UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

TECHNIQUES FOR MONITORING CITY AIR QUALITY

(«Μέθοδοι Επίβλεψης της Ποιότητας του Αέρα στο Αστικό Περιβάλλον»)

By

Dimitra Chalanouli

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

November 2015

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of many people. I wish to express my gratitude to my supervisor, Dr. Sotiris Bersimis who was abundantly helpful and offered invaluable assistance, support and guidance. Special thanks also to external agencies for their indispensable assistance and sharing data I needed. Finally I wish to express my love and gratitude to my beloved family for the understanding through the duration of my study.

ABSTRACT

In this thesis, alarming events of air pollution in the Attica region are being investigated, using ARIMA models and statistical control charts. The research focused on estimating, fitting and forecasting suitable ARIMA models by identifying specific patterns in the time series of pollutants and by using statistical process control techniques to detect possible harmful exceedances. The main objective of this study was to analyze and plot the residuals (forecast errors) of air pollutants, taking nitrogen dioxide (NO_2) and ozone (O_3) daily mean concentrations from 2010-2013 and from 8 different stations located throughout Attica as a case study. The percentage of missing data for each annual time series was around 10% on average, thus multiple imputation techniques were used as an initial step. Corrective actions were taken to monitor such autocorrelated processes including differencing the time series in order to achieve stationarity and remove trend from the data. It was proved that NO_2 and O_3 time series were correlated and only the O_3 time series showed regular peaks (seasonality). After implementing ARIMA models and checking the residuals for correlation and normality, one – step ahead forecasts were produced for NO_2 and O_3 concentrations. Forecast errors were studied and plotted in a $\bar{X}$ chart / MR chart for individual data as an aid to detect outliers. We were mainly interested in the large positive differences between the observed and predicted values (positive forecast errors). Statistical analysis showed that successive large "disturbances" were only occurred for O_3 concentrations at Liosia station, indicating an event that we should pay special attention to, while in all other stations the outliers were significantly low in number, indicating a well – estimated model, close enough to the actual concentrations of 2013.

ΠΕΡΙΛΗΨΗ

Ο κύριος στόχος αυτής της διπλωματικής εργασίας είναι να διερευνήσει τυχόν υπερβάσεις στις τιμές (συγκεντρώσεις) των ρύπων στον Αττικό ουρανό χρησιμοποιώντας μεθόδους ανάλυσης χρονοσειρών όπως για παράδειγμα ARIMA μοντέλα αλλά και διαγράμματα στατιστικού ελέγχου ποιότητας. Χρησιμοποιήθηκαν δεδομένα από δύο κυρίαρχους στην Ελλάδα ρύπους, το διοξείδιο του αζώτου (NO2) και το όζον (O3), όπως μετρήθηκαν από 8 διαφορετικούς σταθμούς στην Αττική. Η ιδιαιτερότητα των δεδομένων αυτών έγκειται στο ότι, πρώτον, είναι έντονο το φαινόμενο των ελλειπουσών τιμών (έως και 10% σε κάθε σταθμό ξεχωριστά) και δεύτερον, όλες οι μετρήσεις των ρύπων είναι αυτοσυσχετισμένες. Για την εξάλειψη του πρώτου χρησιμοποιήθηκαν 2 διαφορετικοί μέθοδοι για την «συμπλήρωση» των τιμών αυτών (multiple imputation) που αφορούν χρονοσειρές ενώ για την εξάλειψη του δεύτερου εφαρμόστηκαν ARIMA μοντέλα. Η τελική χρήση των μοντέλων αυτών βοήθησε, όχι μόνο στην εξάλειψη της αυτοσυσχέτισης, αλλά και της τάσης που εμφάνιζαν αρχικά τα δεδομένα των ρύπων. Έπειτα από την επιλογή των καλύτερα προσαρμοσμένων μοντέλων χρονοσειρών (διαγνωστικός έλεγχος μέσω διαγραμμάτων αυτοσυσχέτισης, Μέσου Τετραγωνικού Σφάλματος και κριτηρίου του Akaike) για δεδομένα από το 2010-2012 («ιστορικό» σύνολο δεδομένων), προχωρήσαμε σε προβλέψεις για τα δεδομένα των προαναφερόμενων ρύπων το 2013 (δεδομένα προς έλεγχο). Ενδιαφέρον δόθηκε στα θετικά σφάλματα πρόβλεψης, δηλαδή στην περίπτωση που η πραγματική τιμή των ρύπων αποκλίνει σημαντικά από την προβλεπόμενη. Αυτό το συμβάν μεταφράζεται ως σημείο εκτός ελέγχου στο διάγραμμα ελέγχου μέσων τιμών για μεμονωμένες τιμές, το οποίο και χρησιμοποιήθηκε για την επιτήρηση της ποιότητας του αέρα στην Αττική. Μετά από την εφαρμογή όλων των παραπάνω μεθόδων αποδείξαμε κυρίως ότι: 1) η εμφάνιση πολλών συνεχόμενων ακραίων τιμών στον σταθμό των Λιοσίων πιθανόν να αποδίδεται σε έντονα καιρικά φαινόμενα που παρατηρήθηκαν εκείνη την περίοδο (1-7 Ιανουαρίου 2013) και 2) στους υπόλοιπους σταθμούς μέτρησης δεν παρατηρήθηκαν πολλά ακραία θετικά σφάλματα πρόβλεψης με αποτέλεσμα να θεωρήσουμε ότι το μοντέλο είναι καλά προσαρμοσμένο στα δεδομένα και δεν διαφέρει πολύ από τις πραγματικές τιμές που παρατηρήθηκαν το 2013.

Table of contents

List of Tables

List of Figures

Chapter 1

Introduction

Air quality is a global issue. In many urban centers around the world, particularly in developing countries, deteriorating air quality is a deepening environmental concern. Poor air quality threatens human health and contributes to environmental damage.
As main sources for the pollution of air, one recognizes: the natural sources, and the man-made sources which are influenced by the industrial development and consequently, the need for increased energy consumption. Primary pollutants are those released directly from the source into the air in a harmful form. Secondary pollutants, by contrast, are modified to a hazardous form after they enter the air or are formed by chemical reactions as components of the air mix and interactive. According to national standards and general international methods the main pollutants to be measured are: NOx (NO, NO_2), CO, SO_2, PM_{10}, O_3, TOC (Total organic carbon) respectively VOC (Volatile organic compounds).

This thesis will initially examine the background philosophy behind air pollution in Europe. In Chapter 2, we briefly discuss air quality management and the eventual setting of National Environmental Standards for Ambient Air Quality. Environmental and statistical management tools to address air quality issues are also introduced. This chapter also includes useful information on background studies conducted, for understanding time series modeling and/or SPC used in air quality monitoring.

In Chapter 3, we introduce the dataset that we will use for analysis. Two pollutants among several assessed throughout the Attica region are chosen to evaluate, the nitrogen dioxide and ozone. They are chosen because they possess special characteristics and are widely known for their effects in human health. Information on the upon characteristics is also described in

detail. Study area also includes the stations used for monitoring the studied pollutants as well as the span of the NO_2 and O_3 concentrations that will be assessed. General information on how the pollutants perform in the Attica region is also given. Finally, we describe the missingness pattern of our data and the multiple imputation techniques as a tool to obtain a full, robust dataset.

In Chapter 4, Autoregressive – Moving Average models (ARIMA) are presented, while further analysis on the model selection steps is given (identification, estimation, forecasting, diagnostic testing). Moreover, the concepts of autocorrelation and stationarity of a time series are introduced and comparison between process monitoring is described. The training datasets introduced in Chapter 3 are examined for autocorrelation and ARIMA models are implemented based on the procedure discussed in theory. Once again the multiple imputation methods are briefly discussed as time series models are implemented for each method separately. Finally, after forecasting, forecast errors are calculated as they are going to be plotted in the next chapter.

In Chapter 5, there is a discussion on traditional statistical process control methods and the difficulties that SPC charts encounter in the presence of autocorrelated data. Basic control charts are quickly reviewed and performance measures such as ARL are introduced. Statistical control charting is used as a guideline to assess expected vs. observed values and serve to augment predictive models when patterns are erratic. As a natural extension of the previous chapter, forecast errors of NO_2 and O_3 ARIMA models are plotted to detect outliers and to understand and improve the process performance as well as forecasting.

The last chapter (Chapter 6) aims to develop the initial interpretations of the previous chapters' results further and link the application of ARIMA models and statistical process control performance as a tool to help detecting outliers as well as improve specification in forecasting. Possible suggestions for future research and limitations are also given at the end of the chapter.

Chapter 2

Motivation of the study

2.1. Introduction

The impact of air pollution on urban climates and air quality monitoring in general have become an important research issue leading to numerous modelling studies related to air quality management and detection of problematic patterns in pollutants' measurements. In this chapter, ambient air quality standards are given as well as general information on air quality in Europe is discussed. The nature of the autocorrelated data such as pollutants, has led us to re – evaluate background studies that modeled such environmental data. An overview of modelling techniques and available software for assessing air quality and environmental data as well as relevant experimental and theoretical studies are briefly discussed. These studies are an incentive and give us guidelines on how to model the data we have in hand and that will be discussed in a following chapter.

2.2. Air Quality Monitoring

Nowadays, automated monitoring networks operate in many European cities providing detailed air quality information on a regular basis. There are several techniques available for monitoring gaseous pollutants (e.g. continuous monitoring using standard gas analyzers, diffusive and pumped sampling using tubes filled with an appropriate adsorbent, grab sampling using canisters) and particulate matter (e.g. filtration and impaction). Each one of them can be associated with a number of advantages and disadvantages that make it suitable or not for a specific application. The response time, which is the time over which the sample is taken, is one of the major factors that determine the suitability of a method. Standard gas

analyzers are sufficiently sensitive and fast to give real time (i.e. typical response time: 1-2 min) measurements of CO, NOX and O_3 concentrations. The results can be then averaged over short time periods and compared to the regulatory standards.

The total number of air quality monitoring stations or sampling locations within a city is limited by practical constraints. Since pollutant concentrations might vary from a street canyon to an urban background area (Palmgren and Kemp, 1999), the selection of monitoring/sampling locations becomes fundamental. The data are often based on one or few monitoring stations placed at critical sites and thus represent microenvironments rather than large urban area. In general, monitoring stations should be located near places of expected air pollution hotspots, but also must be reasonable with respect to population exposure over the averaging times associated with the regulatory values.

2.3. The European Environment until 2010

Presently, airborne particulate matter (PM), ground-level ozone (O_3) and nitrogen dioxide (NO_2) are Europe's most problematic pollutants in terms of harm to health. Effects can range from minor respiratory irritation to cardiovascular diseases and premature death. An estimated 5 million years of lost life per year are due to fine particles ($PM_{2.5}$) alone in Europe.
Air pollution was reduced considerably between 1990 and 2010. This is mainly due to past SO_2 mitigation measures. Nitrogen (N) compounds, emitted as NO_X and ammonia (NH_3), are now the principal acidifying components in our air. In addition to its acidifying effects, N also contributes to nutrient oversupply in terrestrial and aquatic ecosystems, leading to changes in biodiversity. Excessive atmospheric nitrogen in Europe has affected certain sensitive ecosystems and has been decreased slightly between 1990 and 2010. Europe's ambient O_3 concentrations still reduce vegetation growth and crop yields.

Within the European Union (EU), the Sixth Environment Action Program (6EAP) set the long-term objective of achieving levels of air quality that do not give rise to significant negative impacts on, and risks to, human health and the environment. The Thematic Strategy on Air Pollution from the European Commission (EC, 2005) subsequently set interim

objectives for the improvement of human health and the environment through the improvement of air quality to the year 2020.

There has been clear progress made across Europe in reducing anthropogenic emissions of the main air pollutants over recent decades. Nevertheless, poor air quality remains an important public health issue. At present, airborne particulate matter (PM), tropospheric (ground-level) ozone (O_3) and nitrogen dioxide (NO_2) are Europe's most problematic pollutants in terms of causing harm to health. The scale of policy actions undertaken in Europe to specifically address issues concerning air pollution has increased over recent years. Strategies have been developed that require both reduction of emissions at source and reduction of exposures. Local and regional air quality management plans, including initiatives such as low emission zones in cities and congestion charging, must now be developed and implemented in areas of high air pollution. These actions complement measures taken at national level, including, for example, policies setting national emission ceilings, regulating emissions from mobile and stationary sources, introducing fuel quality regulations and establishing ambient air quality standards.

Air quality standards are often unique to a particular country and more complex and established than the respective ones in Greece. Here standards have been established based purely on health and environmental considerations, rather than on political, cultural or social issues many other countries have taken into consideration.

Units of measurement for the ambient air quality standards are expressed in terms of:

- Micrograms ($\mu g = 10^{-6}$ or one one-millionth of a gram) per cubic meter ($\mu g\ m^{-3}$) or milligrams ($mg = 10^{-3}$ or one one-thousandth of a gram) per cubic meter ($mg\ m^{-3}$).
- 1 gram (g) = 1000 milligrams (mg)
- 1 milligram (mg) = 1000 micrograms (μg)
- $1\ m^3 = 1000$ L

2.4. Air Quality Assessment and Management

To reduce the adverse effects of air pollution on health and the environment, various measures are taken at the EU level, including the introduction of fuel quality and product standards. However, in certain areas it is necessary for Member States to take further measures to ensure compliance.

2.4.1. Examples of possible actions by local, regional and national authorities to reduce air pollution in urban areas

Governments can identify their main sources of ambient air pollution, and implement policies known to improve air quality. Improving air quality should be an important consideration in policy planning across different economic sectors (e.g. transport, urban development) to ensure the greatest benefits for health. Examples of such policies are given below.

Transport

- establish low-emission zones that restrict access for more polluting vehicles
- improve transport planning to encourage a shift of transport to less polluting modes including walking, cycling, and public transport
- encourage cleaner fuels and vehicles including use of economic incentives
- renew municipal vehicle fleets to introduce newer, cleaner vehicles
- introduce congestion charging, differentiated parking fees or a city toll
- introduce speed limits and traffic calming measures, for example imposing lower speed limits on main roads
- implement short-term actions such as traffic bans during high pollution episodes
- introduce measures to reduce emissions from non-road vehicles used for example in construction activities

Households, commercial and institutional buildings

- encourage fuel switching from more polluting to cleaner fuels, for example from coal to gas or electricity including use of financial incentives to achieve this
- establish district heating schemes — heat and power cogeneration

- implement rebate schemes that improve the insulation and energy efficiency of buildings
- ensure industrial and commercial combustion sources (including for biomass) are fitted with emission control equipment or replaced

General

- raise the awareness of citizens, provide easy-to-understand information on air quality and health effects of air pollutants
- use air quality forecast and scenario tools to warn the general public and sensitive population groups about episodes of high air pollution

2.5. The characteristics of quality data

Air quality data have certain special characteristics that need to be considered when being analyzed statistically.

1. Data are not usually obtained under sophisticated experimental designs. There are no controlled applications of treatments under carefully administrated randomization scheme. They are more likely obtained through measurements using monitoring networks or specific monitoring stations.
2. Data variability is large and there are many factors affecting it. Observations of air pollutant concentration at a point could be influenced by the emission source characteristics, emission rate, winds, temperatures, precipitation, surface land condition, etc. In addition, different factors may have different importance at different time or place and they may often interact.
3. Skewed data distributions are common. Air quality data often follow heavily skewed probability distributions. The skewness is usually positive, meaning there are many smaller observations and fewer larger ones. This means that the normal distribution assumption required by many classical statistical methods is often violated. The possible existence of outlier observations, in the form of unusually large values also contribute to significant bias in many classical techniques. Hence many resort to

distribution-free methods for analyzing air quality data or model them with heavy-tailed probability distributions like the lognormal, gamma, Weibull, etc.

4. Persistence or auto-correlation often exists in the data. Persistence means the data correlate with themselves at different times. For example, what happens today depends on what happens yesterday. There are a number of possible causes for persistence in air quality data. One possible factor is weather conditions and a possible situation is when a pollution episode lasts over several observational periods. Persistence means that the independent assumption of many classical statistical methods is not satisfied. The effect is that test results may be biased and may lead to erroneous conclusions. One estimate is that with persistence alone, one can detect trends up to 80 percent of the time when no trend actually exists in the data.
5. Monitoring stations provide point measurements. However, for many practical purposes, one needs to convert these point measurements to spatial estimates. This is the realm of spatial statistics. An interesting comparison with hydrometric data is that the point measurement of river discharge, say, is in itself a spatial estimate as it represents the surplus of the water balance within a watershed. There is no comparable parameter for an airshed and the definition of an airshed may be more political and administrative than physical. The precipitation within a watershed also needs to be converted to spatial estimates, however, but the delineation for a watershed is more physical than political/administrative.

2.6. Background studies overview

2.6.1. Time series modeling of environmental data in the literature

Excellent overviews of time series modeling literature, in the environmental field, are given by Prista et al. (2011) and Sunartono et al. (2011). SARIMA multiplicative models were implemented in both cases whereas in the first case the limited size of the data fisheries was adequate to fit and estimate such a model. ACF and PACF play an important role in testing the significance of the model where the use of AIC in selection of the best fitted model

proves to operate well only in in – sample datasets. In Prista et al. significant capabilities of SARIMA models for monitoring landings are explored and it becomes apparent that SARIMA model forecasts include the assumption of persistence of the process that generated the data (approached from a statistical process – control prospective). Therefore large forecast errors can be regarded as indications that there are changes in the fishery process. These forecast errors were used instead of model's forecast to evaluate the performance of the model. The forecasting performance of the multiplicative SARIMA model was compared against other models in both cases. In Sunartono et al. (2011), a subset and an additive model were also used, proving that the latter models give better forecasts at an out – sample dataset than the simple multiplicative model. In Prista et al. (2011), we can see a significant departure of model's forecasts from future observations just by looking at the plotted forecasted errors (residuals) and if they are out – of - control. Detection limits in this case are similar to those of a statistical control process and are calculated as prediction intervals (proved to be "over – optimistic"). Negative forecasted errors indicated underestimation of forecasted landings, thus interest was leaned towards the positive forecasted errors. Generally we observed that SARIMA models provide good short – term forecasts (Zangiacomi et al. 2011, Prista et al. 2011, Sunartono et al. 2011) and it is stated that the most important aspect of time series analysis in general is not the number of parameters, but the degree to which the model approximates the statistical process underlying the data and if it achieves the forecasting objectives.

Yusof et al. (2011) proposed an Extreme Value Distribution (EVD) model for predicting future PM_{10} exceedances. The Cumulative Distribution Function (CDF) (Weibull, gamma and lognormal distributions) and EVD (Frechet and Gumbel distributions) were used to fit the daily maximum concentration of the pollutant. The best distribution that can fit the data was selected based on performance indicators such as MAE, NAE, R2 and IA. Furthermore, the exceedances of a critical PM_{10} concentration were estimated using the best – fitted distributions. This study proved that EVD gives better fit than the CDF for the daily maximum PM_{10} concentration from 2002 to 2006 except for 2005 when there is inconsistency of data recorded due to relocation of sampling site. Therefore, the prediction for future air quality for high PM_{10} concentration was more accurate. It was found out that the exceedances or number of days when PM_{10} concentration is over the standard limits for 2002 is 291 days, 2003 is 224 days, 2004 is 151 days, 2005 is 156 days and 2006 is 9 days.

Mc Nally et al. (2004), proposed alternative models for monitoring bird biodiversity than the usual ARIMA or SARIMA models. General Additive Models (GAMs) may be more difficult to apply but they are more effective and use less compromising assumptions (i.e assumptions of stationarity, invertibility etc) and its modelling framework is well – suited for uncovering trends and unifying datasets (Richards et al. 2010, 2014). The advantages of GAMs in terms of general prediction ability, sensitivity to normal and abnormal data, and clear information for monitoring are indicated by Rongjian et al. (2012) who applied GAM models combined with bootstrap methods for online fault monitoring of glutamate fermentation process with only t, DO, OUR and CER covariates. Environmental processes often generate complex data that are multivariate and potentially nonlinear. Generalized Additive Models (GAMs) are a well-suited modelling framework for uncovering such trends and unifying datasets. This approach allows flexible specification of regression splines to represent the functional relationships between a response variable (the parameter of interest) and a suite of temporal and spatial covariates that can be continuous or discrete using a link function and smooth functions of the covariates.

2.6.2. SPC modeling of environmental data in the literature

Statistical process control monitoring of univariate time series has become the focus of increased attention (Petitgas 2009, Mesnil and Petitgas 2009, Scandol 2003). The use of SARIMA prediction intervals (PI) is similar to that of SPC control charts (Prista et al. 2011) which makes them interesting candidates for the simultaneous monitoring of multiple time series. SARIMA PI's have the advantage of being model-based and do not require extensive historical reference data in order to be calculated. They are also free from the assumption of statistical independence that is a trouble for the estimation of SPC detection limits (Mesnil and Petitgas, 2009).

One commonly used solution is modeling the data through an appropriate time series model and estimating residuals from the model (Alwan and Roberts, 1988). The autoregressive-integrated-moving-average (ARIMA) model is one of the time series models that can be applied for modeling and forecasting the autocorrelated time series data (Alwan and Roberts, 1988). Weisent et al. (2014) also highlighted the advantages of SPC use in time series analysis. He mentioned that traditional models often incorporate smoothing, robust

methods or model outliers to meet assumptions, thereby sculpting a state of statistical control. Improving the ability to distinguish between special causes and common causes (in his case study, for example, outbreaks vs. known seasonality) justifies the effort to identify an "out of control process". Statistical control charting can be used as a guideline to assess expected against observed level of risk and serve to augment predictive models when patterns are erratic.

Xinyao Hu and Xingda Qu (2013) presented a novel pre-impact fall detection model based on the $\bar{X}$ control chart. The fall detection model is individual-specific, since it is constructed using individual historical movement data. The fall detection model demonstrates a high accuracy with up to 94.7% sensitivity and 99.2% specificity. In addition, this model can also provide sufficient time for triggering fall protection device in the pre-impact phase, thus efficient in preventing fall injuries. Xinyao Hu and Xingda Qu (2013) and Lu and Reynolds (1998) also suggested that positive autocorrelation of the process variables can result in severe negative bias in traditional estimators of the standard deviation. This bias produces much tighter control limits than desired and as a result more false alarms are generated. The aforementioned studies concluded that the three sigma control limits provide sufficient power to detect any abnormal changes from an under-controlled process without sounding frequent false alarms.

Christophe Croux et al. (2014) proposed a robust $\bar{X}$ control chart aimed at detecting outliers, in a time series where the robustness of the chart is considered not to harm its performance. Outliers may be present in the “training period”, being the part of the series used to determine the control limits, or in the “test period”, being the part of the series where outliers should be flagged as alarm observations. A large difference between the observed and predicted value implies an unusually large forecasting error, and the corresponding observation is flagged as an outlier in the control chart of the forecast errors. In this paper, the Holt-Winters forecast method is applied, which is a widely used and simple procedure to forecast time series. Every single forecast error is then monitored on an X-chart for individual data and on a robust version of the X-chart. If an isolated outlier occurs in the test sample, the non-robust control chart might yield a sequence of false alarms right after the occurrence of the outlier. The robust approach does not suffer from this drawback. Robust versions of the standard $\bar{X}$ and R chart are given in Rocke (1989) and Tatum (1997), where the mean and scale of the process are estimated robustly for setting up the control limits.

Chapter 3

Athens Pollution Data

3.1. Introduction

The thesis is now focused on the modelling case-study, which is analyzed in detail in Chapters 4 and 5. This chapter emphasizes on the process modelling by presenting the study stations of pollutants' emissions as well as the specific pollutants that are chosen for further evaluation. Existing limits of the pollutants of interest to the modelling case-study (NO_2 and O_3) are examined and analyzed, as their contribution to human health deterioration. Missing measurements on these datasets are handled by using multiple imputation methods, where a review of the aforementioned methods is presented in this chapter. As mentioned in previous chapters, despite PM_{10} being the main factor of pollution of the Athenian air, this thesis case-study concentrates on nitrogen oxide and ozone emissions. Reasons of their assessment are also stated in this chapter.

3.2. Air pollution monitoring stations network in Attica region

3.2.1. Monitoring stations

In 2013, the Department of Air Quality (Ministry of Environment and Energy), operated fourteen air pollution monitoring stations in the Attica region (see. Table 3.1), a station at Oinofyta, and another station in Aliartos (Viotia) for the needs of Cross-Border Pollution Transfer program (EPT).

Table 3.1 shows the location of these stations and their classification.

Table 3.1: Characteristics of air pollution monitoring stations in the Attica region as provided by National Network for Monitoring Air Pollution (NNMAP)

Station				
Location				**Characterization**
Name	**Longitude**	**Latitude**	**Altitude (a.m.s.l.)**	
Athinas	23° 43′ 36′′,63	37° 58′ 41′′,53	100	Urban – roadside[1]
Aristotelous	23° 43′ 39′′,46	37° 59′ 16′′,90	95	Urban - roadside
Geoponiki	23° 42′ 24′′,44	37° 59′ 01′′,05	40	Suburban-Industrial
Liosia	23° 41′ 52′′,23	38° 04′ 36′′,53	165	Suburban-Background
Lykovrisi	23° 47′ 19′′,71	38° 04′ 04′′,35	234	Suburban
Marousi	23° 47′ 14′′,49	38° 01′ 51′′,02	170	Urban - roadside
Nea Smyrni	23° 42′ 46′′,83	37° 55′ 55′′,18	50	Urban-Background[2]
Patisiwn	23° 43′ 58′′,97	37° 59′ 58′′,05	105	Urban - roadside
Piraeus	23° 38′ 42′′,81	37° 56′ 40′′,75	4	Urban - roadside
Peristeri	23° 41′ 18′′,08	38° 01′ 14′′,91	80	Urban-Background
Ag. Paraskeui	23° 49′ 09′′,90	37° 59′ 42′′,39	290	Suburban-Background
Eleusina	23° 32′ 18′′,41	38° 03′ 04′′,86	20	Suburban-Industrial
Thrakomakedones	23° 45′ 29′′,46	38° 08′ 36′′,68	550	Suburban-Background
Korwpi	23° 52′ 44′′,48	37° 54′ 04′′,70	140	Suburban-Background
Oinofyta	23° 38′ 20′′,09	38° 18′ 22′′,39	100	Suburban-Industrial
Aliartos	23° 06′ 36′′,96	38° 22′ 30′′,89	110	Background

[1] Roadside stations are usually located on the pavement of busy streets, avenues or intersections, within few meters distance from the roadway and with their sampling head at 1.5 - 3 m height above ground.
[2]Background stations are placed in parks or other urban locations away from road traffic.

3.2.2. Measured pollutants

Measured pollutants and the methods used are shown in Table 3.2. Measurement of contaminants is done on a continuous basis throughout the course of 24 hours. The response time of automatic analyzers is of the order of one minute, meaning that each analyst gives a

value approximately every minute. The average hourly values of pollutants are calculated with the help of a microprocessor, located at each station and automatically connected to an analyzer.

Table 3.2: Measured pollutants and their methods of measurement

Pollutant	Method of measurement
Carbon Oxide (CO)	Infrared absorption (NDIR)
Nitrogen oxides (NO, NO_2)	Chemiluminescence
Ozone (O_3)	Ultraviolet absorption
Sulfur dioxide (SO_2)	Fluorimetry
Particulate matter (PM_{10}, $PM_{2.5}$)	β – radiation absorption
Benzene (C_6H_6)	Gas Chromatography (GC)

3.3. Temporal fluctuations in concentrations of measured pollutants

3.3.1. Temporal variation of pollutant concentrations in Attica region

The evolution of concentrations of pollutants shows that, although there are fluctuations of average annual pollution concentrations from year to year, there is a trend, a downward trend or a trend of stabilization, depending on the pollutant. This development can be attributed mainly to the technological upgrading of the fleet of cars and public transport, to the implementation of emission control card meter (ECC), the emission control measures pollution from various sources, the use of fuels with better technical specifications, in the operation of fixed rail transport, facilitating the movement of Mass Transmission to penetration of natural gas in the residential, industrial and tertiary sector, the completion of major traffic projects etc.

Specifically for each pollutant, the following are observed:

i) For carbon monoxide, a downward trend in concentrations is generally observed.

ii) For sulfur dioxide, there is a significant downward trend in concentrations associated with reductions in the sulfur content of both diesel fuel and heating and unleaded gasoline.

iii) For benzene, a trend of decreasing concentrations is observed in comparison with previous years.

iv) For nitric oxide, there is a small tendency in declining concentrations.

v) For **nitrogen dioxide**, there is a decline in concentrations in recent years, in most measurement locations.

vi) For **ozone** there is generally a tendency for stabilization of prices with a strong variation from year to year in some stations due to the nature of the pollutant.

vii) For particulate matter (PM_{10}), there is a slight reduction in pollution rates

viii) For particulate matter ($PM_{2.5}$), there is a slight downward trend in concentrations or stabilization.

3.3.2. Effect of meteorological parameters on pollution

Meteorological parameters that affect the formation of air pollution levels are the direction and intensity of the wind, the stability of the atmosphere, and especially for the photochemical pollutants, intensity solar radiation and sunshine duration. Other parameters that modulate the levels of air pollution are the meteorological precipitation and amount of precipitation (rain, snow, etc.), the relative humidity of air and indirectly the temperature.

Lower concentrations of pollutants are observed when winds of the northeast sector are occurred and higher concentrations with the presence of southwestern winds. These values are mainly attributable to the following reasons:

- The closed topography of the basin of Athens, makes more difficult the ventilation and diffusion of pollutants and given the existence of mountains, the prevailing wind is either Northeastern either Southwestern (opening in the northeast between Pendeli and Parnitha mountains and south of the Saronic Gulf). The northeast winds are concise and have a higher average speed over the southwestern winds, contributing in the diffusion of pollutants.

- In case of slight or no accelerated flow, the winds of the south sector result in a local traffic system (sea breeze) which favors the development of high concentrations of secondary (photochemical) pollutants in the basin region.
- Strong winds may affect incrementally the levels of particulate pollution especially in cases of proximity of the station with earthen ground.

3.4. Case Study pollutants - Nitrogen dioxide (NO_2) and Ozone (O_3)

For the case study described in this thesis, two of the pollutants measured in Attica region will be assessed, Nitrogen dioxide (NO_2) and ozone (O_3). These two pollutants, along with PM10 are the most problematic and show in some cases increasing trends. Properties and effects on health and environment of those pollutants are given below in detail.

3.4.1. NO_2 properties

Nitrogen dioxide belongs to a family of highly reactive gases called nitrogen oxides (NOx). These gases form when fuel is burned at high temperatures, and come principally from motor vehicle exhaust and stationary sources such as electric utilities and industrial boilers. A suffocating, brownish gas, nitrogen dioxide is a strong oxidizing agent that reacts in the air to form corrosive nitric acid, as well as toxic organic nitrates. It also plays a major role in the atmospheric reactions that produce ground-level ozone (or smog).

NO_2 is a component of smog and a precursor of ozone. Motor-vehicle exhaust and emissions from other commercial and industrial combustion processes are the major anthropogenic sources of NO_2 (HSDB 2003). Natural sources include forest fires and atmospheric lightning discharges.

3.4.1.1. Health and environmental effects of NO_2

Nitrogen dioxide can irritate the lungs and lower resistance to respiratory infections such as influenza. The effects of short-term exposure are still unclear, but continued or frequent exposure to concentrations that are typically much higher than those normally found in the

ambient air may cause increased incidence of acute respiratory illness in children. EPA's health-based national air quality standard for NO_2 is 0.053 ppm (measured as an annual arithmetic mean concentration). Nitrogen oxides contribute to ozone formation and can have adverse effects on both terrestrial and aquatic ecosystems. Nitrogen oxides in the air can significantly contribute to a number of environmental effects such as acid rain and eutrophication in coastal waters like the Chesapeake Bay. Eutrophication occurs when a body of water suffers an increase in nutrients that leads to a reduction in the amount of oxygen in the water, producing an environment that is destructive to fish and other animal life.

Several studies have been conducted to assess the effects of NO_2 on pulmonary function in asthmatic individuals and patients with chronic lung disease or bronchitis. However, most of the results from studies on pulmonary function and airway hyperactivity in asthmatic humans have been inconclusive and conflicting. Nevertheless, humans with asthma appear to be at greater risk for the respiratory effects of NO_2 exposure than healthy individuals are. For example, it has been reported that asthmatic individuals exposed to NO_2 at 0.3 or 0.5 ppm for $2-4$ h exhibited slight reductions in FEV_1 and specific airway conductance and experienced wheezing and tightness of the chest (Kerr et al.1979; Bauer et al.1985).

3.4.2. O_3 properties

Ozone is an odorless, colorless gas composed of three atoms of oxygen. It occurs both in the Earth's upper atmosphere and at ground level. Ozone can be good or bad, depending on where it is found:

• ***Good ozone (upper level)***. Ozone occurs naturally in the atmosphere 10 to 30 miles above the Earth's surface, where it forms a protective barrier that shields people from the sun's harmful ultraviolet rays. This barrier is sometimes called in the Earth's upper the "ozone layer."

• ***Bad ozone (ground level)***. Because of pollution, ozone can also be found in the Earth's lower atmosphere, at ground level. Ground-level ozone is a major ingredient of smog, and it can harm people's health by damaging their lungs. It can also damage crops and many common man-made materials, such as rubber, plastic, and paint.

Ground-level ozone is not emitted directly into the air but forms when two kinds of pollutants—volatile organic compounds and nitrogen oxides—mix in the air and react chemically in the presence of sunlight. Common sources of volatile organic compounds (often referred to as VOCs) include motor vehicles, gas stations, chemical plants, and other industrial facilities. Solvents such as dry-cleaning fluid and chemicals used to clean industrial equipment are also sources of VOCs. Common sources of nitrogen oxides include motor vehicles, power plants, and other fuel-burning sources.

Photochemical O_3 formation depends mainly on meteorological factors and on the concentrations of NO_X and volatile organic compounds (VOCs). Ozone concentrations in urban areas with high NO_X emissions are generally lower than in the countryside. This is due to the depletion of O_3 through a reaction with nitrogen monoxide (NO), a pollutant especially emitted by traffic — the titration effect. This explains why, in rural areas, where traffic levels and thus concentrations of NO are typically lower, ozone levels are generally higher, though fewer people are exposed.

3.4.2.1. Health effects of O_3

Ozone can irritate the respiratory system. When this happens, you might start coughing, feel an irritation in your throat, and/or experience an uncomfortable sensation in your chest. These symptoms can last for a few hours after ozone exposure and may even become painful.

Ozone can reduce lung function. When scientists refer to "lung function," they mean the volume of air that you draw in when you take a full breath and the speed at which you are able to blow out the air. Ozone can make it more difficult for you to breathe as deeply and vigorously as you normally would.

Ozone can aggravate asthma. When ozone levels are high, more asthmatics have asthma attacks that require a doctor's attention or the use of additional asthma medication.

Ozone can aggravate chronic lung diseases, such as emphysema and bronchitis.

Ozone can inflame and temporarily damage the lining of the lung. Ozone damages the cells that line the air spaces in the lung. Within a few days, the damaged cells are replaced and the old cells are shed. If this kind of damage occurs repeatedly, the lung may change permanently in a way that could cause long-term health effects.

3.5. Data Description

For our case study modeling, we consider NO_2 and O_3 pollutants and we have at our disposal their measurements (concentrations) up to 22 years before, from $1991 - 2013$. As it is mentioned in a section above, NO_2 and O_3 concentrations are measured hourly. These 24 hour – measurements are averaged to one mean daily concentration of NO_2 and O_3 respectively and thus we obtain $n = 365$ (or $n = 366$, depending on the days of the year) observations for each of the years studied. The available datasets have a lot of missing values, complicating any calculations, even the extraction of the mean daily measurements. With a missingness pattern up to 10% for each year, calculation of the mean daily concentrations of NO_2 and O_3 is biased and may alter the true concentrations originally measured.

Out of 16 stations in total in Attica region, we are interested in 8 of those, most of them located close to each other. These stations are: Geoponiki, Athinas, Nea Smyrni, Patisiwn, Liosia, Marousi, Peristeri, Piraeus. We based our choice on different characterization of the stations, as we included urban, suburban, background and industrial locations. Categorization of the aforementioned stations is presented in Table 3.1 above.

It is worth mentioning that the final dataset used for statistical methodology and modeling implementation will not necessarily be of that length ($from 1991 - 2013$) because it is difficult to model such data. Conclusions on the final choice of dataset will be given in detail in following sections.

3.5.1. Handling missing values

The datasets used for further modeling of NO_2 and O_3 concentrations contain a high degree of missingness. Thus, it is essential to impute these missing values in order to obtain a more robust dataset for implementing time series and SPC models. There are several methods of imputation, two of them are described below. Furthermore, general methodology and theoretical background about multiple imputation for time series is given as a base to our subsequent implementation, as well as information on software used.

3.6. Multiple Imputation and time series

Multiple imputation has been shown to reduce bias and increase efficiency compared to listwise deletion. Furthermore, ad-hoc methods of imputation, such as mean imputation, can lead to serious biases in variances and covariances. Unfortunately, creating multiple imputations can be a burdensome process due to the technical nature of algorithms involved. The literature dealing with missing data in time series is nonetheless sparse. Missing data in time series is considered in Little and Rubin (2002); conceptually, the problem can be handled in the same way as in cross sectional data. However, the problem is both harder and more important. Harder, because an additional level of complexity exists when dealing with time series: both contemporaneous and lagged relationships between components need to be considered when imputing a missing data point. With cross sectional data, discarding records with data missing completely at random (MCAR) has no other effect than reducing the available sample. In a time series, each record is unique: dropping it would leave us with a series with holes, unusable for many purposes.

We are going to implement imputation methods to our datasets by using a new package in R, named as Amelia II.

3.6.1. Amelia II and time series imputation methods

Amelia II provides users with a simple way to create and implement an imputation model, generate imputed datasets, and check its fit using diagnostics. The bootstrap-based EMB algorithm included in Amelia II can impute many more variables, with many more observations, in much less time. Amelia II also allows the incorporation of observation and data-matrix-cell level prior information.

The imputation model in Amelia II assumes that the complete data (that is, both observed and unobserved) are multivariate normal. If we denote the (n, k) dataset as D (with observed part D_{obs} and unobserved part D_{mis}), then this assumption is

$$D \sim N_k(\mu, \Sigma)$$

which states that D has a multivariate normal distribution with mean vector μ and covariance matrix Σ. The multivariate normal distribution is often a crude approximation to the true distribution of the data, yet there is evidence that this model works as well as other, more complicated models even in the face of categorical or mixed data.

The essential problem of imputation is that we only observeD_{obs}, not the entirety of D. In order to gain traction, we need to make the usual assumption in multiple imputation that the data are missing at random (MAR). This assumption means that the pattern of missingness only depends on the observed data D_{obs}, not the unobserved data D_{mis}. Let M to be the missingness matrix, with cells $m_{ij} = 1$ if $d_{ij} \epsilon\, D_{mis}$ and $m_{ij} = 0$ otherwise. M is a matrix that indicates whether or not a cell is missing in the data. With this, we can define the MAR assumption as

$$p(M|D) = p(M|D_{obs})$$

Note that MAR includes the case when missing values are created randomly by, say, coin flips, but it also includes many more sophisticated missingness models. When missingness is not dependent on the data at all, we say that the data are missing completely at random (MCAR). Amelia requires both the multivariate normality and the MAR assumption (or the simpler special case of MCAR). Note that the MAR assumption can be made more plausible by including additional variables in the dataset D in the imputation dataset than just those eventually envisioned to be used in the analysis model.

In multiple imputation, we are concerned with the complete-data parameters, $\theta = (\mu, \Sigma)$. When writing down a model of the data, it is clear that our observed data is actually D_{obs} and M, the missingness matrix. Thus, the likelihood of our observed data is $p(D_{obs}, M|\theta)$. Using the MAR assumption, we can break this up

$$p(D_{obs}, M|\theta) = p(M|D_{obs})p(D_{obs}|\theta)$$

We can write the likelihood as

$$L(\theta|D_{obs}) \propto p(D_{obs}|\theta)$$

which we can rewrite using the law of iterated expectations as

$$p(D_{obs}|\theta) = \int p(D|\,\theta) dD_{mis}.$$

With this likelihood and a at prior on, we can see that the posterior is

$$p(\theta|D_{obs}) \propto p(D_{obs}|\theta) = \int p(D|\theta)dD_{mis}$$

The main computational difficulty in the analysis of incomplete data is taking draws from this posterior. The EM algorithm is a simple computational approach to finding the mode of the posterior. Amelia's EMB algorithm combines the classic EM algorithm with a bootstrap approach to take draws from this posterior. For each draw, it bootstraps the data to simulate estimation uncertainty and then run the EM algorithm to find the mode of the posterior for the bootstrapped data. Once we have draws of the posterior of the complete-data parameters, we make imputations by drawing values of D_{mis} from its distribution conditional on $D_{obs}$and the draws of θ.

3.6.2. Imputation analysis in theory

In order to combine the results across *m* data sets, first decide on the quantity of interest to compute, such as a univariate mean, regression coefficient, predicted probability, or first difference. Then, the easiest way is to draw $1/m$ simulations of q from each of the m data sets, combine them into one set of m simulations, and then to use the standard simulation-based methods of interpretation common for single data sets. Alternatively, we can combine directly and use as the multiple imputation estimate of this parameter, $\bar{q}$, the average of the *m* separate estimates, $q_j,\ j = 1, \dots, m$

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j$$

The variance of the point estimate is the average of the estimated variances from within each completed data set, plus the sample variance in the point estimates across the data sets (multiplied by a factor that corrects for the bias because $m < \infty$). Let $SE(q_j)^2$ denote the estimated variance (squared standard error) of q_j from the dataset *j*, and $S_q^2 = \sum_{j=1}^{m}\left(q_j - \bar{q}\right)^2/(m-1)$ be the sample variance across the m point estimates. The standard error of the

multiple imputation point estimate is the square root of $SE(q)^2 = \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + S_q^2(1 + \frac{1}{m})$.

Many variables that are recorded over time within a cross-sectional unit are observed to vary smoothly over time. In such cases, knowing the observed values of observations close in time to any missing value may enormously aid the imputation of that value. However, the exact pattern may vary over time within any cross-section. There may be periods of growth, stability, or decline; in each of which the observed values would be used in a different fashion to impute missing values. Also, these patterns may vary enormously across different cross-sections, or may exist in some and not others. Amelia can build a general model of patterns within variables across time by creating a sequence of polynomials of the time index.

With Amelia we can add covariates to the model that correspond to time and its polynomials. These covariates will help better predict the missing values. If cross-sectional units are specified, these polynomials can be interacted with the cross-section unit to allow the patterns over time to vary between cross-sectional units. When k is set to 0, this interaction simply results in a model of fixed effects where every unit has a uniquely estimated constant term. Amelia does not smooth the observed data and only uses this functional form, or one you choose, with all the other variables in the analysis and the uncertainty of the prediction, to impute the missing values.

When the data to be analyzed contain a high degree of missingness or very strong correlations among the variables, or when the number of observations is only slightly greater than the number of parameters $p(p + 3)/2$ (where p is the number of variables) results from your analysis model will be more dependent on the choice of imputation model. This suggests more testing in these cases of alternative specifications under Amelia. This can happen when using the polynomials of time interacted with the cross section are included in the imputation model. Amelia has a number of methods of setting priors within the imputation model. For example, adding a ridge prior will help with numerical stability by shrinking the covariances among the variables toward zero without changing the means or variances. However, like many Bayesian methods, it reduces variance in return for an increase in bias that one hopes does not overwhelm the advantages in efficiency. In general, we suggest keeping the value on this prior relatively small and increase it only when necessary. As it will be shown below in

detail, a prior of 1% of the number of observations n will be used as a reasonable starting value for imputing our datasets.

3.6.3. Implementation of imputation methods to our dataset

As far as our data is concerned, we have daily mean measurements of NO_2 and O_3 up to 20 years before $(1991 - 2013)$ from 8 different stations in Attica region. More specifically these are: Geoponiki, Athinas, Nea Smyrni, Patisiwn, Liosia, Marousi, Peristeri, Piraeus.

The imputation is implemented both for the whole dataset (all stations together) and for each station's measurements separately. The first imputation method is only implemented for theoretical purposes and further analysis of the NO_2 and O_3 measurements will be based on the latter.

Firstly, we consider 6 different training datasets to impute. We begin with 20 years of measurements for NO_2 and O_3 respectively and we continue with training datasets of $15, 10, 5, 3$ and 1 years. The imputation methods used are two of the most significant methods used for time series imputation. The first one is a –simpler– time series imputation with polynomials of time and the second one is a time series imputation with inclusion of a ridge prior for high missingness pattern, as there are a lot of missing values across some of the stations. As we stated above, a prior of 1% of the number of observations n will be used. With imputing the data for all of the stations, 3 and 5 chains of imputations are created with the first and the second method respectively (this is repeated for each of the 6 training datasets). As it will be shown further in detail in chapter 4, the training datasets that have the smaller MSE and AIC values are those of $5, 3$ and 1 year. Hence, our more realistic choice for studying the behavior of the NO_2 and O_3 concentrations are the 5 or 3 years training datasets and we will continue merely with their further description in the next chapter.

3.6.3.1. Covariates in multiple imputation

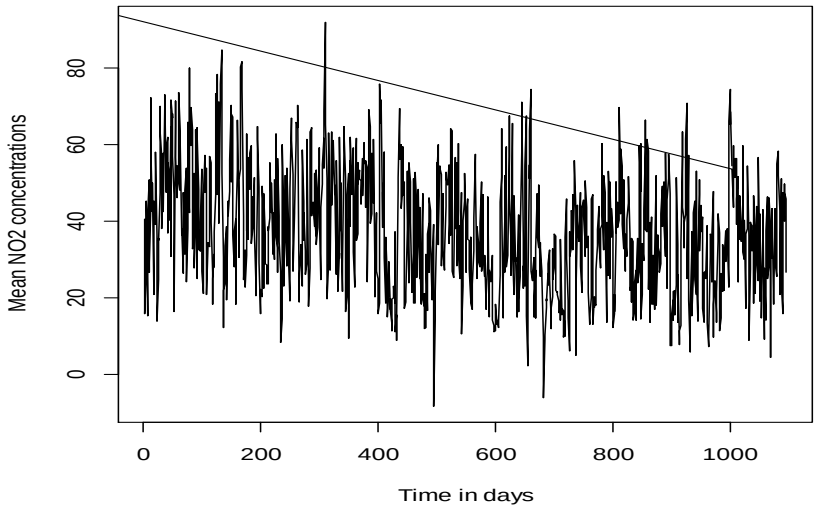
If we add to our data covariates like temperature, humidity and wind velocity, the imputation method is improved, producing more accurate imputed values. Due to this

improvement, we are able to implement more robust and simpler ARIMA models which will give better MSE and AIC values (reference to ARIMA models implemented in this case study and their general use will be described in Chapter 4). In the imputation method implemented by Amelia package, we use temperature as the cross – sectional argument, including all the other variables in the process (humidity and wind velocity as the rest of the covariates in the model).

For the purposes of our analysis, we prefer studying O_3 and NO_2 for each station separately. This analysis is decided to be undergone for 3 consecutive years ($2010 - 2012$). We also implement the same imputation methods and we construct suitable ARIMA models until we find the best ones fitted. The imputation method chosen is based on best fitted ARIMA model, meaning the one with smaller MSE and AIC. It is also chosen based on its simplicity (simpler ARIMA models with small p, d and q are easier to explain), but these will be discussed in Chapter 4. The same imputation procedure was followed separately for 2013 concentrations of NO_2 and O_3. As we will observe in detail in another chapter, NO_2 and O_3 measurements of 2013 will be used as Phase II data in SPC modeling in order to detect values of the pollutants that deviate from the in-control measurements.

Figure 3.1 shows an example of the resulted imputed dataset of NO_2 measurements of Geoponiki station. As it will be analyzed furtherly in Chapter 4, there seems to be a trend in our data, but no specific seasonality is captured.

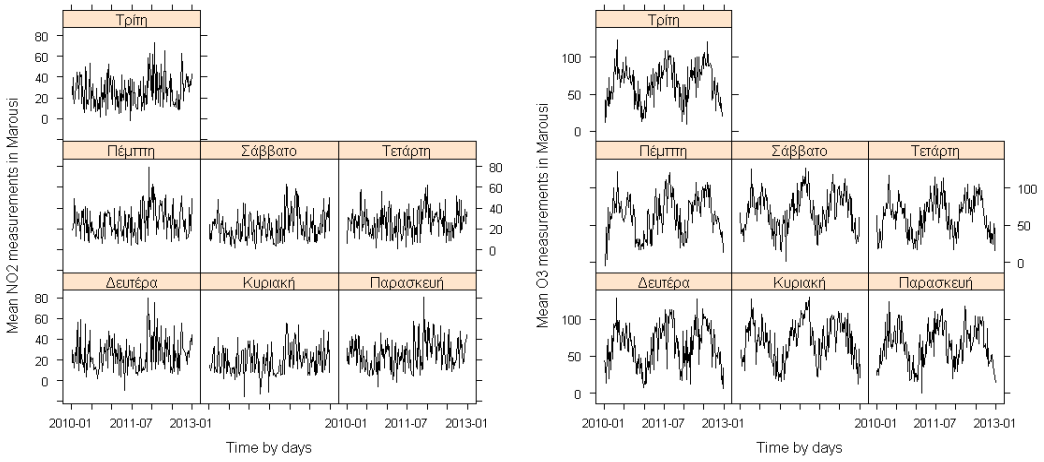
Fig 3.1: Time series plot of the NO_2 concentrations in Geoponiki station after imputation



A special characteristic of air quality data (as mentioned in Chapter 2) is that the assumption of normality is not satisfied. Thus, fitting other models such as GLM, GAM or Gamma Distribution models might be a more realistic choice. However, we continue with ARIMA models because they are simpler and widely used in bibliography.

In figure 3.2, time series of NO_2 and O_3 concentrations in Marousi station from $2010 - 2012$ (training dataset) for each day of the week are illustrated. According to the plots, higher concentrations of NO_2 are observed during Friday and Monday, while high concentrations of O_3 are observed during the weekend. Note that there is a seasonality pattern for O_3 time series data. This is verified by the nature of the pollutant, as ozone concentrations are usually higher during the spring and summer, when the temperature and the humidity are high.

Fig 3.2: NO_2 and O_3 time series of Marousi station from $2010 - 2012$ (training dataset) for each day of the week respectively

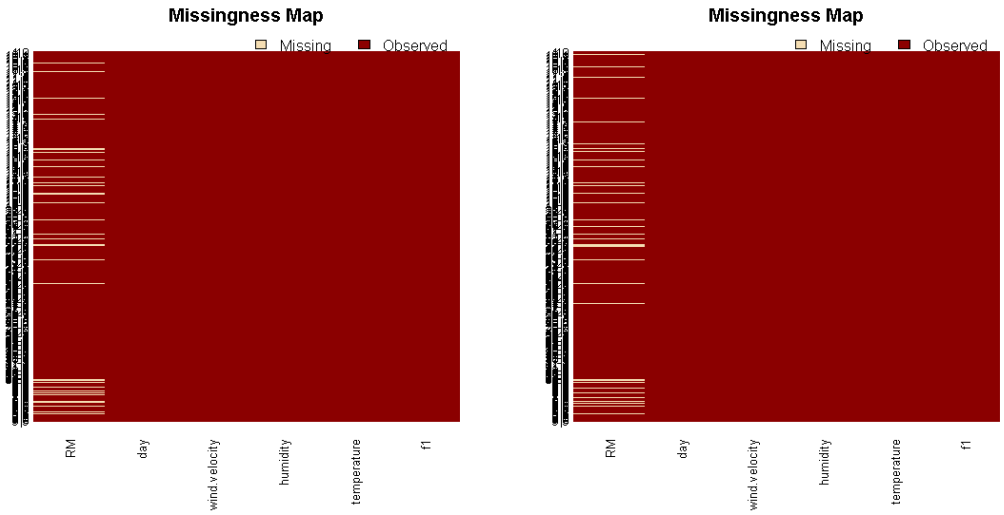


3.6.3.2. Missingness maps

One useful tool for exploring the missingness in a dataset is a missingness map. This is a map that visualizes the dataset a grid and colors the grid by missingness status. The column of the grid is the variables and the rows are the observations, as in any spreadsheet program. This tool allows for a quick summary of the patterns of missingness in the data.

Figure 3.3 shows missingness maps for station Geoponiki datasets for NO_2 and O_3 daily concentrations respectively. It is clear that the variable RM (mean NO_2 and O_3 measurements) is the only variable missing in both datasets.

Figure 3.3: Missingness maps for station Geoponiki (NO_2 and O_3 daily concentrations)



The missingness map is important for understanding the patterns of missingness in the data and can often indicate potential ways to improve the imputation model or data collection process.

3.7. Sources of Data and software used

NO_2 and O_3 hourly concentrations are provided from the department of Air Quality of the Ministry of Environment and Energy in several excel files, one for each year. Temperature, humidity and wind velocity data originate from Athens observatory and they are daily measurements used as covariates in modeling the previous NO_2 and O_3 concentrations. The latter dataset has no missing values, so no imputation methods were used. Wind direction was another variable provided, but it was no included in the imputation formula and the model, as it is more difficult than the other covariates to interpret. Statistical programming language R 3.1.2 was used for multiple imputation methods and the implementation of ARIMA models as well as of the control charts. More specifically, the package "Amelia II" was solely used for multiple imputation and missingness maps, whereas "forecast" package was used for generating time series plots, ACF plots, ARIMA estimation and forecasting. Other packages used for evaluating our case study were "stats", "tseries", "qcc" "spc" and "lattice".

Chapter 4

Time Series Modeling

4.1. Introduction

We turn now to the construction of time-series models. Our objective is to develop models that "explain" the movement of a time series by relating it to its own past values and to a weighted sum of current and lagged random disturbances. While many functional forms can be used, we will focus on linear models. In the first section of this chapter we introduce mixed autoregressive – moving average models. In these models the process is a function of both its past values and lagged random disturbances, as well as a current disturbance term. Even if the original process is nonstationary, it often can be differenced one or more times to produce a new series that is stationary and for which a mixed autoregressive – moving average model can be constructed. This model can be used to produce a forecast one or more periods into the future, after which the forecasted stationary series can be integrated one or more times to yield a forecast for the original time series.

After presenting the theoretical background of time series modeling, we apply ARIMA models to our datasets of daily NO_2 and O_3 concentrations, we estimate the parameters of the best fitted models after choosing them based on several accuracy methods (MSE, AIC) and a diagnostic check of the residuals is performed in order to see if there is any persistent autocorrelation left on the data. Finally two methods of forecasting are introduced to our data (one – step ahead forecasts without re – estimation of the model and one – step ahead forecasts with re – estimation of the model).

4.2. Mixed autoregressive – moving average models

Many stationary random processes cannot be modeled as purely moving average or as purely autoregressive, since they have the qualities of both types of processes. The logical extension of the moving average model and the autoregressive model is the mixed autoregressive – moving average process of order (p, q). We denote this process as ARMA (p, q) and represent it by

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \qquad (4.1)$$

We assume that the process is stationary, so that its mean is constant over time and is given by

$$\mu = \varphi_1 \mu + \cdots + \varphi_p \mu + \delta$$

or

$$\mu = \frac{\delta}{1 - \varphi_1 - \cdots - \varphi_p}$$

This gives a necessary condition for the stationarity of the process, that is,

$$\varphi_1 + \varphi_2 \ldots + \varphi_p < 1$$

The simplest mixed autoregressive – moving average process is ARMA (1,1):

$$y_t = \varphi_1 y_{t-1} + \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

The autocorrelation function of this process is given by

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{(1 - \varphi_1 \theta_1)(\varphi_1 - \theta_1)}{1 + \theta_1^2 - 2\varphi_1 \theta_1}$$

where γ_1 and γ_2 are covariance and autocovariance respectively.

For displacement k greater than 1,

$$\rho_k = \varphi_1 \rho_{k-1} \ , \quad k \geq 2$$

Thus, the autocorrelation function begins at its starting value ρ_1 (which is a function of both φ_1 and θ_1 and then decays geometrically from that starting value. This reflects the fact that the moving average part of the process has a memory of only one period.

Often it is convenient to write or describe time lags by using the backward shift operator B . The operator B imposes a one – period time lag each time it is applied to a variable. Thus, $\varepsilon_t = \varepsilon_{t-1}$, $B^2\varepsilon_t = \varepsilon_{t-2}$, … , $B^n = \varepsilon_{t-n}$.

Using this operator we can now rewrite eq. (4.1) as

$$\left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right) y_t = \delta + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t \qquad (4.2)$$

4.3. Homogeneous nonstationary processes: ARIMA models

In practice, many of the time series we work with are nonstationary, so that the characteristics of the underlying stochastic process change over time. In this section we construct models for those nonstationary series which can be transformed into stationary series by differencing them one or more times. We say that y_t is a homogeneous nonstationary of order d if

$$w_t = \Delta^d y_t$$

is a stationary series. Here Δ denotes differencing, i.e., $\Delta y_t = y_t - y_{t-1}$, $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$ and so forth.

After we have differenced the series y_t to produce the stationary series w_t, we can model w_t as an ARMA process. If $w_t = \Delta^d y_t$ and w_t is an ARMA (p, q) process, then we say that y_t is an integrated autoregressive – moving average process of order (p, d, q) or simply ARIMA (p, d, q). We can write the equation for the process ARIMA (p, d, q), using the backward shift operator, as

$$\varphi(B)\Delta^d y_t = \delta + \theta(B)\varepsilon_t$$

4.3.1. Selection of ARIMA models (Box – Jenkins approach)

ARIMA models are usually fitted by using a sequence of three general steps collectively known as Box – Jenkins (BJ) method: 1) identification of the model, 2) estimation of the model and 3) a diagnostic check of the model.

4.3.1.1. Autocorrelation function

In the identification stage, a model structure (p, d, q) is selected by comparisons of sample ACF and PACF with theoretical ACF/PACF profiles of AR, MA and ARMA processes. The autocorrelation function tells us how much correlation there is (and by implication how much independency there is) between neighboring data points in the series y_t. Suppose the stochastic process is simply $y_t = \varepsilon_t$ where ε_t is an independently distributed random variable with zero mean. The autocorrelation function for this process is given by $\rho_0 = 1, \rho_k = 0$, $k > 0$. The process $y_t = \varepsilon_t$ is called white noise, and there is no model that can provide a forecast any better than $y_{t+l} = 0$ for all l. Thus if the autocorrelation function is zero or close to zero for $k > 0$, there is a little or no value in using a model to forecast the series. The ACF function is purely theoretical so we must calculate an estimate of the ACF, called sample autocorrelation function.

To test whether a particular value of the autocorrelation function ρ_k is equal to zero, we use a result obtained by Bartlett (1937). He showed that if a time series has been generated by a white noise process, the sample autocorrelation coefficients are distributed approximately according to a normal distribution with mean 0 and standard deviation $1/\sqrt{T}$ (T = number of observations in the time series). To test the joint hypothesis that all the autocorrelation coefficients are zero, we use the Ljung – Box – Pierce test or also known as Portmanteau test.

The autocorrelation function can also be used to test whether a series is stationary. If $\hat{\rho}_k$ does not fall off quickly as k increases, this is an indication of nonstationarity. On the other hand, if it declines rapidly, this is an indication of stationarity.

4.3.1.2. Seasonality and the autocorrelation function

Often seasonal peaks are easy to spot by direct observation of the time series. However, if the time series fluctuate considerably, seasonal peaks may not be distinguishable from the other fluctuations. Recognition of seasonality is important because it provides information about regularity in the series that can help us in forecasting. We can identify seasonality by observing regular peaks in the autocorrelation function, even if they cannot be distinguished in the time series itself.

4.3.2. Estimating and Forecasting with time series models

In this section we show how the parameters of an ARIMA model are estimated. If the model contains moving average terms, this involves the application of a nonlinear estimation method. After this we describe diagnostic checking, a procedure used to test whether the model has been specified correctly (i.e., whether p, d and q have been chosen correctly). Once a time series model has been estimated and checked, it can be used for forecasting. In this section we explain how to use the general ARIMA model

$$\varphi(B)\Delta^d y_t = \theta(B)\varepsilon_t$$

to obtain a forecast of y_t for period $t + l$ (that is l periods ahead with $l \geq 1$). We denote this forecast by $\widehat{y_t}(l)$ and call it the origin – t forecast for lead time l.

4.3.2.1. Model Estimation

Suppose a tentative specification of the time-series model has been made, i.e., values of p, d and q have been chosen for the ARIMA model,

$$\varphi(B)\Delta^d y_t = \varphi(B)w_t = \theta(B)\varepsilon_t$$

With $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$. Now estimates must be obtained for the p autoregressive parameters $\varphi_1, \ldots, \varphi_p$ and the q moving average parameters $\theta_1, \ldots, \theta_q$. As in the case of the regression model, we choose parameter values that will minimize the sum of squared differences between the actual time series $w_t = \Delta^d y_t$ and the fitted times series $\widehat{w_t}$.

We can rewrite the equation above in terms of the error term series ε_t^2:

$$\varepsilon_t = \theta^{-1}(B)\varphi(B)w_t$$

The objective in estimation is to find a set of autoregressive parameters ($\varphi_1, \ldots, \varphi_p$) and a set of moving average parameters ($\theta_1, \ldots, \theta_q$) that minimize the sum of squared errors

$$S(\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q) = \sum_t \varepsilon_t^2$$

We denote the sets of parameters that minimize the previous equation by $\hat{\varphi}_1, \dots, \hat{\varphi}_p$ and $\hat{\theta}_1, \dots, \hat{\theta}_q$ and denote the residuals associated with these parameter values by $\hat{\varepsilon}_t$, so that $\hat{\varepsilon}_t = \hat{\theta}^{-1}(B)\hat{\varphi}(B)w_t$. Thus,

$$S(\hat{\varphi}_1, \dots, \hat{\varphi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q) = \sum_t \hat{\varepsilon}_t^2 \qquad (4.3)$$

If moving average terms are present, $\varepsilon_t = \theta^{-1}(B)\varphi(B)w_t$ is nonlinear in the parameters and so a method of nonlinear estimation must be used in the minimization of Eq. (4.3). In addition, the first error term in the series, ε_1, depends on the past and unobservable values $w_0, w_{-1}, \dots, w_{-p+1}$ and $\varepsilon_0, \varepsilon_{-1}, \dots, \varepsilon_{-q+1}$.

After the model has been estimated, a procedure of diagnostic checking is used to test whether the initial specification was correct. We would expect the residuals $\hat{\varepsilon}_t,\ t = 1, \dots, T$, to resemble closely the true errors ε_t, which by assumption are uncorrelated. We then test whether these residuals are indeed uncorrelated. If they are not, we will want to respecify the model (i.e., choose new values for p, d and q), estimate this new model and perform another diagnostic check. Once the model has been checked to our satisfaction, it can be used for forecasting.

4.3.2.2. Diagnostic Testing

The process of diagnostic checking usually involves two steps. First the autocorrelation function for the simulated series (i.e., the time series generated by the model) can be compared with the sample autocorrelation function of the original series. If the two autocorrelation functions seem very different, we may doubt the validity of the model and a re-specification may be in order. If the two autocorrelation functions are not significantly different (and this will most often be the case), one can analyze the residuals of the model.

We have assumed that the random error terms ε_t in the actual process are normally distributed and independent. Then, if the model has been specified correctly, the residuals $\hat{\varepsilon}_t$ should resemble a white noise process. In particular we would expect the residuals to be nearly uncorrelated with each other so that a sample autocorrelation function of the residuals would be close to 0 for displacement $k \geq 1$.

The residuals of the model are:

$$\varepsilon_t = \theta^{-1}(B)\varphi(B)w_t$$

The sample autocorrelation function $\hat{r}_k$ of the residuals is calculated by

$$\hat{r}_k = \frac{\sum_t \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_t \hat{\varepsilon}_t^2}$$

A very convenient test based on statistical results is obtained by Box and Pierce and can be applied to this sample autocorrelation function. If the model is correctly specified, then for large k, the residual autocorrelations $\hat{r}_k$ are themselves uncorrelated, normally distributed random variables with mean 0 and variance$1/T$, where T is the number of observations in the time series.

The statistic Q composed of the first K residual autocorrelations $\hat{r}_1, \ldots, \hat{r}_k$ is

$$Q = T\sum_{k=1}^{K} \hat{r}_k^2$$

This statistic is a sum of squared independent normal random variables, each with mean 0 and variance $1/T$, and is therefore itself approximately distributed as chi – square. The first few autocorrelations will have a variance slightly less than $1/T$ and may themselves be correlated. Box and Pierce demonstrate that the approximation is quite close and that the statistic Q will be distributed as $X^2(K-p-q)$. Therefore, a statistical hypothesis test of the model's accuracy can be performed by comparing the observed value of Q with the appropriate points from a chi – square table.

4.3.2.3. Computing a forecast

The computation of the forecast $\widehat{y_t}(l)$ can be done recursively by using the estimated ARIMA models. This involves first computing a forecast one period ahead, then using this forecast to compute a forecast two periods ahead, and continuing until the *l*-period forecast has been reached. Let as write the ARIMA (p, d, q) model as

$$w_t = \varphi_1 w_{t-1} + \cdots + \varphi_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} + \delta$$

with $y_t = \Sigma^d w_t$

To compute the forecast $\widehat{y_t}(l)$, we begin by computing the one – period forecast of w_t, $\widehat{w}_t(1)$. To do so, we write the equation above with the time period modified:

$$w_{t+1} = \varphi_1 w_t + \cdots + \varphi_p w_{t-p+1} + \varepsilon_{t+1} - \theta_1 \varepsilon_t - \cdots - \theta_q \varepsilon_{t-q+1} + \delta$$

We then calculate our forecast $\widehat{w}_t(1)$ by taking the conditional expected value of w_{t+1}:

$$\widehat{w}_t(1) = E(w_{t+1}|w_t, \dots) = \varphi_1 w_t + \cdots + \varphi_p w_{t-p+1} - \theta_1 \hat{\varepsilon}_t - \cdots - \theta_q \hat{\varepsilon}_{t-q+1} + \delta$$

where the $\hat{\varepsilon}_t$, $\hat{\varepsilon}_{t+1}$, etc. are observed residuals. The expected value of ε_{t+1} is 0. Now, using the one period forecast $\widehat{w}_t(1)$, we can obtain the two period forecast, $\widehat{w}_t(2)$:

$$\widehat{\mathrm{w}}_\mathrm{t}(2) = \mathrm{E}(\mathrm{w}_{\mathrm{t+2}}|\mathrm{w}_\mathrm{t}, \dots) = \varphi_1 \widehat{\mathrm{w}}_\mathrm{t}(1) + \varphi_2 \mathrm{w}_\mathrm{t} + \cdots + \varphi_\mathrm{p} \mathrm{w}_{\mathrm{t-p+2}} - \theta_2 \hat{\varepsilon}_\mathrm{t} - \cdots - \theta_\mathrm{q} \hat{\varepsilon}_{\mathrm{t-q+2}} + \delta$$

The two – period forecast is then used to produce the three period forecast and so on until the l – period forecast $\widehat{w}_t(l)$ is reached:

$$\widehat{w}_t(l) = \varphi_1 \widehat{w}_t(l-1) + \cdots + \varphi_1 w_t + \cdots + \varphi_p w_{t-p+1} - \theta_1 \hat{\varepsilon}_t - \cdots - \theta_q \hat{\varepsilon}_{t-q+1} + \delta$$

Once the differenced series w_t can be forecasted, a forecast can be obtained for the original series y_t simply by applying the summation operation tow_t, that is, by summing w_t d times. Suppose that $d = 1$. Then our l – period forecast of y_t is given by

$$\hat{y}_t(l) = y_t + \widehat{w}_t(1) + \widehat{w}_t(2) + \cdots + \widehat{w}_t(l)$$

However, if the model for y_t were ARIMA with $d = 2$, the l – period forecast $\hat{y}_t(l)$ would be given by

$$\begin{aligned}\hat{y}_t(l) &= y_t + [\Delta y_t + \widehat{w}_t(1)] + [\Delta y_t + \widehat{w}_t(1) + \widehat{w}_t(2)] + \cdots + [\Delta y_t + \widehat{w}_t(1) + \cdots + \widehat{w}_t(l)] \\ &= y_t + l\Delta y_t + l\widehat{w}_t(1) + (l-1)\widehat{w}_t(2) + \cdots + \widehat{w}_t(l)\end{aligned}$$

The procedure is similar for larger values of d.

4.3.2.3.1. Minimum mean square error forecasts

Besides the expression of $\widehat{w}_t(l)$ in terms of the difference equation, it can also be expressed as an infinite weighted sum of current and previous shocks ε_j:

$$\widehat{w}_t(l) = \sum_{j=-\infty}^{t+1} \psi_{t+l-j}\varepsilon_j = \sum_{j=0}^{\infty} \psi_j\, \varepsilon_{t+l-j} \quad (4.4)$$

Where $\psi_0 = 1$ and the ψ weights may be obtained by equating coefficients in $\varphi(B)(1 - \psi_1 B + \psi_2 B^2 + \cdots) = \theta(B)$

Suppose we would like to make a forecast $\widehat{w}_t(l)$ of w_{t+l} which is to be a linear function of current and previous observations w_t, $w_{t-1}, w_{t-2}, \ldots$. Then it will also be a linear function of current and previous shocks $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\ldots$

The best forecast is:

$$\widehat{w}_t(l) = y_l^* \varepsilon_t + y_{l+1}^* \varepsilon_{t-1} + y_{l+2}^* \varepsilon_{t-2} + \cdots$$

where the weights $y_l^*, y_{l+1}^*, \ldots$ are to be determined. Using Eq. (4.4), the mean square error of the forecast is

$$E[w_{t+l} - \widehat{w}_t(l)]^2 = (1 - \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_\varepsilon^2 + \sum_{j=0}^{\infty} \{\psi_{l+j} - \psi_{l+j}^*\}\sigma_\varepsilon^2$$

which is minimized by setting $\psi_{l+j}^* = \psi_{l+j}$

We have then:

$$w_{t+l} = (\varepsilon_{t+l} + \psi_1\varepsilon_{t+l-1} + \cdots + \psi_{l-1}\varepsilon_{t+1}) + (\psi_1\varepsilon_t + \psi_{l+1}\varepsilon_{t-1} + \cdots) = \hat{\varepsilon}_t(l) + \widehat{w}_t(l)$$

Where $\hat{\varepsilon}_t(l)$ is the error of the forecast $\widehat{w}_t(l)$ at lead time *l*.

So we have $\hat{\varepsilon}_t(l) = \varepsilon_{t+l} + \psi_1\varepsilon_{t+l-1} + \cdots + \psi_{l-1}\varepsilon_{t+1}$ and since $E[\hat{\varepsilon}_t(l)] = 0$, the forecast is unbiased.

Also the variance of the forecast error is

$$V(l) = var[\hat{\varepsilon}_t(l)] = (1 - \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_\varepsilon^2$$

4.3.2.3.2. The residuals as one step ahead forecast errors

By using the equation $\hat{\varepsilon}_t(l) = \varepsilon_{t+l} + \psi_1\varepsilon_{t+l-1} + \cdots + \psi_{l-1}\varepsilon_{t+1}$, the one step ahead forecast error is

$$\hat{\varepsilon}_t(1) = w_{t+1} - \widehat{w}_t(1) = \varepsilon_{t+1}$$

Hence, the residuals ε_t which generate the process, and which are introduced as a set of independent random variables or shocks, turn out to be the one step ahead forecast errors.

It follows that, for a minimum square error forecast, the one step ahead forecast errors must be uncorrelated. This is eminently sensible, for if one step ahead errors where correlated then the forecast error ε_{t+1} could be predicted from available forecast errors $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots..$ If the prediction so obtained was $\hat{\varepsilon}_{t+1}$, then $\widehat{w}_t(1) + \hat{\varepsilon}_{t+1}$ would be a better forecast of w_{t+1} than was $\widehat{w}_t(1)$.

4.3.2.3.3. Correlation between the forecast errors

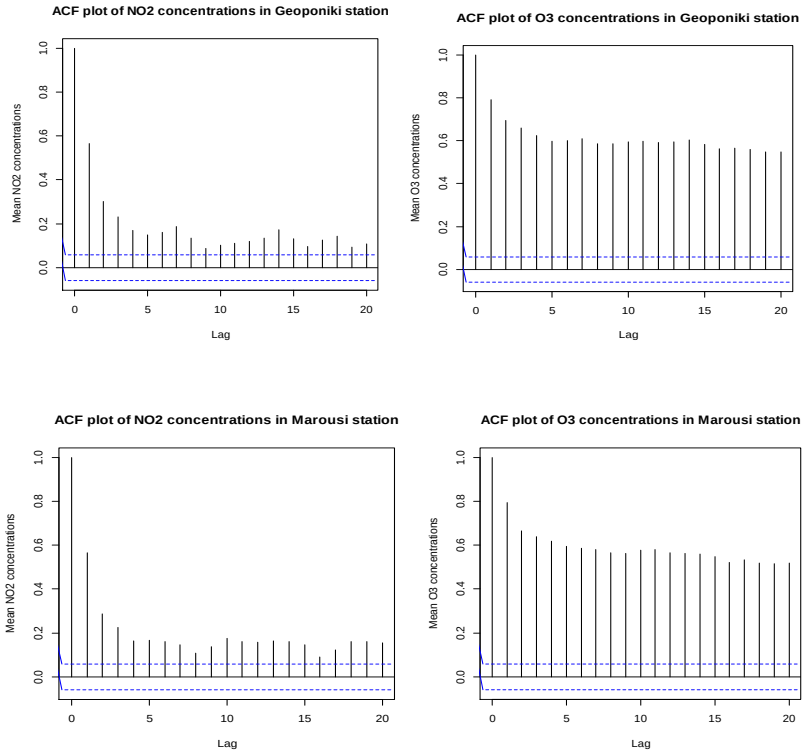
Although the optimal forecast errors at lead time 1 will be uncorrelated, the forecast errors for longer lead times in general will be correlated. There will be often a tendency for the forecast function to lie either wholly above or below the values of the series when they eventually come to hand.

4.4. Implementation of ARIMA modeling in our case study

As it is mentioned in Chapter 3, two methods of multiple imputation for time series were chosen and used to impute the existing missing values in our dataset. The first is a basic imputation method with polynomials of time and the second one is a more complex method including a ridge Bayesian prior for high missingness in the data. After imputation with both methods, ARIMA models are used to describe and model our autocorrelated data.

Before starting with implementation of ARIMA models, one should state that it would not be needed to implement such models if it wasn't for the autocorrelation in the initial time series. This becomes more obvious if we take a look at ACF of the time series for two of the stations studied.

Fig 4.1: ACF plot of NO_2 and O_3 concentrations of Geoponiki and Marousi station respectively



The autocorrelation is asserted if the ACF plot demonstrates a ‘‘tail off’’ pattern (i.e. the value of ACF decreases gradually). This gradual decrease is more apparent in the ACF plots of O_3 concentrations, so need for fitting an ARIMA model is again verified. Same patterns as in the previous ACF plots are met in all stations of both pollutants. In our case we started with multiple training datasets in order to choose the exact dataset used for further modeling (see chapter 3). An ARIMA model is implemented for each training dataset. The best fitted ARIMA model for each dataset is the one that has the smaller Mean Square Error (MSE) and Akaike Information Criterion (AIC). Among the above, other accuracy methods are calculated such as MAE, MAPE and the first autocorrelation function (ACF_1). The MSE is an absolute error measure that squares the errors (the difference between the actual historical data and the forecast – fitted data predicted by the model) to keep the positive and negative errors from canceling each other out. The measure also tends to exaggerate large errors by weighting the large errors more heavily than smaller errors by squaring them, which can help when comparing different time series models. The MSE is calculated by simply taking the average of σ_ε^2. MAPE is usually preferred as an accuracy method because it is easier to explain due to its nature (percentages are more illustrative than a single number). Nevertheless, MSE is widely used and applied in time series analysis and we will focus on its value during our

ARIMA model selection. However, a more accurate method is Mean Standardized Square Error (MSSE), as it is – in contrast with MSE – a well-known goodness-of-fit measure. As it will be mentioned at the end of the Chapter, we will eventually study the forecast errors (not the standardized errors), therefore the use of MSSE is optional and not chosen to describe.

Tables with the aforementioned training datasets and the best fitted ARIMA models are given below for each of the pollutants studied (NO_2 and O_3). Some of the accuracy methods mentioned above are given for each model.

Table 4.1: ARIMA modeling for training datasets 20, 15 and 10 years after imputation for O_3

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
	ARIMA(1,1,2) model with drift	***ARIMA(2,1,2) model***
Training dataset: 20 years	MSE=349.306	MSE=351.998
	MAE=13.47947	MAE=13.49657
	MAPE=79.43283	MAPE=86.21512
	ACF1= -0.0003919	ACF1= -0.0004205
	AIC=503443.6	AIC=503890.1
	ARIMA(3,1,1) model	***ARIMA(2,1,2) model***
Training dataset: 15 years	MSE=376.6468	MSE=369.7991
	MAE=14.05007	MAE=13.97266
	MAPE=91.83109	MAPE=112.22
	ACF1=-2.282 e-05	ACF1=-0.0003965
	AIC=382918.8	AIC=382117.6
	ARIMA(1,1,1) model	***ARIMA(2,1,1) model***
Training dataset: 10 years	MSE=271.5288	MSE=324.1066
	MAE=11.97117	MAE=13.06121
	MAPE=51.19714	MAPE=79.08824
	ACF1=-0.001326	ACF1=0.0002293
	AIC=123328	AIC=251769.7

Table 4.2: ARIMA modeling for training datasets 5, 3 and 1 years after imputation for O_3

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
	ARIMA(1,1,2) model	***ARIMA(1,1,1) model***
Training dataset: 5 years	MSE=269.8493	MSE=280.7562
	MAE=11.96484	MAE=12.0894
	MAPE=57.36891	MAPE=73.18189
	ACF1=2.858 e-05	ACF1=-0.002579
	AIC=123239.3	AIC=123921.2
	ARIMA(2,1,1) model	***ARIMA(3,1,1) model***
Training dataset: 3 years	MSE=261.3764	MSE=249.694
	MAE=11.86008	MAE=11.6649
	MAPE=60.16327	MAPE=47.80428
	ACF1=-2.9285 e-06	ACF1=0.0002919
	AIC=73627.72	AIC=73229.17
	ARIMA(1,1,1) model	***ARIMA(1,1,1) model***
Training dataset: 1 year	MSE=221.8664	MSE=248.7733
	MAE=11.30404	MAE=11.83802
	MAPE=45.64328	MAPE=43.30092
	ACF1=0.0027055	ACF1=0.0058968
	AIC=24126.38	AIC=24461.54

Table 4.3: ARIMA modeling for training datasets 20, 15 and 10 years after imputation for NO_2

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
	ARIMA(2,1,2) model with drift	***ARIMA(4,1,1) model***
Training dataset: 20 years	MSE=483.1178	MSE=481.5203
	MAE=16.49996	MAE=16.46872
	MAPE=92.03711	MAPE=88.09654
	ACF1= -0.0002015	ACF1= -0.00011128
	AIC=522219.4	AIC=522036.5
	ARIMA(4,1,3) model	***ARIMA(5,1,3) model***
Training dataset: 15 years	MSE=438.0565	MSE=434.1856
	MAE=15.80929	MAE=15.76995
	MAPE=73.44416	MAPE=69.85942
	ACF1=9.7773 e-05	ACF1=0.00010804
	AIC=389521.2	AIC=389135.5
	ARIMA(1,1,1) model	***ARIMA(5,1,4) model***
Training dataset: 10 years	MSE=367.4793	MSE=366.7685
	MAE=14.45436	MAE=14.40566
	MAPE=313.0083	MAPE=108.2935
	ACF1=-3.082 e-06	ACF1=0.0001221
	AIC=255437	AIC=255425.8

Table 4.4: ARIMA modeling for training datasets 5, 3 and 1 years after imputation for NO_2

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
	ARIMA(2,1,1) model	***ARIMA(3,1,1) model***
Training dataset: 5 years	MSE=287.3266	MSE=282.5294
	MAE=12.63176	MAE=12.5533
	MAPE=69.31308	MAPE=63.71348
	ACF1=7.619 e-05	ACF1=1.383 e-05
	AIC=124165.2	AIC=123921.2
	ARIMA(3,1,1) model	***ARIMA(2,1,2) model***
Training dataset: 3 years	MSE=227.7676	MSE=229.7386
	MAE=11.38323	MAE=11.39705
	MAPE=72.00521	MAPE=59.57715
	ACF1=-7.7842 e-05	ACF1=0.0010569
	AIC=72432.96	AIC=72508.44
	ARIMA(4,1,3) model	***ARIMA(5,1,5) model***
Training dataset: 1 year	MSE=195.0416	MSE=194.4181
	MAE=10.61737	MAE=10.56236
	MAPE=67.89437	MAPE=52.5116
	ACF1=0.000141788	ACF1=-0.002809
	AIC=23759.95	AIC=23756.55

As it is shown above, the training datasets that have the smaller MSE and AIC values are those of $5, 3$ and 1 year. Hence, our more realistic choice for studying the behavior of the NO_2 and O_3 measurements are the 5 or 3 years training datasets and we continue merely with their further description. Adding to our data covariates like temperature, humidity and wind velocity, the imputation method is improved, producing more accurate imputed values. Due to this improvement, we implement more robust and simpler ARIMA models which give better MSE and AIC values.

The tables below show the improved generated ARIMA models for each of the imputed training dataset:

Table 4.5: ARIMA models and accuracy methods for training datasets after imputation for O_3, including temperature, humidity and wind velocity as covariates in the imputation

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
	ARIMA(2,1,1) model	***ARIMA(1,1,1) model***
Training dataset: 3 years	MSE=184.879	MSE=186.5377
	MAE=10.44724	MAE=10.4919
	MAPE=37.4003	MAPE=42.2987
	ACF1= 0.0005581	ACF1= 0.0118249
	AIC=70594.35	AIC=70670.58
	ARIMA(1,1,2) model	***ARIMA(1,1,2) model***
Training dataset: 5 years	MSE=198.0541	MSE=204.2407
	MAE=10.67242	MAE=10.79233
	MAPE=41.72928	MAPE=39.83648
	ACF1=0.0001992	ACF1= 8.6798 e-05
	AIC=118720.6	AIC=119170

Table 4.6: Accuracy methods for training datasets after imputation for NO_2, including temperature, humidity and wind velocity as covariates in the imputation

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
	ARIMA(2,1,1) model	***ARIMA(1,1,1) model***
Training dataset: 3 years	MSE=176.8661	MSE=174.962
	MAE=10.29646	MAE=10.24747
	MAPE=46.213	MAPE=48.40924
	ACF1= 0.0015613	ACF1= 0.021046
	AIC=70206.87	AIC=70110.03
	ARIMA(3,1,1) model	***ARIMA(1,1,2) model***
Training dataset: 5 years	MSE=208.3834	MSE=209.6785
	MAE=11.11108	MAE=11.1643
	MAPE=45.19361	MAPE=47.90671
	ACF1=8.675 e-06	ACF1= -0.0005795
	AIC=119466	AIC=119554.5

For the purposes of our analysis, we would prefer studying NO_2 and O_3 for each station separately. This analysis is decided to be undergone for 3 consecutive years $(2010 - 2012)$.

We also implement the same imputation methods and we construct suitable ARIMA models until we find the best ones fitted.

The imputation and time series analysis of O_3 pollutant for all the observed stations is shown below in conclusion. The imputation method chosen is based on best fitted ARIMA model, meaning the one with smaller MSE and AIC and are red - coloured. It is also chosen based on its simplicity (simpler ARIMA models with small p, d and q are easier to explain).

Table 4.7: Time series analysis for stations Geoponiki, Liosia, Nea Smyrni and Piraeus (training dataset=3 years) after imputation for O_3, including temperature, humidity and wind velocity as covariates in the imputation

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
Stations	***ARIMA(1,1,2) model***	***ARIMA(2,1,2) model***
Geoponiki	MSE=149.7408	MSE=150.6177
	MAE=9.6247	MAE=9.6971
	MAPE=43.9514	MAPE=31.70068
	ACF1= 0.0001011	ACF1= -0.0004591
	AIC=8602.39	AIC=8608.5
	ARIMA(2,1,2) model	***ARIMA(2,1,2) model***
Liosia	MSE=165.5898	MSE=167.3235
	MAE=9.81305	MAE=9.8549
	MAPE=21.6765	MAPE=23.1185
	ACF1=0.008833	ACF1=0.001275
	AIC=8714.67	AIC=8725.8
	ARIMA(1,1,2) model	***ARIMA(1,1,2) model***
N. Smyrni	MSE=213.4512	MSE=214.2195
	MAE=11.2747	MAE=11.2315
	MAPE=23.901	MAPE=22.9276
	ACF1=-0.001382	ACF1=-0.001139
	AIC=8932.92	AIC=8936.81
	ARIMA(1,1,1) model	***ARIMA(1,1,1) model***
Piraeus	MSE=211.4866	MSE=199.7064
	MAE=11.1593	MAE=11.0791
	MAPE=41.505	MAPE=43.3183
	ACF1=0.0014026	ACF1=-0.007197
	AIC=8978.66	AIC=8915.87

Table 4.8: Time series analysis for stations Marousi, Peristeri, Athinas and Patisiwn (training dataset=3 years) after imputation for O_3, including temperature, humidity and wind velocity as covariates in the imputation

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
Stations	***ARIMA(2,1,1) model***	***ARIMA(1,1,2) model***
Marousi	MSE=226.6085	MSE=243.8432
	MAE=11.9823	MAE=12.3594
	MAPE=30.8739	MAPE=33.8056
	ACF1=0.002233	ACF1=-0.0006219
	AIC=9055.79	AIC=9136.11
	ARIMA(1,1,2) model	***ARIMA(1,1,2) model***
Peristeri	MSE=178.1392	MSE=190.7559
	MAE=10.3456	MAE=10.7643
	MAPE=27.2275	MAPE=23.7805
	ACF1=-0.0005528	ACF1=0.0001003
	AIC=8792.22	AIC=8867.18
	ARIMA(1,1,2) model	***ARIMA(1,1,2) model***
Athinas	MSE=109.3231	MSE=112.4558
	MAE=8.0536	MAE=8.07973
	MAPE=72.9952	MAPE=33.9552
	ACF1=-0.0001679	ACF1=1.219e-05
	AIC=8257.81	AIC=8288.75
	ARIMA(1,1,2) model	***ARIMA(2,1,1) model***
Patisiwn	MSE=148.0442	MSE=145.7414
	MAE=9.3686	MAE=9.2099
	MAPE=62.5460	MAPE=67.3221
	ACF1=-8.59e-05	ACF1=0.0001529
	AIC=8589.82	AIC=8572.67

The correspondent tables for NO_2 pollutant are given below, following the same hypotheses and modeling as for O_3.

Table 4.9: Time series analysis for stations Geoponiki, Liosia, Nea Smyrni and Piraeus (training dataset=3 years) after imputation for NO_2, including temperature, humidity and wind velocity as covariates in the imputation

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
Stations	***ARIMA(2,1,2) model***	***ARIMA(2,1,2) model***
Geoponiki	MSE=165.2317	MSE=161.7043
	MAE=10.385	MAE=10.2752
	MAPE=37.0458	MAPE=36.3445
	ACF1= -0.001381	ACF1= -0.002676
	AIC=8713.48	AIC=8689.9
	ARIMA(1,1,2) model	***ARIMA(2,1,2) model***
Liosia	MSE=120.4405	MSE=119.9827
	MAE=8.7026	MAE=8.6804
	MAPE=49.8253	MAPE=49.9819
	ACF1=-0.001609	ACF1=-0.002654
	AIC=8364.46	AIC=8360.31
	ARIMA(1,1,2) model	***ARIMA(1,1,1) model***
N. Smyrni	MSE=153.2934	MSE=157.5273
	MAE=9.6162	MAE=9.7008
	MAPE=57.1765	MAPE=55.234
	ACF1=-0.0009018	ACF1=0.01285
	AIC=8573.18	AIC=8600.84
	ARIMA(1,1,1) model	***ARIMA(1,1,1) model***
Piraeus	MSE=120.2233	MSE=125.4519
	MAE=8.6564	MAE=8.8112
	MAPE=23.7342	MAPE=25.3373
	ACF1=0.001912	ACF1=-0.01117
	AIC=8360.7	AIC=8407.42

Table 4.10: Time series analysis for stations Marousi, Peristeri, Athinas and Patisiwn (training dataset=3 years) after imputation for NO_2, including temperature, humidity and wind velocity as covariates in the imputation

Imputation methods	TS imputation method with polynomials of time	TS imputation method including ridge prior for high missingness
Stations	***ARIMA(1,1,2) model***	***ARIMA(1,1,2) model***
Marousi	MSE=135.9852	MSE=132.3015
	MAE=9.1971	MAE=9.0772
	MAPE=71.9039	MAPE=86.2664
	ACF1=-0.003975	ACF1=-0.002925
	AIC=8497.44	AIC=8467.39
	ARIMA(1,1,2) model	***ARIMA(1,1,2) model***
Peristeri	MSE=134.1675	MSE=149.1041
	MAE=9.0908	MAE=9.5288
	MAPE=41.872	MAPE=42.2776
	ACF1=-0.003884	ACF1=-0.001988
	AIC=8482.9	AIC=8598.55
	ARIMA(1,1,2) model	***ARIMA(2,1,2) model***
Athinas	MSE=111.0823	MSE=112.8345
	MAE=8.1228	MAE=8.1479
	MAPE=18.4628	MAPE=18.4843
	ACF1=-0.002321	ACF1=-0.0006003
	AIC=8275.64	AIC=8294.8
	ARIMA(1,1,2) model	***ARIMA(1,1,2) model***
Patisiwn	MSE=316.9037	MSE=333.4002
	MAE=14.0285	MAE=14.3628
	MAPE=26.9314	MAPE=23.4269
	ACF1=-0.003492	ACF1=-0.001918
	AIC=9424.06	AIC=9479.64

As it is mentioned above, the best fitted ARIMA models were chosen based on MSE and AIC values. ARIMA models for both imputation methods were quite similar in most of the stations. The smallest MSE values are observed in ARIMA models for station in Athinas which probably means that the autocorrelation of the specific data is quite high. It is well known that MSE depends on r^2, and the higher the correlation, the lower the MSE. This becomes more obvious by the following formula:

$$MSE = (n-1/n-2)\, S_y^2(1-r^2)$$

Additionally, a relatively low AIC value of an ARIMA model means that the specific model is well – fitted (see applied ARIMA models of NO_2 and O_3 for Athinas station). Besides, AIC is mostly a measure of goodness – of – fit.

ARIMA models are usually fitted by using the three general steps below:

1) Identification of the model
2) Estimation of the model
3) Diagnostic check of the model

4.4.1. Identification of the model

At the identification stage, the model structure is selected by comparing the ACF and PACF (a sample of them is given above) with respective theoretical profiles. More specifically, for NO2 in station Geoponiki, both ACF and PACF tailed off at lag 2 suggesting using an ARIMA $(2, 1, 2)$ model. Most time series in practice can be adequately modeled by an ARIMA (p, d, q) model with p, d and $q \leq 2$, as simpler models are easier to explain and evaluate.

Nevertheless, we used the *auto.arima* function in R (package "forecast") to fit a model, with the function above taking into consideration all the identification techniques.

4.4.2. Parameter estimation of our model

With the implementation of ARIMA models for each station and pollutant separately (NO_2 and O_3 respectively) comes also the estimation of each model's parameters. In Tables 4.11 and 4.12 that follow, parameter estimation of NO_2 and O_3 concentrations is illustrated for each station studied.

Table 4.11: Parameter estimation of ARIMA models for NO_2 concentrations for each station separately

Station	ARIMA model (p, d, q)	φ_p	θ_q
Geoponiki	(2, 1, 2)	-0.111, 0.267	-0.318, -0.657
Liosia	(2, 1, 2)	0.153, 0.061	-0.540, -0.401
Nea Smyrni	(1, 1, 2)	0.330	-0.827, -0.103
Piraeus	(1, 1, 1)	0.317	-0.961
Marousi	(1, 1, 2)	0.246	-0.657, -0.288
Peristeri	(1, 1, 2)	0.276	-0.632, -0.325
Athinas	(1, 1, 2)	0.496	-0.932, -0.025
Patisiwn	(1, 1, 2)	0.359	-0.753, -0.209

Table 4.12: Parameter estimation of ARIMA models for O_3 concentrations for each station separately

Station	ARIMA model (p, d, q)	φ_p	θ_q
Geoponiki	(1, 1, 2)	0.386	-0.348, -0.043
Liosia	(2, 1, 2)	1.394, -0.416	-1.905, 0.911
Nea Smyrni	(1, 1, 2)	0.154	-0.609, -0.225
Piraeus	(1, 1, 1)	0.253	-0.912
Marousi	(2, 1, 1)	0.491, -0.122	-0.878
Peristeri	(1, 1, 2)	0.194	-0.639, -0.202
Athinas	(1, 1, 2)	0.206	-0.754, -0.127
Patisiwn	(2, 1, 1)	0.622, -0.092	-0.934

An example of interest is that of O_3 concentrations model for Liosia station. The first autoregressive parameter $\boldsymbol{\varphi_1}$ is greater than 1, but for a AR(2) process the stationarity conditions are applied and verified, as $1.394 + (-0.416) = 0.978 < 1$. Note also that $-0.416 - 1.394 = -1.810 < 1$, thus the process is stationary. As far as the MA part of the process is concerned, MA processes are always stationary. If we check the process for invertibility, we will observe a rather unnatural situation. According to the invertibility conditions for a MA(2) process, $\theta_2 - \theta_1 < 1$, which is not valid in our case as $0.911 - (-1.905) = 2.816 > 1$. This means that the weight given to observation Y_{t-j} increases with j and the influence of previous $Y_t's$ on the current observation increases with age. It is worth

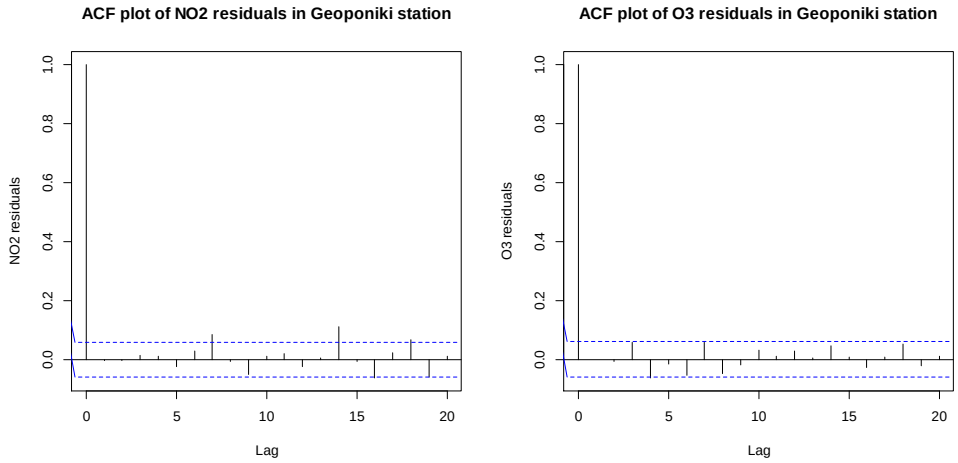
mentioning that in the case the assumption of invertibility is not satisfied, ACF plot cannot give us information about the model, as there is probably another (better-fitted) model with the same ACF pattern (not 1-1 relationship between model-ACF plot).

Having a look at the rest of the ARIMA models and their parameters, conditions of stationarity and invertibility are satisfied. It is worth mentioning that time series appear to be stationary after taking first differences for each of the processes above. Overdifferencing the series may cause other problems, i.e. the variance of the overdifferenced model will be inflated, thus obscuring the actual structure of the data. Besides, it is illogical to give more weight to older than to more recent observations.

4.4.3. Diagnostic testing

After specifying the model parameters, the diagnosis testing is carried out to assess whether the autocorrelation of the original NO_2 and O_3 measurements has been eliminated adequately. This is done by analyzing the ACF of the residuals. Autocorrelation is considered to be removed if the ACF of the residuals does not differ significantly from zero for all lags greater than ten. Otherwise, the ARIMA modeling procedure is repeated until no autocorrelation is found. Figure 4.2 illustrates the ACF plots of NO_2 and O_3 residuals respectively as an example of autocorrelation elimination.

Fig 4.2: ACF plot of NO_2 and O_3 residuals of Geoponiki station



Ljung – Box – Pierce statistic is also used as a more valid statistical test to examine if there is any remaining autocorrelation in the residuals. In Table 4.13 and Table 4.14 it is illustrated that, neither NO_2 residuals, nor O_3 residuals show any sample of remaining autocorrelation (p-value of Box – Lung statistic is greater than 0.05 as estimated by the function *Box.test* in R). Therefore, we can now proceed in forecasting the ARIMA models in order to construct the control charts.

Table 4.13: Box – Ljung test of the NO_2 residuals performed for each station

Station	Box – Ljung* test (p-value)
Geoponiki	$p = 0.248$
Liosia	$p = 0.740$
Nea Smyrni	$p = 0.538$
Piraeus	$p = 0.723$
Marousi	$p = 0.127$
Peristeri	$p = 0.614$
Athinas	$p = 0.669$
Patisiwn	$p = 0.647$

*Null hypothesis of the test: $H_0: \rho_i = 0,\ i = 1, \dots, K$

Table 4.14: Box – Ljung test of the O_3 residuals performed for each station

Station	Box – Ljung* test (p-value)
Geoponiki	$p = 0.690$
Liosia	$p = 0.377$
Nea Smyrni	$p = 0.549$
Piraeus	$p = 0.266$
Marousi	$p = 0.699$
Peristeri	$p = 0.688$
Athinas	$p = 0.869$
Patisiwn	$p = 0.125$

*Null hypothesis of the test: $H_0: \rho_i = 0,\ i = 1, \dots, K$

4.4.4. Forecasting the fitted residuals

The autocorrelation assessment results revealed that all the NO_2 and O_3 time series were autocorrelated. After ARIMA modeling, non-autocorrelated fitted residuals were obtained for all the mean NO_2 and O_3 concentrations. The approach followed on our case study is to estimate the model on a single set of training data, and then compute one-step forecasts on the remaining test data. This can be handled by applying the fitted model to the whole data set,

and then extracting the "fitted values" which are simply one-step forecasts. As far as our data is concerned, we chose to compute one – step forecasts without re - estimation of all coefficients.

Therefore, we have i.e for NO2 concentrations in Geoponiki station a fitted ARIMA $(2, 1, 2)$ model for years $2010 - 2012$ ($n = 1096$ measurements) and we evaluate year 2013 (365 days) of model forecasts, using the fitted ARIMA model for the training dataset as the forecast origin.

An alternative approach is to extend the training data and re-estimate the model at each iteration, before each forecast is computed. This is what is called "time series cross-validation" because it is analogous to leave-one-out cross-validation for cross-sectional data. In our case, when we re – estimate an ARIMA model to $[1: 1097]$ (our data for 3 years are 1096 total), fitted values will essentially be from a different model coefficient than an ARIMA model estimated on $[1: 1096]$ points. If we forecast the next 365 values (year 2013) with refitting the model but also using the latest data available for each date we have:

Forecast for $(t + 1)$ uses the model fitted from 1 to t and the time series from 1 to t

Forecast for $(t + 2)$ uses the model fitted from 1 to t and the time series from 1 to $t + 1$

…….

and so on.

After obtaining one – step – ahead forecasts for the original data $\hat{y}_h,\ h = 1, \dots, 365$ we would like to assess the model performance by comparing these forecasts with daily concentrations observed during 2013 (y_h). This is done by evaluating daily forecast errors $(e_h = y_h - \hat{y}_h)$.

MSE of the forecasted errors is now almost 5% smaller in all NO_2 and O_3 datasets than the MSE calculated for the residuals of the training dataset. For MAPE as an accuracy method, the same conclusion applies. Mean absolute percentage error is 8% smaller (i.e for NO_2 forecasted errors in Geoponiki station) in relation to the respective measure in the historical dataset.

Chapter 5

Statistical Process Monitoring of Time Series Residuals

5.1. Introduction

Statistical process monitoring (SPM) is the collection of methods for recognizing special causes and bringing a process into a state of control and reducing variation about a target value. The need of monitoring specific processes led to its great development and improvement.

The most valuable tool of SPM is control charts. In this chapter we apply well-known control charts to the forecast errors and the limits of these charts are computed based on the forecast errors of the training period (Phase I). Prediction errors for the test data (January 2013 – December 2013) are plotted in the same chart based on the in – control (when achieved) mean and standard deviation of the training sample (Phase II). A large difference between the observed and predicted value implies an unusually large forecasting error, and the corresponding observation will be flagged as an outlier in the control chart of the forecast errors. These outliers and more specifically the positive forecast errors are of scientific interest and will be under examination at this chapter.

5.2. Introduction to Statistical Process Control

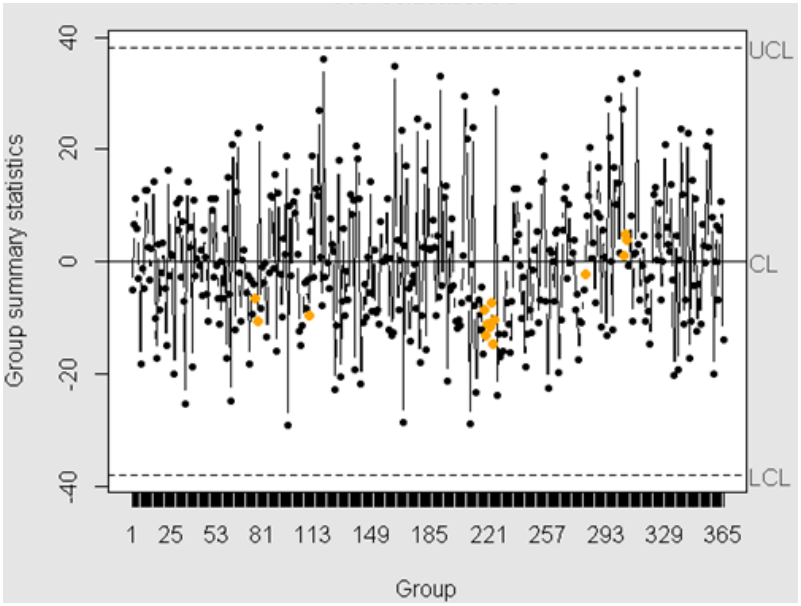
SPC charts give a graphical representation of the process providing the ability to any manager with or without the knowledge of statistics to immediately understand if the process is in-control or not. The wide use and popularity of control charts stems from many reasons. First of all, it is proved that they are able to improve productivity and decrease the cost. Their effective prevention of defect items is also valuable. The use of control charts helps to keep the process under control. Finally, the diagnostic information of control charts is significant as it allows for changes in the process. The research in the area of control charts is active for decades now. Although someone would expect that there would be a decreasing interest in this area after all these years, we observe exactly the opposite. There is an increasing interest for this tool since it has proved its value in practice.

When we have a production process there is usually a target value. We want our process to achieve this target for every product. However, in every process there is an inherent random variability. Therefore, no matter how good we design the whole procedure, we expect to be close to the target value but not always on this value. The existence of this variability affects our process. There are two different "versions" of this variability. The common cause (chance cause) variability is the natural variability every process experiences. Its existence is due to randomness as we can find purely random variability from one product to another. A process that operates with only common cause variability is said to be in-control. The special cause (assignable cause) variability is a result of factors that are not purely random. These factors cause heterogeneity in the process and as a result they affect it, leading to low quality product. A process that operates in the presence of special causes of variability is said to be out-of-control. This type of variability can be detected with control charts giving us the ability to remove its effect and reduce the overall variability. As a result, removing special causes leads to an improvement of the quality of the product. Common cause variability is what remains of the variability after every component of special cause has been removed. In order to remove common cause variability we have to alter the process itself.

Special causes of variability can be divided in two different groups; transient special causes and persistent special causes. Transient special causes are those causes that affect a process for a short time until their reappearance in a future point in time. Persistent special causes are those causes that when they occur they stay in the process until they are detected and removed.

A control chart is a graphical representation of a characteristic of the process under investigation. It is used as the main tool to identify special causes of variability in a process. On the horizontal axis we have the number of the sample drawn from the process or the time that the sample was inspected. On the vertical axis we have the value of the characteristic measured for each sample or for the time of the horizontal axis. A straight line connects the successive points indicating the level of the characteristic in time or in successive samples. There are also three usually straight lines that stand for the upper control limit (UCL) the center line (CL) and the lower control limit (LCL). An example of a control chart is given in Figure 5.1. This chart depicts a random choice of data from Geoponiki station:

Figure 5.1: $\overline{X}$ control chart of random NO_2 daily measurements from station Geoponiki



We assume that a process operates under control when the line connecting the sequence of points does not cross UCL or LCL. When a point is plotted outside these limits we assume that the process is in an out-of-control state and corrective actions must be taken in order to remove the assignable cause that led to this problem. The values of UCL and LCL are chosen usually in such a way that when the process is in-control the probability of a point plotting outside these limits is very small. However, there are some cases that even when all the points plot inside the control limits we characterize the process as being in an out-of-control state. Such cases are for example when we see a series of nine successive points plotting all above (below) the center line or when we see six successive points in a row steadily increasing or

decreasing. We have to state here that the removal of any cause is not the objective of a control chart. A control chart simply indicates that an assignable cause may exist.

In the literature, two distinct phases of control charting practice have been discussed. In Phase I, charts are used for retrospectively testing whether the process was in-control when the first subgroups were being drawn. In this phase, the charts are used as aids to the person who implement them, in bringing a process into a state of statistical control. Once this is accomplished, the control chart is used to define what is meant by statistical control. This is referred as the retrospective use of control charts. During this phase we are studying the process very intensively. The data collected are then analyzed in order to answer the question "were the data collected are from an in-control process?". In Phase II, control charts are used for testing whether the process remains in-control when future subgroups are drawn. In this phase, the charts are used for monitoring the process for any change from an in-control state. At each sampling stage, we ask the question "has the state of process changed?". The meaning of in-control, in this phase, is usually determined by the values of the process parameters e.g., the mean and standard deviation for univariate continuously distributed variables. The values of the parameters are either given to the practitioner or they are estimated from the historical data known to be under control from Phase I. Note that in this phase the data is not taken as being from an in-control process unless the data provide evidence against no change in the process. Using these data to define what is meant by the process being in-control might lead to use an out-of-control process to define a state of statistical in-control.

In a control chart we have two objectives. Firstly, when a process is in-control, we want our chart to signal very rarely (low number of false alarms). In statistical terms we want the chart to operate with the planned probability of the statistic computed to plot outside the control limits if we are in-control. Secondly, when a process is out-of-control, we want the chart to signal as soon as possible. In statistical terms we want the probability of the statistic computed to plot in-control if we are out-of-control to be as small as possible. Different measures for evaluating the performance of a chart, concerning the previous two objectives, have been proposed. The most known measure is the average run length (ARL), which is based on the run length (RL) distribution. The number of observations (individual data), or samples (data in subgroups), needed for a control chart to signal is a run length or alternatively one observation of the RL distribution. The mean of the RL distribution is the

ARL, which is actually the average number of observations needed for a control chart to signal. Usually, along with the ARL, the standard deviation of the run length (SDRL) is computed. Alternatively, the ARL is expressed as the average number of observations to signal (ANOS). A measure similar to the ARL is the average time to signal (ATS), which is the average time needed for a control chart to signal and it is actually a product of the ARL and the sampling interval used in the case of fixed sampling.

From the preceding discussion we see that all these measures are related to the ARL. However the sole use of the ARL has been criticized (see, e.g. Woodall (1983)). The disadvantage of the ARL is the skewness of the run length distribution in the out-of-control case and in non-normality and as a result the misleading conclusions one can draw based on the ARL. An alternative measure is the median run length (MRL), which is more credible since it is less affected by the skewness.

5.3. Univariate Shewhart Control Charts

The most known control charts are the Shewhart type control charts. They owe their name to Walter Shewhart who established them in 1931. They are used to detect transient special causes in a process. This property is the result of the fact that Shewhart Control Charts are memoryless. In the following we present some of the Shewhart control charts for variables.

5.3.1. Control charts for the mean for data in subgroups

Assume that we have a variable that is normally distributed with mean μ and standard deviation σ. We assume that μ and σ are both known. Let $x_1, x_2, \dots, x_n$ be a sample of $n$independent and identically distributed observations drawn from our production process. Then the average of this sample $\bar{x}$ is distributed as a normal variable with mean μ and standard deviation $\sigma/\sqrt{n}$. Therefore, we can use as control limits for each sample

$$UCL = \mu + Z_{a/2}\sigma/\sqrt{n}$$

$$LCL = \mu - Z_{a/2}\sigma/\sqrt{n}$$

where UCL and LCL are the upper and lower control limits respectively, $Z_{a/2}$ is the inverse of the normal cumulative distribution function for probability $\alpha/2$ and α is the probability

that an in-control sample will plot outside these limits. If all the points (samples) plot inside the control limits we claim that we have an in-control process. This plot is a Phase II Shewhart chart for the mean.

However, in real world we usually do not know the values of μ and σ . Consequently, we have to estimate them. Therefore, the control limits in such a case will not be fixed numbers, but rather random variables. The control limits in this case for Phase I Shewhart chart for the mean are

$$\widehat{UCL} = \hat{\mu} + k\hat{\sigma}/\sqrt{n}$$
$$\widehat{LCL} = \hat{\mu} - k\hat{\sigma}/\sqrt{n}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the estimates for the mean and the standard deviation, respectively and k is a constant used to specify the width of the control limits usually taken to be equal to 3. If a point plots above $\widehat{UCL}$ or below $\widehat{\text{LCL}}$ we have an indication that this point (sample) is from an out-of-control process. Let $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$ be the sample means from samples each with n observations. Then, an estimate for the mean is $\hat{\mu} = \bar{\bar{x}}$, the average of all the sample means. If the process is in-control this estimator is normally distributed with mean μ and variance σ^2/mn . For the standard deviation three different estimators have been proposed. The first one is based on the range. Let $R_1, R_2, \dots, R_m$ denote the range for each of the m samples and R the average of these ranges. Then, a control charts' unbiased estimator is given by $\bar{R}/d_2$. The estimated control limits for the $\bar{X}$ chart are given by

$$\widehat{UCL} = \bar{\bar{x}} + k\bar{R}/(d_2\sqrt{n})$$
$$\widehat{LCL} = \bar{\bar{x}} - k\bar{R}/(d_2\sqrt{n})$$

where d_2 is the mean of the random variable R/σ and is a function of the sample size$n$. Details on the derivation of d_2 along with its values for different sample sizes can be found in textbooks, see e.g. Montgomery (2001).

A second version of the estimated control limits for the mean is based on a different unbiased estimator for the standard deviation. Let $S_1, S_2, \dots, S_m$denote the standard deviation for each of the m samples and $\bar{S}$ their average. An unbiased estimator for σ is $\bar{S}/c_4$. The control limits will be

$$\widehat{UCL} = \bar{\bar{x}} + k\bar{S}/(c_4\sqrt{n})$$
$$\widehat{LCL} = \bar{\bar{x}} - k\bar{S}/(c_4\sqrt{n})$$

The ARL for Shewhart charts is given by

$$ARL = 1 - P_r \text{ (a point plots outside the control limits)}$$

It has to be stressed though that this relationship holds in the case of known parameters. If the parameters are unknown and as a result they have to be estimated a different relationship holds.

5.3.2. Control charts for individual data

Let $X_i,\ i = 1, \dots, n$ represent independent and identically distributed observations from a $N(\mu, \sigma^2)$ process. If the parameters μ and $\sigma^2$are known, the Phase II X chart control limits are

$$UCL = \mu + 3\sigma$$
$$LCL = \mu - 3\sigma$$

Usually, these parameters are not known and they have to be estimated. In this case, the variability is usually controlled using moving ranges. Nevertheless, Rigdon et al. (1994) has recommended either against the use of the moving range chart or its use together with the classical Xchart. Moreover, it is shown that a moving range control chart does not contribute significantly to the identification of out-of-control situations. For these reasons we do not present it here. Therefore, the use of the X control chart for monitoring both the process mean and standard deviation is recommended. The Phase I control limits of the X control chart are

$$\widehat{\text{UCL}} = \bar{\bar{X}} + 3\hat{\sigma}$$
$$\widehat{\text{LCL}} = \bar{\bar{X}} - 3\hat{\sigma}$$

where $\bar{\bar{X}}$ is an unbiased estimate of the mean of the process and $\hat{\sigma}$ is an estimate of the standard deviation σ of the process. Usually, the estimate of the standard deviation used is $\overline{MR}/d_2$ where $\overline{MR}$ denotes the average of the moving ranges and d_2 is the usual function of the sample size n used to make the estimator unbiased. However, a preferable estimate of σ is s/c_4 where c_4 is defined the same way as in the case of rational subgroups and s is the standard deviation of the observations. Sullivan and Woodall (1996a) proposed a Phase I control chart for independent observations that uses the log-likelihood function and is used to detect shifts in both the mean and the variance. This chart is shown to have better performance in comparison to the X chart or the combined X and MR chart. Moreover, it performs well for detecting sustained shifts in the distribution but not that well for outliers.

5.3.3. CUSUM and EWMA charts

Other widely known control charts with multiple applications especially in time series data are the CUSUM and EWMA charts. CUSUM charts are used to identify persistent causes in a variable instead of Shewhart charts. This ability is attributed to the fact that they have a memory as they are based on successive sums of the observations minus a constant. Generally, we can say that CUSUM charts are able to detect small to moderate shifts whereas Shewhart charts are able to detect large shifts. The optimality of the CUSUM is for detecting a shift to a single specific out-of-control distribution. The CUSUM that is optimal for detecting one particular shift is not optimal for detecting a different shift. For a different shift a different CUSUM will be optimal. However, while a CUSUM for detecting a shift of one standard deviation is optimal only for this shift, it performs nearly as well as the optimal CUSUM for all shifts that are not too far from one standard deviation. The EWMA chart is used as the CUSUM chart to detect persistent shifts in a variable. Its ability is to signal faster than the Shewhart charts for small and moderate shifts but not that fast for large shifts. Generally, we can say that its performance is similar to the performance of the CUSUM chart.

5.4. Performance of Control Charts for Uncorrelated Data

A classical SPC chart can be thought of as a sequence of tests of hypothesis where the goal is to find evidence against the hypothesized Shewhart model. Therefore, it is natural to think of the probabilities associated with type I and type II errors as performance measures of how well the chart works. Consider the simple case of a Shewhart Y chart. For this chart, the probability of a false alarm is

$$\alpha = P\{\textit{Type I error}\} = P\ \{\textit{one } Y_i \textit{ falls outside limits} \mid \textit{process is in control}\}$$

Since $Y \sim N(\mu, \sigma^2/n)$, assuming that the in-control mean is some known value μ_0, we have

$$a = 2P\{\bar{Y} < LCL | \mu = \mu_0\} = 2P\left\{\frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{LCL - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right\} = 2\Phi(-k)$$

where $\Phi(.)$ denotes the standard normal distribution function. Thus the false alarm rate is determined exclusively by the width of the control limits.

5.5. Effect of Autocorrelation on SPC Chart Performance

SPC control charts assume that if in control, the process has a constant mean and is completely uncorrelated. An important practical question is to investigate what happens with the performance of SPC charts as the process exhibits more and more serial correlation. Positive autocorrelation at low lags is common because given the advances in sensor technologies, measurements are taken closer together in time. In discrete-part manufacturing, this sometimes implies that every part is measured. Observations that were generated close in time will tend to be similar; hence positive correlation at low lags will result.

5.6. Time-Series Approaches to SPC of Autocorrelated Data

Many have proposed modeling the inherent autocorrelation that exists in process control data obtained in modern manufacturing. The argument is that if the autocorrelation is considered as normal for the process, once an ARIMA (p, d, q) model is to fit to the data, the model becomes the in-control model. As long as the new observations behave according to the model, one-step ahead forecast errors (similar to the residuals but computed as the difference between prediction by model and actual) should form a white noise sequence. Thus the forecast errors would conform to the assumptions behind Shewhart's model and we would be able, in principle, to detect shifts and other unpredictable disturbances typical of SPC. The main problem with this approach is that the models are just approximations, so the monitoring scheme will be sensitive to estimation errors in the parameters.

A similar approach was proposed by Montgomery and Mastrangelo (1991), who monitors the one-step-ahead forecasts of the process on the same plot as the data, using the IMA $(1, 1)$ as a generic model and time-varying control limits. The IMA $(1, 1)$ model $Y_t = \theta\varepsilon_{t-1} + \varepsilon_t$ is fit to the data by estimating the value of θ using a least squares criterion. An EWMA with parameter $\lambda = 1 - \theta$ is then used as a generic forecasting technique for the process. Although this approach appears to suffer from the same problems as the first approach mentioned

above, a number of additional monitoring devices based on tracking signals have been tested extensively and seem to improve the performance of the method. Obviously, the method works better the closer the process is to an IMA$(1, 1)$ process [e.g., when we have an ARMA $(1, 1)$ process with a value of φ close to1].

An alternative to the previous approaches is to monitor the original autocorrelated data. Unless ρ_1 is large, shift detection in the original series is better accomplished with a CUSUM chart applied to the series itself rather than to the forecast errors of some model. If the autocorrelation is low to moderate, there is no significant difference between monitoring the residuals and monitoring the observations directly using a standard SPC chart with modified limits. In fact, for large shifts, a Shewhart chart of the observations is better than a Shewhart chart on the residuals in terms of ARLs. For small shifts, an EWMA of the observations slightly outperforms an EWMA of the residuals EWMA charts and CUSUM charts perform similarly in the presence of autocorrelation.

Regardless of the remedial actions taken to monitor an autocorrelated process, we have to be careful in reading performance statistics related to charts for this type of process. The effect of autocorrelation on a control chart scheme depends crucially on (1) the actual model that the process really follows, (2) the model that we consider to describe in-control operation, (3) the way we estimate the in-control model, and (4) the type of disturbance that the chart is supposed to detect. Considerations 1 to 3 are particularly important because the effect if the uncertainties in the parameter estimates and in the form of the estimated model will affect the performance of the chart procedure. A very large data set is necessary before the performance of an EWMA chart with estimated process parameters approaches that of a chart based on known parameters. This applies to the case when the residuals are monitored, and perhaps more important, it also applies to the case when the limits are adjusted using the parameter estimates.

5.7. Data Analysis

In our case study, the statistical process control chart is used to monitor the real-time NO_2 and O_3 concentrations time series and detect the occurrence of any alarming values of the pollutants, but this part will be discussed below in detail. The statistical process control chart is based on the ARIMA models that are estimated in a previous chapter whereas the $\overline{X}$ control charts are constructed following three steps: (1) autocorrelation assessment, (2) ARIMA modeling and (3) control chart construction.

Historical data of NO_2 and O_3 concentrations are used to establish the Shewhart statistical process control chart for fall detection by specifying the individual-specific control limits (CLs) based on the Shewhart three sigma control theory. The historical data for each pollutant are the forecasted residual time series ε_t obtain from ARIMA models after one – step – ahead forecasting without re – estimating the model fitted for the training dataset $(2010 - 2012)$.

5.7.1. Evaluation and results

The autocorrelation assessment results in Chapter 4 revealed that all the pollutants' time series were autocorrelated. After ARIMA modeling, non - autocorrelated fitted residuals were obtained for all the NO_2 and O_3 time series. The control chart needs to be robust, meaning that the presence of outliers in the series should not harm its performance. Outliers may be present in the "training period", being the part of the series used to determine the control limits, or in the "test period", being the part of the series where outliers should be flagged as alarm observations.

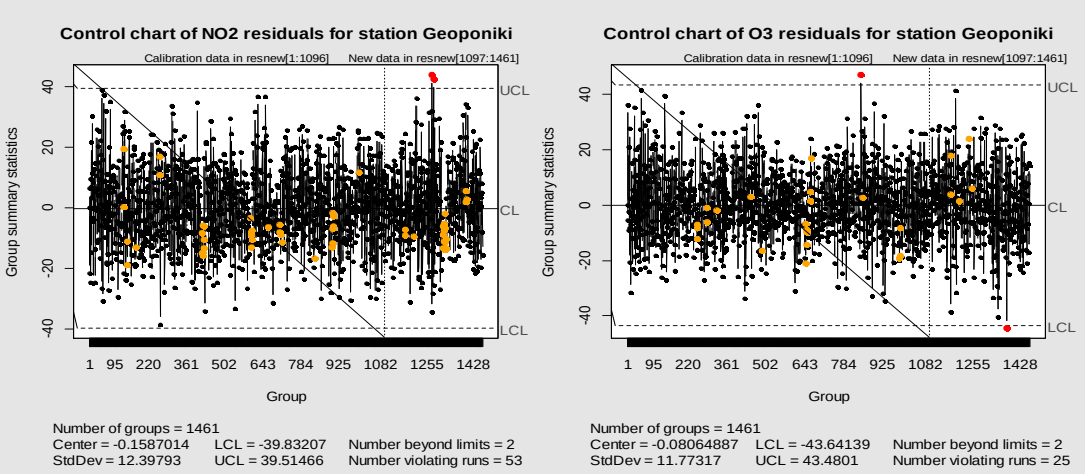
The control chart we propose, the $\bar{X}$ chart for individual data, belongs to the class of special-cause control charts, where forecast errors of the ARIMA prediction method are subject to regular control chart techniques. If an observation has a large deviation from its predicted value, an unexpected event has occurred, and an alarm should be given. A large difference between the observed and predicted value implies an unusually large forecasting error, and the corresponding observation will be flagged as an outlier in the control chart of the forecast errors.

Control limits to monitor the new incoming observations (concentrations of the pollutants for 2013) are constructed from the training sample. Since the size of the collected NO_2 and O_3 data, and therefore the size of the forecasted errors is undoubtedly big, historical limits are not difficult to estimate.

The package "qcc" from R is used for most of the calculations, whereas the package "spc" was considered for additional research purposes (as far as other rules from out of control limit points are considered).

Fig 5.2 illustrates the results of using $\bar{X}$ chart in monitoring the behavior of NO2 and O_3 forecasted errors of Geoponiki station. The vertical dashed line in the chart indicates the end of the training period. It is of crucial importance that the process is completely in control during the training period, otherwise the forecast errors are inflated by the presence of outliers in the training period. In case any of the control charts constructed is not completely in control during $2010 - 2012$, outliers in the test period (2013) may lead to wider control limits.

Fig 5.2: Control chart of NO_2 and O_3 forecast errors for station Geoponiki



As it is shown in the figure above, the control chart of NO_2 forecast errors seems to be completely in control for $2010 - 2012$ (training dataset), thus the forecast errors are unbiased. Changes in the underlying process have occurred. There are two time points where there is a significant departure of forecasts from future observations. Individual forecast errors that could signal an alarm are 43.92 and 42.59 for –as it is found by the time series– June

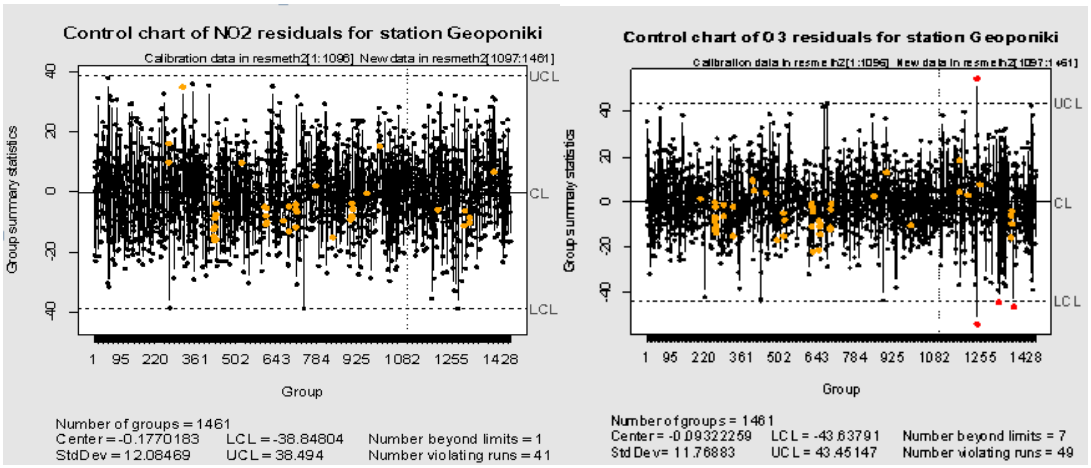
27^{th} 2013 and July 1^{st} 2013 respectively. Looking back on the observed values of NO_2 in Geoponiki station, we notice that in June 27^{th} the true value of NO_2 was 80.35 while it was predicted to be 36.43. The same applies for the July 1^{st} where the true value of NO_2 was 59.03 while it was predicted to be 16.44. It is widely known that the $\bar{X}$ chart can detect faster large shifts from the mean than other charts.

As far as the control chart of O_3 forecast errors is concerned, the in – control assumption of Phase I data is not satisfied, thus the control limits estimated by training data may be wider than usual. In Phase II (O_3 concentrations during 2013) there are no forecast errors out of the upper control limit, but just one out the LCL (the predicted value of O_3 at this exact day was higher that the real observed one). Note that sudden increases in daily O_3 concentrations (positive forecast errors) are more "acceptable" than sudden decreases (negative forecast errors), so no focus on positive e_t will be given.

Another forecasting approach that was mentioned in Chapter 4 is to re-estimate the fitted model at each iteration, before each forecast is computed (one – step – ahead forecasts with re – estimation). We will use this approach as an example to Geoponiki station to compare the control charts generated with the charts studied above (one – step – ahead forecast errors without re – estimation of the model).

Figure 5.3 illustrates the control charts of NO_2 and O_3 forecast errors for station Geoponiki with re – estimation of the model. The control charts of NO_2 and O_3 forecast errors seem to be completely in control for $2010 - 2012$ (training dataset – Phase I). There are no changes occurred in NO_2 chart, as there is not any alarming deviation between observed and predicted concentration during 2013 (Phase II data). As far as the control chart of O_3 forecast errors is concerned, there is one positive forecast error out – of – control equal to 54.46 observed at June 3^{rd} 2013. The actual O_3 concentration value of the time series at this time point is 89.83 while the respective predicted value is 35.37. Again, negative observed forecast errors are of no value to our case study.

Fig 5.3: Control chart of NO_2 and O_3 forecast errors for station Geoponiki with re – estimation of the model at each iteration.



Note that generated forecast errors by re – estimating the fitted model at each step for each daily concentration of NO_2 and O_3 respectively are more adaptive and take into consideration the whole course of time series values when forecasting one – step – ahead for each time point. We also observe that the limits appear to be slightly tighter than those of the previous method.

Control charts for the rest of the stations (Marousi, Peristeri, Nea Smyrni, Athinas, Liosia, Patisiwn, Piraeus) for both NO_2 and O_3 forecast errors are given in figures 5.4 until 5.10 below.

Fig 5.4: Control chart of NO_2 and O_3 forecast errors for station Marousi

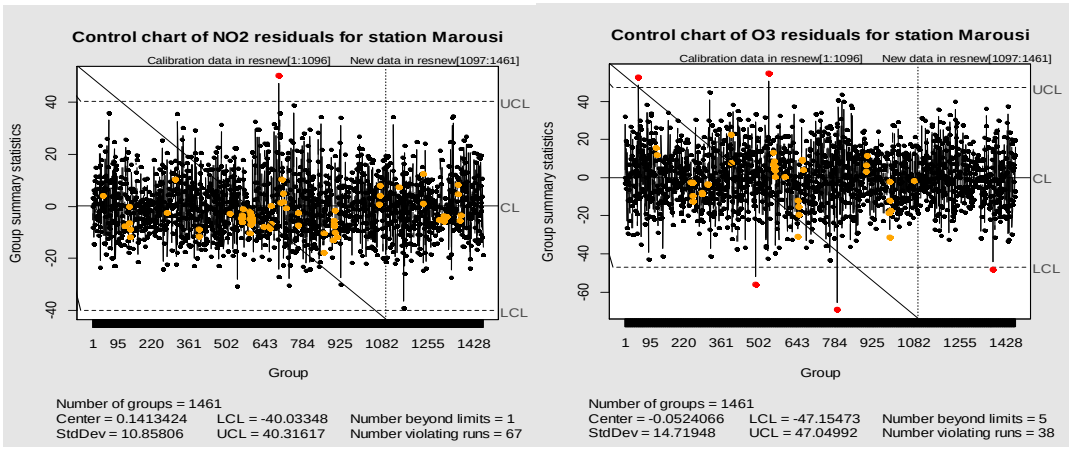


Fig 5.5: Control chart of NO_2 and O_3 forecast errors for station Peristeri

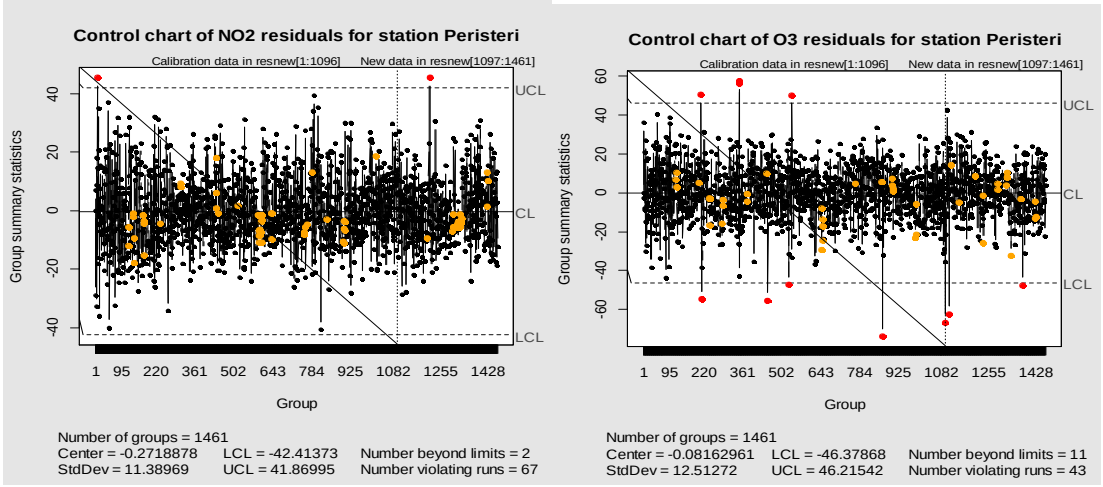


Fig 5.6: Control chart of NO_2 and O_3 forecast errors for station Nea Smyrni

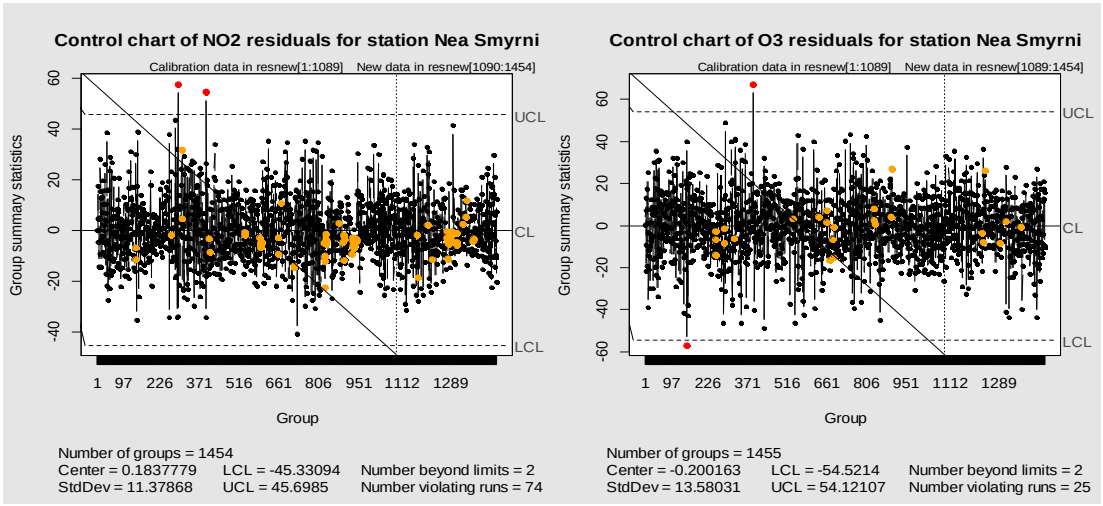


Fig 5.7: Control chart of NO_2 and O_3 forecast errors for station Athinas

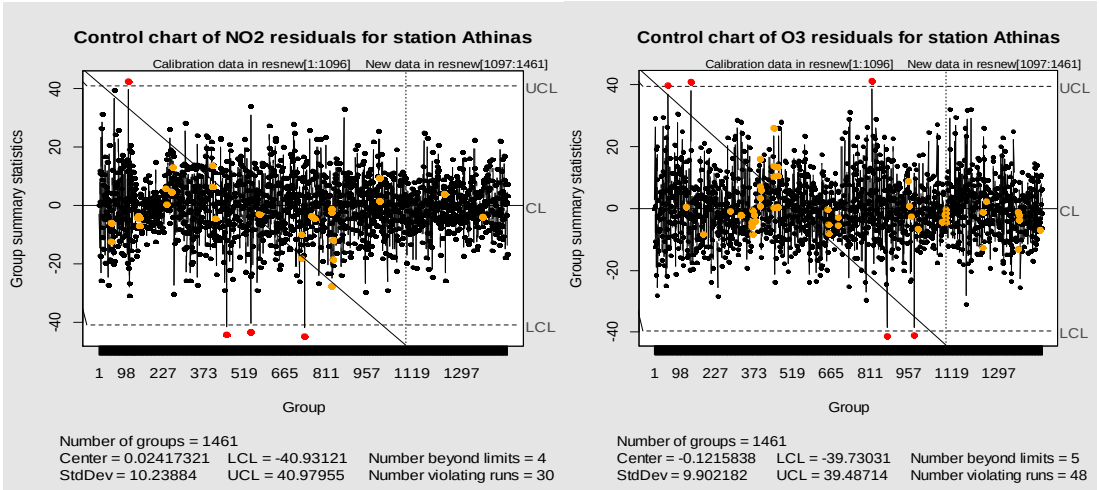


Fig 5.8: Control chart of NO_2 and O_3 forecast errors for station Liosia

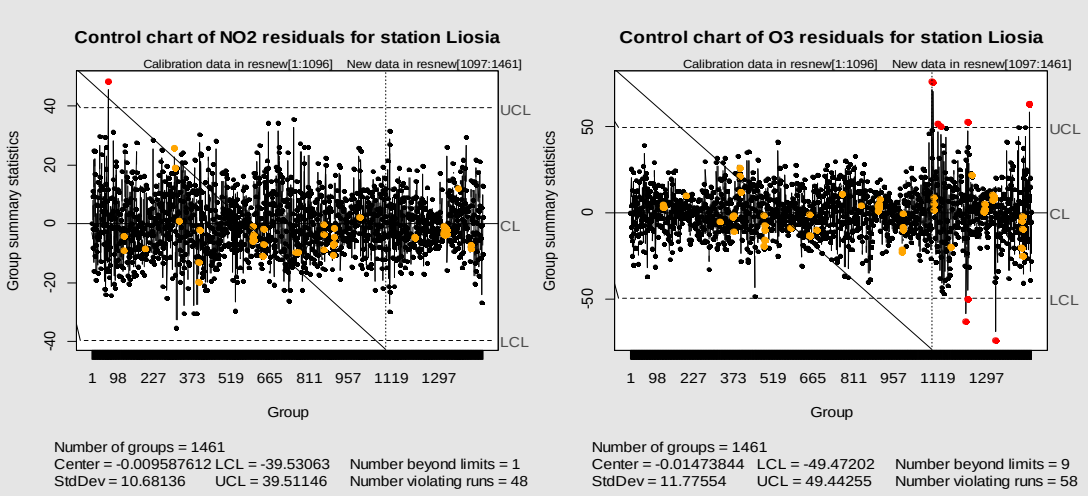


Fig 5.9: Control chart of NO_2 and O_3 forecast errors for station Patisiwn

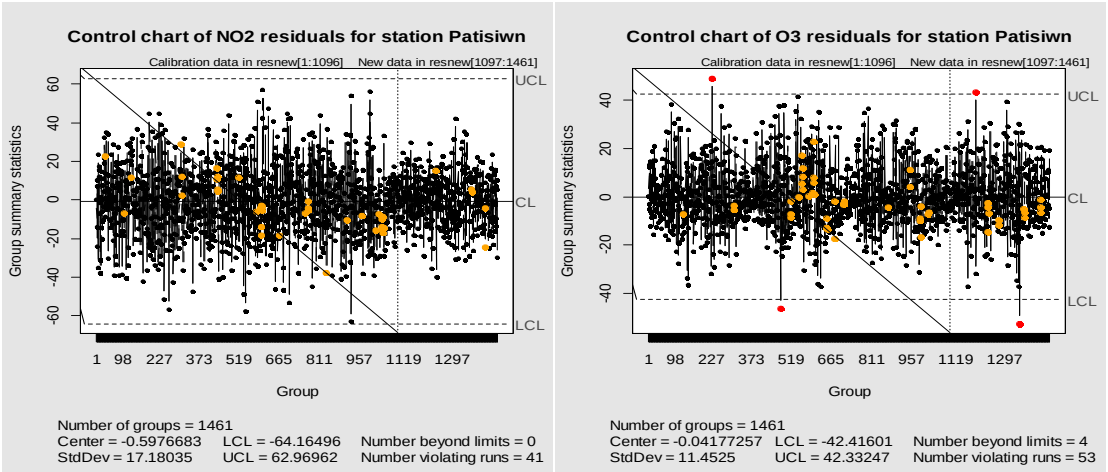
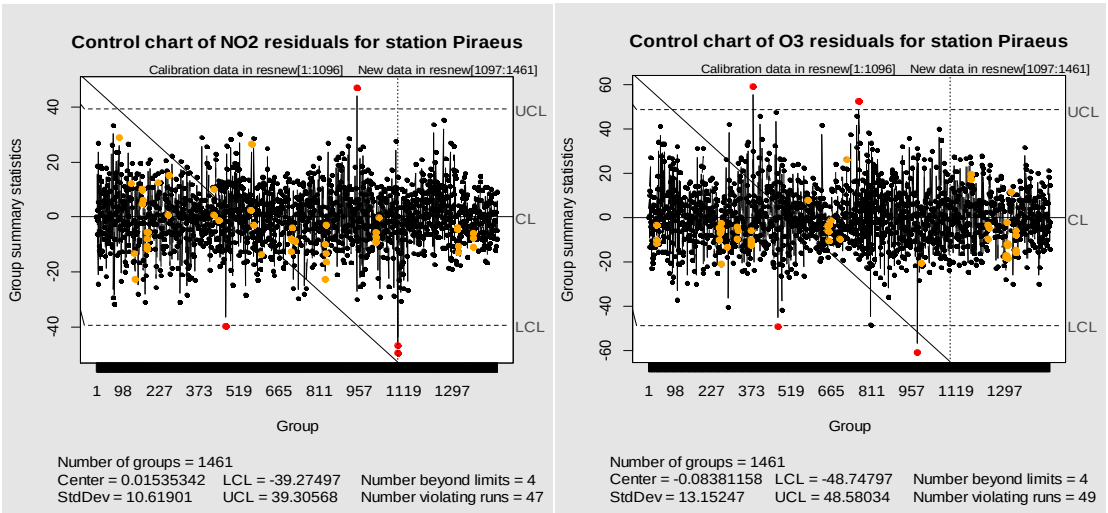


Fig 5.10: Control chart of NO_2 and O_3 forecast errors for station Piraeus



As we can see from the figures above, except for O_3 forecast errors of Liosia station, NO_2 and O_3 forecast errors of Phase II data (2013) are trivial and sometimes no forecast errors are observed. Control charts of NO_2 forecast errors of Patisiwn station and O_3 forecast errors of Liosia station seem to be more robust as the training data used for estimating and forecasting the test data are in-control after Phase I. One might say that the control chart of O_3 forecast errors of Nea Smyrni does not operate well. If we look back to O_3 integrated time series model, we will observe that there is still a remaining seasonal pattern. Hence, a control chart with different limits where seasonal peaks occur may be more appropriate.

Table 5.1 summarizes the outliers (positive forecast errors) occurred after using the NO_2 and O_3 concentrations of 2013 as Phase II data.

Table 5.1: Summary of the positive forecast errors (out of the UCL limit in the $\bar{X}$ chart)

	Positive forecast errors e_t		**Actual (Predicted) values of the time series Y_t**		**Date of the outlier occurred**	
Station	***NO_2***	***O_3***	***NO_2***	***O_3***	***NO_2***	***O_3***
Geoponiki	43.92 42.59	54.46	80.35 (36.43) 59.03 (16.44)	89.83 (35.37)	27 June 2013 1 July 2013	3 June 2013
Peristeri	45.33		74.25 (28.92)		29 April 2013	
Liosia		75.82 74.93 51.37 49.92 52.22 62.84		111.67 (35.85) 143.55 (68.66) 137.53 (86.16) 145.20 (95.28) 125.25 (73.07) 115.84 (53.03)		1 January 2013 7 January 2013 24 January 2013 5 February 2013 10 May 2013 25 December 2013
Patisiwn		43.17		61.4 (18.24)		7 April 2013

A special case of outlier detection is the O_3 concentrations of Liosia station. There were successive large "disturbances" during January 2013 (1st, 7th and 24th of this month), indicating that some external event occurred around these time points. From the respective

control chart, it is clear that the variability of the forecast errors increases somewhere after the end of 2012, thus it is expected a series of points to be out – of – control. The six outliers detected in Table 5.1 may be interpreted as indications of a model change and this is the contribution of the proposed procedure, that it can pinpoint those observations that deserve further investigation.

We also observe that NO_2 exceedances occurred between April and July. Note that nitrogen dioxide plays a major role in the production of ground-level ozone (or smog) which is known to be escalated during this time period with fluctuations in temperature and humidity being the primary cause.

In conclusion, the problem with the training data being out – of – control and the use of its mean and standard deviation as an input to plot Phase II data is that in this case, outliers affect the estimation of the control limits of the chart. Use of 3σ control limits seems to perform well but may not be suitable enough for plotting these forecast errors.

Chapter 6

Conclusions

In this thesis we have presented and studied the ARIMA models and the main univariate $\overline{X}$ control charts. Our scope was to deal with alarming level shifts in NO_2 and O_3 daily concentrations for several stations by plotting the forecast errors of the models fitted. The first problem under consideration was the high missingness pattern of the data in hand. This issue was solved by implementing two multiple imputation methods for time series data in order to have the advantage of comparing them and choose the best one fitted for our datasets. A second problem was autocorrelation of the time series data and which will be the best ARIMA model fitted to eliminate this autocorrelation. ACF plots reassured us of the existed autocorrelation as they didn't cut – off at lag 10. Differenced time series were then determined and it was found that the ACF plot for the first order differenced time series had now a ''cut – off'' pattern. Accuracy measures such as MSE and AIC were used to find the best fitted model. The model with minimum AIC was selected as the model that was closest to the statistical process generating the data. Parameter estimation of the selected models was performed in order to assess the stationarity and invertibility conditions of the time series. Diagnostic checking of the process residuals as the last step of ARIMA models selection was conducted. Ljung – Box test proved that the residuals of the ARIMA models used had no trace of autocorrelation (all the p-values were significantly high).

The selection of forecasting method was another problem examined. One – step ahead forecasts were used for $h = 365$ in total which would be the NO_2 and O_3 predicted concentrations from January 2013 until December 2013. This type of forecasts were calculated a) without refitting the model for our training dataset $(2010 - 2012)$ and b) with re – estimation of the model for $2010 - 2012$ at each iteration randomly. The last method was used as an example of comparison with the basic method of no re – estimation only for

Geoponiki station. ARIMA model performance was assessed by comparing h – step forecasts of NO_2 and O_3 measurements with their respective observed values and this was done by assessing the daily forecast errors $e_t = Y_t - \hat{Y}_t$.

Finally, the one-step-ahead NO_2 and O_3 forecast errors calculated are plotted on a $\bar{X}$ control chart for all stations studied. Control limits to monitor new incoming observations are constructed from the training period of length $n = 1096$ ($n = 365$ served as a test sample). For example, in Geoponiki station, we noticed two large outliers, one near the end of June 2013 and another at the beginning of July 2013. As for the specific days that the alarming values occurred, the NO_2 concentrations may have been mispredicted due to the covariates inserted for imputing the original series (observed values). Therefore, the ARIMA model generated afterwards was already biased by environmental covariates. For example, maximum temperature in these two days was low (29 degrees of Celcium) in relation to other days in June and July respectively. These low values may have resulted to lower predicted concentrations of the pollutant than the real observed values. As far as O_3 concentrations in Liosia station is concerned, we observe from the residual plot that three successive forecast errors occurred at January 2013. Looking back at general weather characteristics during this time period, we notice that in this month, particularly high temperatures for this time of the year were observed. More specifically at January 1^{st} (time of the first outlier) the weather was quite warm in relation to previous days while at January 7^{th} (time of the second outlier) the temperature decreased significantly with heavy snowfalls even in the center of Athens. At January 24^{th}, high precipitation levels were occurred in contrast to the previous phenomena. Such weather variations may be responsible for these distinctive outliers, and once again the significance of the environmental covariates to the predicted concentrations of the pollutants is verified.

There are several open questions, which we consider as areas for future research. If a structural break appears in the time series, the control chart should be readjusted and new control limits need to be computed. Another limitation of the presented approach is the normality assumption on the forecast errors (excluding outliers). After performing Jarque – Bera test for normality of the residuals we noticed that the NO_2 and O_3 distributions of many stations of the case study deviated from the normal distribution. Thus, more robust control charts need to be constructed especially when outliers are appeared in the training sample because it will help decrease the false detection rate which is much higher when the training

data are not cleaned from outliers (overestimating the forecasted model). Finally, we should take into consideration the correlation between the two pollutants (NO_2 and O_3) studied in the thesis (O_3 is essentially formed by NO_2!) and propose a multivariate control chart in order to monitor them jointly.

APPENDIX

Fig I: Residuals versus fitted values of NO_2 pollutant for each studied station

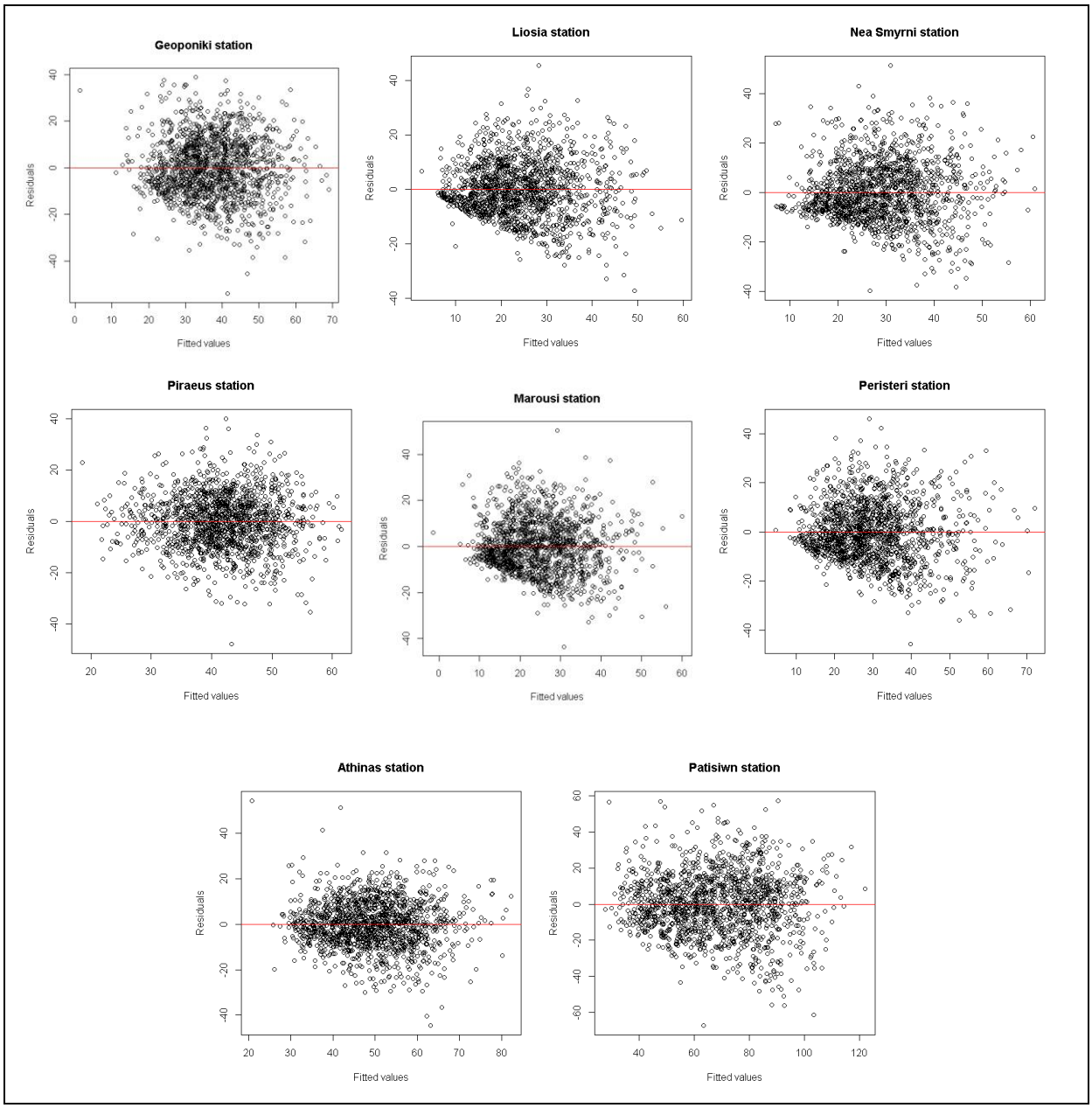
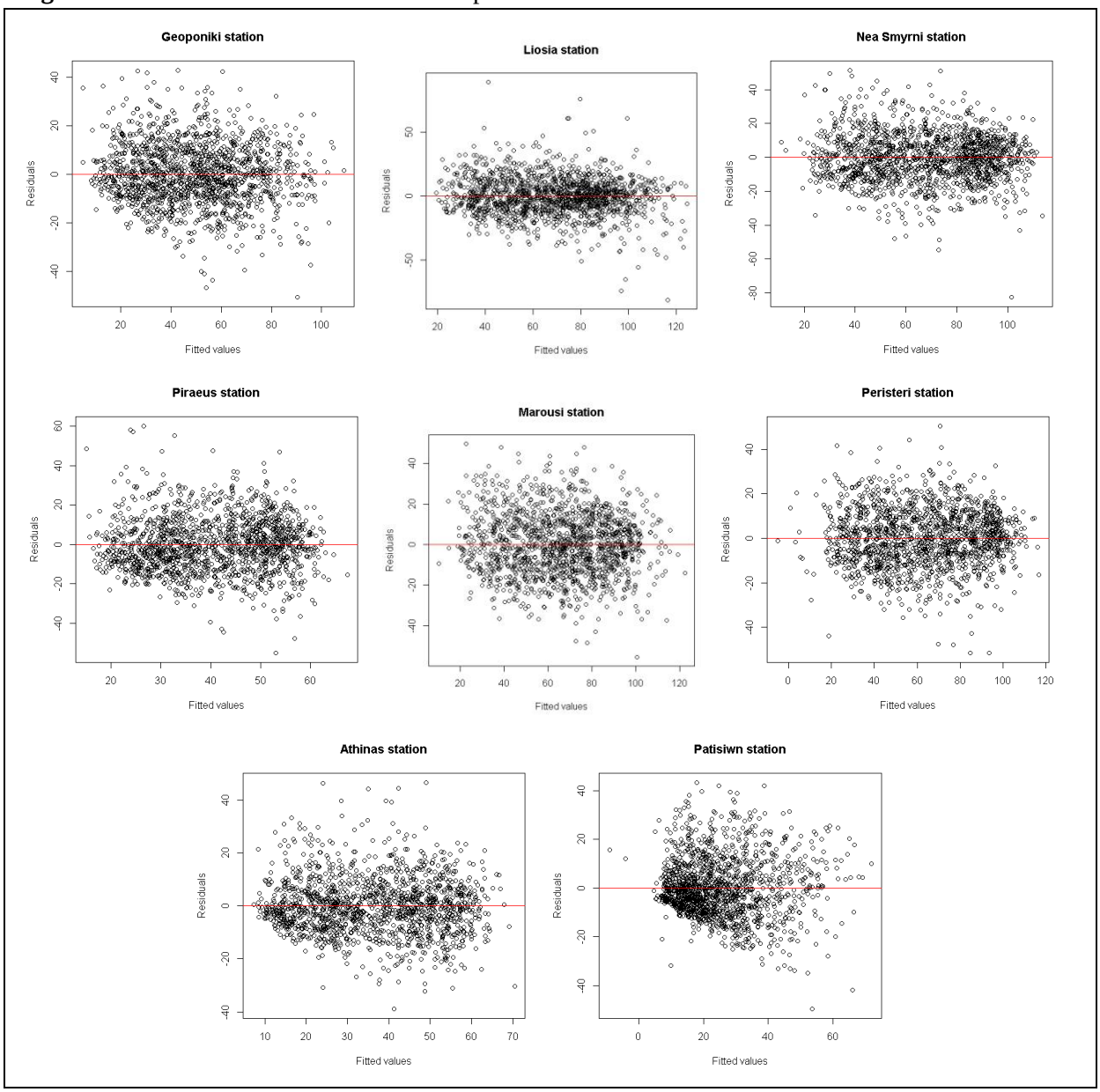


Fig II: Residuals versus fitted values of O3 pollutant for each studied station



References

Books

Box, G. E. P., G. M. Jenkins, 1976. Time series analysis: forecasting and control Revised Edition, Holden – Day, Oakland California

Enrique del Castillo, 2002. Statistical Process Adjustment for Quality Control, John Wiley and Sons, Inc.

R.S. Pindyck, D.L. Rubinfeld, 1998. Econometric models and economic forecasts, 4th edition, McGraw Hill

Montgomery, D. C.; Mastrangelo, C. M. (1991). Some Statistical Process Control Methods for Autocorrelated Data. Journal of Quality Technology, Vol. 23, pp. 179-193.

National Research Council of the National Academies**,** Acute Exposure Guideline Levels for Selected Airborne Chemicals, Volume 19

Roderick J. A. Little, Donald B. Rubin, 2002. Statistical Analysis with Missing Data, Wiley Series in Probability and Statisics, 2nd Edition

Papers

Xinyao Hu, Xingda Qu, 2013. An individual-specific fall detection model based on the statistical process control chart, Elsevier¶

Lu, C.W., Reynolds Jr., M.R., 1999. Control charts for monitoring the mean and variance of autocorrelated processes. Journal of Quality Technology 31, 259–274

Christophe Croux, Sarah Gelper, Koen Mahieu, 2014. Robust control charts for time series data, K.U Leuven, Erasmus University–Rotterdam*

Noor Faizah Fitri Md Yusof, Nor Azam Ramli, Ahmad Shukri Yahaya, IJEP 2011. Extreme Value Distribution for prediction of future PM10 exceedences, Clean Air Research Group - School of Civil Engineering, Universiti Sains Malaysia*

Nuno Prista, Norou Diawara, Maria Jose Costa, Cynthia Jones, 2011. Use of SARIMA models to assess data poor fisheries: a case study with a sciaenid fishery off Portugal, Fish. Bull., 109:170-185

Sunartono, 2011. Time series forecasting by using Seasonal Autoregressive Integrated Moving Average: Subset, Multiplicative or Additive Model, Journal of Mathematics and Statistics 7 (1): 20-27

Edson Zangiacomi Martinez, Elisangela Aparecida Soares da Silva, Amaury Lelis Dal Fabbro, 2011. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of Sao Paolo, Brazil, Sociedade Brasileira de Medicina Tropical 44 (4): 436-440

Ralph Mc Nally, Murray Ellis and Geoff Barrett, 2004. Avian biodiversity monitoring in Australian rangelands, Austral Ecology 29, 93-99

Farshid S. Ahrestani, Mark Hebblewhite and Eric Post, 2013. The importance of observation versus process error in analysis of global ungulate populations, Scientific Reports 3:3125

Mesnil, B., and Petitgas, 2009. Detection of changes in time series of indicators using CUSUM control charts, Aquat. Living Resour. 22: 187 – 192

Petitgas, P., 2009, The CUSUM out-of-control table to monitor changes in fish stock status using many indicators. Aquat. Living Resour. 22: 201 – 206

Scandol, J. 2003, Use of cumulative sum (CUSUM) control charts of landed catch in the management of fisheries. Fish. Res. 64: 19 – 36

Weisent J., Seaver W., Odoi A., 2014. The importance of climatic factors and outliers in predicting regional monthly campylobacteriosis risk in Georgia, USA, 58 (9)

Alwan C. L., Roberts V. H., 1998. Time-Series Modeling for Statistical Process Control, Journal of Business and Economic Statistics, 6(1): 87-95

Tsay S. Ruey, 1988. Outliers, Level Shifts, and Variance Changes in Time Series, Journal of Forecasting, Vol. 7, I-20

Ionel Ioana and Francisc Popescu, Methods for online monitoring of air pollution concentration, Politechnica university of Timisoara, Romania)[*]

Planning and implementing a real – time air pollution monitoring and outreach program for your community – The AirBeat project of Roxbury Massachusetts, EPA, 2010

Palmgren, F.; Berkowicz, R.; Egeløv, A.; Hertel, O.; Kemp, K.; Larsen, Søren Ejling, 1999. Experimental studies of air pollution in street canyons, Proceedings, Transport and chemical transformation in the troposphere. Vol. 2. ed. / P.M. Borrell; P. Borrell. Southampton: WIT Press, p. 811-815.

Rocke, David M., 1989. Robust Control Charts, ASQC and the American Statistical Association, University of California, pp. 173-184

Tatum, Lawrence G., 1997. Robust Estimation of the Process Standard Deviation for Control Charts, American Statistical Association and ASQC, The City University of New York, pp. 127-141

Rigdon, Edward E. (1994), Demonstrating the Effects of Unmodeled Random Measurement Error, Structural Equation Modeling, 1 (4), 375-80.

Sullivan, Joe H.; Woodall, William H., 1996. A Comparison of Multivariate Control Charts for Individual Observations, Mississippi State University, University of Alabama, Tuscaloosa, pp. 398-408

[¶]Scientific web-library including well-known journals

[*]Papers released as a part of Phd research