.2
Fr. Sá Carneiro	31.587	132.342	99.819	-94.509	138.215	0.978	48.2	73.5
Frossos	-4.027	27.432	19.494	-153.726	174.025	0.714	79.2	94.3
Fundão	0.809	3.635	2.119	-18.271	50.976	0.584	82.7	92.9
Ílhavo	-2.066	7.515	5.812	-93.485	109.305	0.529	87.3	96.8
Instituto Geofísico	-2.925	18.948	12.562	0.641			73.0	95.5
Laranjeiro	-9.047	57.177	32.097	-211.047	225.582	0.736	56.7	95.1
Loures	-10.097	41.725	28.560	-249.651	265.452	0.830	84.8	95.0
Mem Martins	-5.687	24.018	15.519	-246.212	261.663	0.829	44.9	87.6
Olivais	-11.206	67.843	41.193	-161.127	176.562	0.734	70.5	93.9
Paio Pires	-6.566	33.168	22.244	-149.791	165.719	0.610	61.9	94.0
Pe Moreira Neves	-7.146	35.630	24.418	-78.079	93.38	0.560	83.2	93.9
Quebedo	-16.563	35.004	27.402	-228.087	237.639	0.655	73.9	94.9
Quinta Marquês	-5.085	22.168	15.644	-247.755	264.369	0.744	76.5	94.5
Restelo	-6.908	27.732	20.477	-189.835	206.697	0.662	49.7	93.8
Sonega	1.492	4.091	2.308	7.026	31.144	0.489	74.2	89.9
Vermoim	-24.432	24.678	24.433	-115.878	115.882	0.524	98.3	99.9

Table C.11: SARFIMA performance measures for each station - NO_x.

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Aveiro	-14.194	20.893	18.267	-194.855	200.887	0.530	22.9	57.8
Av. Liberdade	-12.191	19.635	16.922	-89.042	96.16	0.409	25.8	54.5
Chamusca	-3.460	12.152	8.910			0.497	24.0	51.5
Custóias	-14.516	19.294	16.960			0.619	65.7	99.9
Entrecampos	-8.208	14.691	12.421	-78.341	86.527	0.406	34.5	68.8
Ermesinde	-11.405	20.069	16.164			0.538	35.1	70.9
Ervedeira	-4.815	12.879	10.380			0.465	36.2	75.4
Escavadeira	-8.161	14.989	12.762	-103.254	112.625	0.456	29.4	65.9
Estarreja	-7.305	20.559	15.656			0.502	45.1	82.4
Fornelo Monte	-1.039	14.584	8.205			0.676	46.6	93.0
Fr. Bartolomeu	-14.912	17.591	16.096			0.497	58.7	100.0
Fr. Sá Carneiro	-11.571	18.315	15.558			0.541	43.6	69.0
Frossos	-5.403	13.874	11.759			0.565	54.1	95.1
Fundão	-0.709	13.187	8.445			0.598	32.0	70.0
Ílhavo	-6.686	16.413	12.175			0.470	38.2	78.7
Instituto Geofísico	-6.509	14.287	10.806			0.502	59.4	98.1
Laranjeiro	-6.220	14.150	11.526	-99.275	110.508	0.451	30.3	60.1
Loures	-6.363	12.865	10.618	-110.405	120.548	0.439	40.3	75.3
Meco	-14.49	20.090	17.247			0.553	48.9	79.4
Mem Martins	-5.508	10.970	8.951	-78.013	86.933	0.406	41.7	77.6
Monte Velho	-3.939	16.940	11.596			0.489	67.5	93.1
Olivais	-7.674	14.561	12.127	-130.385	139.946	0.476	33.3	64.4
Quebedo	-6.391	15.851	12.672	-102.115	114.321	0.447	39.2	71.7
Vermoin	-15.754	21.512	19.153			0.688	44.0	99.2

Table C.12: SARFIMA performance measures for each station - PM₁₀.

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Chamusca	-1.832	7.218	5.246			4.097	32.5	69.7
Entrecampos	-0.735	8.985	6.806			2.474	54.1	81.8
Ervedeira	-7.456	10.213	9.167			2.878	39.1	99.1
Estarreja								
Fundão	-1.981	7.141	5.020			3.328	36.3	93.1
Olivais	-2.627	8.693	6.804			2.586	44.6	86.2
Terena	0.884	9.731	7.455			1.460	57.6	91.2

Table C.13: SARFIMA performance measures for each station - PM_{2.5}.

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Arcos	-0.642	25.916	20.102	-92.571	114.729	0.329	51.8	73.5
Beato	3.322	25.589	20.394	-78.046	106.879	0.375	49.9	74.1
Chamusca	0.382	24.393	18.560	-13.670	31.899	0.261	44.7	67.2
Custóias	1.515	27.485	23.009			0.508	39.7	66.3
Entrecampos	2.814	25.964	21.295	-349.521	382.732	0.477	50.8	78.6
Ermesinde	0.876	27.960	23.255	-166.013	198.21	0.537	42.4	67.8
Ervedeira	-2.010	26.094	21.263			0.373	36.7	58.8
Escavadeira	3.506	26.954	21.358	-67.616	96.213	0.372	52.4	78.8
Estarreja	-6.129	29.447	24.151			0.522	48.9	71.2
Fornelo Monte	-3.431	22.839	16.507	-15.283	27.082	0.219	60.3	84.1
Frossos	0.269	31.924	27.912			0.673	27.7	55.1
Fundão	-3.767	25.802	20.619	-51.532	70.016	0.316	46.5	68.5
Ílhavo	-4.545	27.809	22.617			0.446	37.8	60.2
Instituto Geofísico	3.032	25.708	20.557	-69.323	99.031	0.415	42.5	66.6
Laranjeiro	2.660	27.195	21.951	-82.147	111.246	0.397	45.8	72.0
Loures	0.814	27.597	22.295	-112.043	140.015	0.404	47.2	73.0
Meco	1.636	27.382	22.937			0.470	38.7	64.0
Mem Martins	-0.853	23.113	17.587	-28.115	46.192	0.262	50.1	75.5
Monte Chãos	12.392	24.641	19.052	8.554	23.356	0.282	51.2	77.8
Olivais	3.318	27.309	22.141	-131.068	163.053	0.437	48.3	74.1
Paços Ferreira	-11.432	22.608	19.467	-164.115	178.179	0.467	55.1	91.9
Paio Pires	0.753	26.264	21.146	-131.283	158.295	0.402	46.5	70.3
Restelo	-3.915	25.252	19.837	-93.914	112.578	0.338	51.7	75.4
Sonega	21.351	32.228	25.937	18.508	30.850	0.448	35.5	61.0
Terena	-5.197	21.798	18.000			0.393	44.1	69.3
VNTelha	0.251	24.335	20.131	-87.81	116.374	0.446	49.4	77.2

Table C.14: SARFIMA performance measures for each station - O_3 .

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Entrecampos	−0.480	1.481	1.252			0.822	86.7	98.8
Ervedeira	−0.436	5.385	3.666			0.999	84.0	94.7
Escavadeira	−3.956	4.070	3.966			0.748	100.0	100.0
Estarreja	−3.586	5.100	4.495			0.697	98.1	99.6
Fundão	−0.443	1.111	0.905			0.930	73.0	95.1
Ílhavo	−1.482	1.582	1.536			0.946	99.5	99.8
Lavradio	−6.569	6.655	6.574			0.867	100.0	100.0
Mem Martins	−0.258	0.538	0.463			0.819	93.5	98.7
Olivais	−0.018	1.138	0.876			0.953	75.4	96.3
Paio Pires	−0.859	2.489	1.697			0.934	89.8	97.9
Quebedo	−0.266	5.698	0.717			1.068	95.4	99.4
Sonega	−0.003	7.534	3.632			0.672	86.0	99.6

Table C.15: SARFIMA performance measures for each station - SO₂.

Station	ME	RMSE	MAE	MPE	MAPE	MASE	PI 80%	PI 95%
Arcos	0.004	0.098	0.066	−12.762	30.587	0.291	77.0	93.8
Aveiro	−0.082	0.208	0.158			0.542	59.1	96.6
Av. Liberdade	−0.086	0.202	0.159	−52.519	62.075	0.381	60.5	96.1
Entrecampos	−0.038	0.232	0.157	−47.836	63.218	0.460	49.4	92.4
Fr. Sá Carneiro	−0.145	0.282	0.243	−81.890	92.484	0.448	56.3	97.0
Laranjeiro	0.000	0.141	0.075	−14.142	28.309	0.280	78.1	94.9
Olivais	−0.032	0.184	0.118	−45.344	58.903	0.420	66.1	94.9
Quebedo	−0.065	0.150	0.115	−61.136	69.671	0.399	63.8	94.3

Table C.16: SARFIMA performance measures for each station - CO.