UNIVERSIDADE D
COIMBRA

Maria José Guerra

FORECASTING OF AIRBORNE POLLUTANTS WITH ARIMA MODELS

Dissertação no âmbito do Mestrado em Química Avançada e Industrial,
variante Química-Física Experimental e Teórica,
orientada pelo Professor Doutor Jorge Costa Pereira
e apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologia
da Universidade de Coimbra

Setembro de 2018

UNIVERSITY OF COIMBRA

MASTERS THESIS

Forecasting of Airborne pollutants with ARIMA models

Author:
Maria José GUERRA

Supervisor:
Dr. Jorge COSTA PEREIRA
Dr. Pedro CARIDADE

A thesis submitted in fulfillment of the requirements
for the degree of Master's in Advanced and Industrial Chemistry Chemistry

Department of Chemistry
Faculty of Science and Technology of the University of Coimbra

September 5, 2018

"How are you holding up? Because I'm a POTATO!"

GLaDOS - Portal 2

Abstract

Exposure to airborne pollutants has serious health, environmental and economic impacts. Real-time monitoring and forecasting of the air quality has become of necessity in order to protect citizens. Due to the intricacies of the physical and chemical processes that govern airborne pollutant formation and transport, producing forecasts is not a trivial task. Computational and expertise costs are very high when resorting to physically based models. Stochastic models provide more parsimonious approaches with the same accuracy potential.

The Box-Jenkins methodology, that originally developed for econometrics, was used in the past to forecast successfully some airborne pollutants. Using that methodology several Autoregressive Integrated Moving Average model was fitted to six pollutants, regulated in the European Union, and the goodness of fit tested with residual analysis and a statistical test.

The fitted models were used to produce out-of-sample forecasts and their accuracy tested. In order to evaluate the models forecast robustness, a time series cross-validation procedure was implemented . The results presented and compared for every pollutant.

Resumo

Exposição a poluente aéreos acarreta grandes impactos nos sectores da saúde, ambiente e economia. Monitorização da qualidade do ar em tempo real e emissão de previsões tornaram-se numa necessidade para a protecção dos cidadãos. Devido à complexidade dos processos físicos e químicos que descrevem a formação e transporte de poluentes, fazer previsões dos mesmos não é uma tarefa trivial. Os custos computacionais e de perícia são extremamente alto quando são utilizados modelos baseados fisicamente. Modelos estocásticos providenciam uma abordagem mais económica e têm o potencial de exibir a mesma precisão e exactidão.

O método de Box-Jenkins, desenvolvido originalmente para econometria, foi utilizado no passado para a previsão bem sucedida de algumas espécies de poluentes aéreos. Fazendo recurso a essa mesma metodologia vários modelos Auto-regressivos Integrados com Média Móvel foram ajustados a seis poluentes, regulados dentro da União Europeia, e a qualidade dos ajustes testadas através de análise residual e um teste estatístico.

Os modelos ajustados foram utilizados para produzir previsões para além da amostra e a sua qualidade testada. Para melhor avaliar a robustez das previsões, vários procedimentos de validação cruzada para séries temporais foram implementados. Foram apresentados e comparados os resultados obtidos para todos os poluentes.

Agradecimentos

Gostaria em primeiro lugar de agradecer aos meus orientadores, Professor Doutor Jorge Costa Pereira e o Doutor Pedro Caridade pelas discussões, sugestões e apoio dado ao longo deste projecto. Nada disto seria possível sem vocês. Também a toda à equipa da SpaceLayer em especial à Dra Carla Gouveia-Caridade. Sem o seu apoio este desafio nunca me teria sido proporcionado.

Um gigante obrigado aos meus pais, que patrocinaram não só a minha educação mas a minha vida. Apoiaram-me em tudo. Mesmo em coisas que não concordaram e possibilitaram-me ser uma pessoa mais educada, completa e humana. Devo-vos tudo. Mãe, tinhas razão. Ainda não sei em quê. Mas provávelmente tinhas razão.

Para as minhas irmãs, que foram os meus exemplos e que me mostraram tantas possibilidades na vida. Não estamos presos a nada a não ser a nós próprios. Obrigada também por partilharem e nutrirem toda a minha loucura. É de família. Espero poder cuidar de vocês como cuidaram de mim.

Aos meus amigos, André, Eduardo e Miguel. Obrigada pelos incontáveis cafés, conversas, discussões, risos e momentos. Vocês também passaram por aqui comigo e ajudaram-me a suportar a carga. André, depois não te esqueças de nós quando fores famoso. Miguel, volta estás perdoado. Eduardo, espero sobrevivas aos próximos 9 meses.

Finalmente, para o Renan. Não há palavras para te agradecer por tudo$^{(\text{mesmo os ralhetes})}$. Obrigada por acreditares mais que eu. Por não me deixares desistir mesmo quando eu tudo o que queria fazer era ser um burrito. Afinal acho que valeu a pena.

Para a Ângela.

Contents

List of Figures

List of Tables

List of Abbreviations

ACF	**A**uto**c**orrelation **F**unction
ADF	**A**ugmented **D**ickey-**F**uller
AIC	**A**kaike **I**nformation **C**riterion
ANN	**A**rtifical **N**eural **N**etwork
AR	**A**uto**r**egressive
ARIMA	**A**uto**r**egressive **I**ntegrated **M**oving **A**verage
LOESS	**Lo**cal **r**egression
MA	**M**oving **A**verage
MAD	**M**edian **A**bsolute **D**eviation
MLE	**M**aximum **L**ikelihood **E**stimation
PACF	**P**artial **A**uto**c**orrelation **F**unction
QQ	**Q**antile-**Q**uantile
TS	**T**ime **S**eries

Chapter 1

Introduction

1.1 Airborne pollution impacts and future solutions

Pollutant gases and particulate matter have always been present naturally in the atmosphere. But ever since human activity boomed with the Industrial Revolution, creating industry that depended on the energy from burning coal, they have become a serious health problem [1]. During the 1950's, a smog crisis in London killed thousands of people and caused health problems to thousands more [2] leading legislators to understand that emissions of said gases needed to be regulated. After successfully implementing said regulations it was presumed until the 1980's that the levels of pollution experienced in Europe were no harmful. Since then long-term epidemiological studies reveled that even exposure to moderate to low concentrations of air pollution have short and long-term effects on human health [1].

The 2017 report by the European Environment Agency – *Air quality in Europe* – states that air pollution still seriously impacts the European population health and economy. It is responsible for losses in life span, increases in medical costs and consequently hindering productivity and finally, premature death [3].

In Portugal alone, the cost of treating childhood asthma amounts to 0.9% of the total health-care expenses with 75% of these represented by the treatment of acute episodes [4]. Air pollution is a serious contributor to the heightening and triggering of respiratory conditions [5], meaning that a possible solution would be interconnected to a change in policies in order to diminish the number of acute cases that require emergency treatment [3–5].

To counteract the impacts of air pollution, the European report explains that solutions that involve technological advances and structural and behavioural changes must be found and implemented across global and local levels, economic sectors and enlist the public in what is only achievable by a joined effort [3]. It is in response to the European necessity to take immediate action, that Sensor Observation of Urban Life, SOUL, by SpaceLayer Technologies came to life.

Relying in satellite data provided by the Copernicus Earth Observation Programme and a network of moving sensors through out a city, capture local concentrations of pollutants. By sending real-time geo-tagged information about the air quality to a mobile application, it enables the user to protect himself against acute exposition to harmful pollutants, either by taking alternative routes or preventive medication.

The present work is a first approach to produce n-step ahead forecasts of several common airborne pollutants, regulated in the European Union, in order to ultimately generate alerts.

We will not dwell in the spatial aspect of the problem mentioned earlier hence, only satellite data will be used. The data for the present work was provided by the Copernicus program and contains hourly concentrations about six pollutants:

Carbon Monoxide - CO,

Sulfur Dioxide - SO_2,

Nitrogen Dioxide - NO_2,

Ozone - O_3,

Particulate matter - PM_{10} and $PM_{2.5}$.

The main anthropogenic sources for these pollutants is fuel combustion. Tropospheric O_3 is a secondary pollutant, as it is not directly emitted into the atmosphere. Rather it results from a photochemical reaction between precursors gases such as, NO_2, NO and non-methane volatile organic compounds. The precursors main source is nevertheless, fuel combustion [3].

1.2 Air Quality forecasting

Considering the impact that air pollution has in health, the environment and even economy, air quality forecasting has become extremely relevant. Many advances have been made since the 1960's but this problem still poses a challenge due to the complexity of the phenomena involved [6]. There are several, more or less sophisticated, tools available to produce air quality forecast. These can be broadly grouped as: simple empirical techniques, statistical models and physically-based approaches.

Simple empirical techniques include methods like, persistence and climatology. The persistence method assumes that the pollution levels of one day are the same that the day before and climatology uses averages of historical data to produce a forecast. As one can extrapolate, these methods are not highly accurate and fail if there are changes in the air quality.

Statistical models use the fact that the different variables are correlated to produce forecasts. In this group are included methods like regression models, artificial neural networks and fuzzy logic. These approaches demand a large volumes of different types of data. Both pollutants concentrations and other variables that correlate to these e.g. UV index or temperature [6].

Although more accurate than simple empirical models and not very computationally expensive, there are a few disadvantages. They can only describe the conditions in which the data used was collected and cannot be generalized. For that reason they also fail to describe any behavior not present in the historical data. Fundamentally, simplifications of several important meteorological and chemical processes are used, hindering the accuracy of the models. They also lack the capacity to explain the underlining processes since they do not use any physical or chemical relations. Despite the handicaps mentioned, some artificial neural networks were able to outperform a physically-based model in [7, 8].

Finally, the most complex and computationally expensive approach, but also the most accurate, are physically-based models. The processes that describe pollutants formation and

accumulation are explicitly resolved. They are capable of forecasting in time and space and provide an understanding of the pollutants processes. Because they do not depend on the characteristics of historic data, they perform well under unusual condition and in unknown or unmonitored zones.

The major difficulty of implementing these types of models are the costs associated. In order to build such a model, it is necessary an extensive knowledge of the sources of pollution and of the processes that determine how the pollutants form and travel. Being an imperfect representation of reality, some approximations have to be made. Knowledge of the processes might be insufficient or the complexity to high to handle. Also, their accuracy depends on the accuracy of the model inputs. If the monitoring of the input data or approximations are not done correctly, it leads to bias and inaccuracy in the forecasts [6, 9].

Particularly in Portugal, the Portuguese Environment Agency, provides a one day ahead forecast of the air quality index. The information is provided to this agency by two Universities, of Aveiro and NOVA in Lisbon [10]. One of the models used to forecast is a physical-based model specifically, CHIMERE. A chemical transport model extensively studied and used in Europe [11–13]. The other model is a statistical model. It makes multivariate regression using classification and regression trees [14, 15] which is a machine learning method. Both models forecast O_3 and PM_{10} concentrations which are subsequently used to calculate the air quality index forecast.

1.3 Objectives for the present work

The data available for each pollutant, is an hourly measurement of the maximum concentration. Since there are no other variables a statistical approach was chosen to deal with the forecast problem. Because the data is recorded as a function of time it is possible to resort to time series modeling.

A Time Series (TS) is a sequence of data where the variable is indexed according to time order. Most commonly, a TS consists of a set observations collected sequentially in time [16]. When working with other types of data, usually, the order is irrelevant. As a matter of fact, it is often good practice to randomize the order of the data points e.g. when training a Artificial Neural Network (ANN), in order to eliminate any dependency effects. This is not the case when analyzing a TS. The data is dictated by time and randomizing a TS would mean to lose information about it. Moreover, this dependency can be crucial to better estimate the appropriate model [17].

For the present work, the purpose of the model will be to produce accurate forecasts of pollutants one day ahead. There are several options to model a TS. For this thesis, Autoregressive Integrated Moving Average (ARIMA) models were chosen. These are essentially regressions to past values of the variable itself, making them suitable for the present work.

The indicated family of models was originally introduced in the field of econometrics but have since been used in several other fields of study. Applying these models to forecast pollutants has been done before with moderate success [7, 18–20].

The models were first introduced by Box and Jenkins who developed the methodology that is still used to build ARIMA models.

The Box-Jenkins method is comprised of three main steps:

- Model selection;
- Parameter estimation;
- Model checking.

The model selected for this work is ARIMA. This is a more general form that can incorporate mixed or purely autoregressive moving average models as will be explained in Chapter 3.

Before taking the second step mentioned above, it is necessary to examine the data visually. We will inspect the data in search of trend and seasonal components, any abrupt changes in behavior and the presence of outliers. These family of models is constrained to stationary data [17]. If trend and seasonal components are present, it is necessary to remove them in order to coerce the TS into a stationary series. The values produced from the transformations are referred to as residuals. There are several transformations that can help to achieve this goal and they will be discussed further ahead.

After ensuring stationarity of the data, parameter estimation can effectively begin. Analyzing the autocorrelation and partial autocorrelation plots of the residuals, informs what parameters could possibly perform a good fit. Parameter estimation will be done by a search algorithm to provide the best fit. The algorithm tests the fit of several models and selects the best one. We then assess the model analyzing errors and perform residual analysis of the model residuals[1]. If the results are not satisfactory it is necessary to return to the first step. This cycle is repeated until an adequate model is achieved.

Upon completing the former steps we proceed in using the selected model to forecast the data. Finally, using cross-validation, we will determine the model that produces the most accurate and robust forecasts [17, 21–23].

[1]Not to be confused with the residuals that originate from applying a mathematical transformation to the original data.

Chapter 2

Statistical Concepts

Prior to discussing the practical model implementation, it is necessary to define some concepts that are necessary to understand decisions and interpretations made in Sections and Chapters ahead. The way to make informed decisions about how to use these concepts, is to understand the theoretic fundamentals behind them. We would like to highlight that the information in this Chapter is merely introductory.

2.1 Stochastic Processes and Stationarity

A stochastic process is, by definition, a group of random variables $\{Y_\theta\}$, where θ belongs to an index set Θ. Considering the data for the present work was recorded in fixed intervals, the TS is discrete. Hence, Θ becomes a set of integers representing particular time points. The index θ is now replace by $n \in \{1, \ldots, N\}$ and the stochastic process becomes $\{Y_n\}$ [24, 25]. Since a stochastic process is random, it can also be described as a process that develops according to probabilistic rules.

A stochastic or a probability model, is one that can be used to calculate the probability of a future value, falling between two specific limits [17]. From this point forward it will be implied that the stochastic processes and models are discrete.

These processes are said to be strictly, or strongly, stationary when the joint probability distribution is invariant under a time shift, k [17, 25].

Let $\{Y_n\}$ be a stochastic process and $\mathcal{F}_Y(Y_{n+k}, \ldots, Y_{N+k})$, the function of the joint distribution of $\{Y_n\}$ at times $n+k, \ldots, N+k$. A process $\{Y_n\}$ is strictly or strongly stationary if for all n, for all k and for all $Y_n, \ldots, Y_N$ the following conditions verifies:

$$\mathcal{F}_Y(Y_{n+k}, \ldots, Y_{N+k}) = \mathcal{F}_Y(Y_n, \ldots, Y_N). \tag{2.1}$$

This means that the joint distribution of any set of observations is not a function of time [16, 17]. Strictly stationary natural phenomena are very uncommon. Furthermore, the latter condition is very hard to verify. In practice, less restrictive criteria is imposed and the concept of strong stationarity is replaced with weak stationarity [16, 26].

A process is said to be weakly stationary if condition (2.1) is verified for the first and second joint moments for any time point and shift, n and k, respectively. As a consequence of the latter condition the variance and mean of the process are constant over time.

The covariance between two random variables y_n and x_n is:

$$\mathrm{Cov}(y_n, x_n) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_y)(x_n - \mu_x)\ , \quad \mu_y = \frac{1}{N}\sum_{n=1}^{N} y_n. \tag{2.2}$$

For a weakly stationary process the mean is constant over time, $\mu_n = \mu_{n+k} = \mu$. Then, using Equation (2.2), one can define the autocovariance which will only depend on the interval between time points [16, 26] i.e., the lag k:

$$\gamma_k = \frac{1}{N}\sum_{n=1}^{N-k}(y_n - \mu)(y_{n+k} - \mu). \tag{2.3}$$

2.1.1 Assessing Stationarity

When inspecting a TS one looks for homogeneous aspects: the series is approximately horizontal, has constant variance and there are no predictable long-term pattern. It is not always obvious if a data set is stationary or not. Hence, some tools are required to make an accurate assessment. Next, we discuss some of these tools.

Autocorrelation

Autocorrelation is the dependence of a variable with itself, shifted in time. A plot of the autocorrelation function versus lag is used in TS analysis to infer about the characteristics of a TS.

Using the definition for autocovariance given by Equation (2.3) the autocorrelation, ρ_k, can be calculated. It is defined as:

$$\rho_k = \frac{\gamma_k}{\gamma_0}. \tag{2.4}$$

Autocorrelation of a variable plotted as a function of lags is referred to as the Autocorrelation Function (ACF). Another useful function is the Partial Autocorrelation Function (PACF). The partial autocorrelation between the lagged values, n and $n + k$, of a TS is similar to the autocorrelation but the effects of intermediate lags, $n + 1, \ldots, n + k - 1$, are not accounted for. If the data is stationary the value of the lags will fall to zero quickly and if the data is non-stationary the values will decay slowly [17, 21, 27].

ACF and PACF plots of the data also help to identify parameters of the model. It is not always straightforward to identify what parameters would best explain a TS, but combining the analysis of both plots offers a reasonable starting point. These types of plots will later be used in Chapter 4 to help identify non-stationarity and, to assess what model parameters could one expect for a given pollutant.

Augmented Dickey-Fuller Test

Another tool for testing stationarity is unit-root tests like Augmented Dickey-Fuller (ADF). This test uses an autoregressive model applied to the data. If there is an unit-root present in the

model the characteristics of the TS will be a function of time and the stationarity condition is not met [28]. Let:

H_0: The model has an unit-root.

H_1: The model does not have an unit-root and hence, the TS is stationary.

Regarding the present work, the sample size is superior to 500 data points. This means, that to reject the null hypothesis, at the 99% confidence level, the ADF results for the data will need to be inferior to the critical value of -3.98 and a p-value inferior to 0.05 [28].

2.1.2 Extracting Non-Stationarity

Many TS models are built assuming that the condition of stationarity is met, but real-world data is more often that not non-stationary [16]. One way of overcome this is to decompose the data in order to remove the components that make it non-stationary.

Assuming additive components we can define a TS, y_n, as:

$$y_n = S_n + T_n + R_n. \tag{2.5}$$

In the latter definition, S_n,T_n and R_n are the seasonal, trend-cycle and remainder components, respectively.

Seasonality,S_n, is defined by the impact of seasonal factors like, months, holidays or days of the week, on the data. It has fixed and know frequency. A cycle,T_n, is similar to a seasonal pattern but with no fixed frequency. Usually a cyclic pattern is also longer than a seasonal one. The trend is observed as long term variations in the TS as a whole [27].

It is possible to model these components separately in order to subtract them to the TS and obtain the remainder, which includes all the unexplained features of the TS [16]. Next we will present some methods to model these components.

Moving Average Smoothing

One way of estimating the trend, T_n, in order to smooth the data is to apply a moving average (MA) [16]. A moving average order m is defined as:

$$\hat{T}_n = \frac{1}{m} \sum_{j=-k}^{k} y_{n+j}. \tag{2.6}$$

Here, $k = \frac{m-1}{2}$. For even orders, the MA is not symmetric and it becomes more practical to use a centered MA. This is equivalent to using a MA order of 2 after the first MA of even order [29].

Seasonal-Trend Decomposition based on LOESS

Another method to estimate the TS components is local regression (LOESS). It works by constructing a function that results from local fits to subsets of the data with a chosen length. The local fits are performed using polynomials of first or second degree, depending on the

curvature of the data. The fitting of the local polynomial is also weighted. This means that closer observations will have a greater influence on the local fit and, as the distance increases, the observations will have less influence on the local fit. All the procedure is implemented in R and described in [30].

Differencing

Differencing is the process of subtracting consecutive data points generating a new TS with the values of the differences. First order differencing is defined by:

$$y'_n = y_n - y_{n-1}. \tag{2.7}$$

This new time series,y'_n, will have $N-1$ values since it is not possible to obtain a difference for y_1, the first observation. To obtain higher order differencing it is only necessary to repeat this process e.g., to obtain second order differencing the differences of y'_n are calculated and so on.

Differencing the data helps eliminate changes in level and hence, helps stabilize the mean and decrease the effects of a trend and seasonality [17, 27].

Square root and Natural Logarithm

These transformations are special cases of a Box-Cox transformation. They are designed to stabilize the variance of the data, additivity of effects and symmetry of the density [27, 31].

2.2 Models and Fitting

The previous section was dedicated to exposing statistical concepts that are needed in order to apply the family of models chosen. Now, the models themselves and how to evaluate their fit will be described.

2.2.1 Autoregressive Models

Autoregressive (AR) models are those where a linear combination of past values of given variables are used to forecast that same variables [27]. The evolution of these processes happens by regressing the variables past values towards the mean and then adding noise [26, 27].

An AR process of order p, AR(p), can be defined as follows:

$$y_n = c + \phi_1 y_{n-1} + \phi_2 y_{n-2} + \ldots + \phi_p y_{n-p} + \varepsilon_n. \tag{2.8}$$

Where $\phi_1, \ldots, \phi_p$ are coefficients and ε_n is a white noise term. The models adjust the coefficients by regressing to past values. Changing these coefficients produces different TS patterns [26, 27].

These types of processes are discrete representations of ordinary differential equations making them very useful in climate research [26]. Additionally, they are very flexible and can handle various types of time-series behaviors.

2.2.2 Moving Average Models

Moving average (MA) models, not to be confused with moving average smoothing explained previously, are a type of regression model where the past forecast errors are used to estimate the coefficients [27]. A MA model order q, MA(q), is given by the following equation:

$$y_n = c + \theta_1 \varepsilon_{n-1} + \theta_2 \varepsilon_{n-2} + \ldots + \theta_q \varepsilon_{n-q} + \varepsilon_n. \tag{2.9}$$

Here, $\theta_1, \ldots, \theta_q$ are coefficients and $\{\varepsilon_n\}$ is a white noise process with mean zero and variance σ^2. Similar to AR models, different TS patterns are obtained by altering the coefficients [27].

MA and AR models are not unique. In fact, it is their non-uniqueness that allows for the construction of more sophisticated models described later. The non-uniqueness makes it possible to describe a stationary AR process, with arbitrary precision, using an infinite MA [26]. This can be more easily illustrated for an AR(1) model[1]:

$$\begin{aligned} y_n &= \phi_1 y_{n-1} + \varepsilon_n \\ &= \phi_1(\phi_1 y_{n-2} + \varepsilon_{n-1}) + \varepsilon_n \\ &= \phi_1^2 y_{n-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_n \\ &= \phi_1^3 y_{n-3} + \phi_1^2 y_{n-2} + \phi_1 \varepsilon_{n-1} + \varepsilon_n + \ldots \end{aligned}$$

If the condition $|\phi_1| < 1$ is met, the coefficient value , ϕ_1^k, decreases as the index k increases. If $k \to \infty$ it is possible to write:

$$y_n = \varepsilon_n + \phi_1 \varepsilon_{n-1} + \phi_1^2 \varepsilon_{n-2} + \phi_1^3 \varepsilon_{n-3} + \ldots. \tag{2.10}$$

which is a MA(∞) process. The previous condition is called the invertibility condition. The same is valid for a MA(q) process as long as the invertibility conditions are met [26, 27]. These conditions are similar to the ones of the AR model e.g., for a MA(1) model the constraint would be $|\theta_1| < 1$. For higher model orders the conditions are more complex.

The R algorithm, used in the present work, was designed to account for these constraints of stationarity and invertibility [27].

Invertibility has use beyond the non-uniqueness of AR and MA models. Take for example a MA(1) process, $y_n = \varepsilon_n + \theta_1 \varepsilon_n - 1$. Using the AR($\infty$) representation, the most recent error, ε_n, can be approximated by a sum of current and past observations:

$$\varepsilon_n = \sum_{j=0}^{\infty} (-\theta)^j y_{n-j}. \tag{2.11}$$

If $|\theta| > 1$, the weight increases with the lag and more distant observations will have more influence than the adjacent ones. If $|\theta| = 1$, the weight is constant and the distance of an observation will not matter since they are weighted the same. The two previous cases are very illogical hence, it is required that $|\theta| < 1$. It is only when this condition verifies that the most recent observations weight more in the error than the more distant ones [27].

[1] For simplicity we assume that the AR(1) process has no constant.

2.2.3 Autoregressive Integrated Moving Average Models

The previous models discussed are special cases of the more general Autoregressive integrated moving average models (ARIMA). Here integration refers to reversing the differencing transformation. This type of model is defined by orders p, d and q. The parameter p refers to the order of AR model used, the d to the differencing order and the q to the order of the MA model. Using the backshift notation, where $By_n = y_{n-1}$, the general form of a nonseasonal ARIMA can be written as:

$$(1 - \phi_1 B - \ldots - \phi_p B^p)(1 - B)^d y_n = c + (1 + \theta_1 B + \ldots + \theta_q B^q)\varepsilon_n. \tag{2.12}$$

Where $\{\varepsilon_n\}$ is a white noise process with mean zero and variance σ^2. If there is a significant influence from seasonal effects, these can be included in the model. A seasonal ARIMA model is constructed simply by including seasonal $(P, D, Q)_m$ parameters, where m is the seasonal period. , a seasonal ARIMA model can by described as[2]:

$$(1 - \phi_p B^p)(1 - \Phi_P B^{m \times P})(1 - B)^d (1 - B^m)^D y_n = (1 + \theta_q B^q)(1 + \Theta_Q B^{m \times Q})\varepsilon_n. \tag{2.13}$$

The seasonal terms of the model are similar to the nonseasonal but use the backshift of the seasonal period. These terms are simply multiplied to the existent ones. The invertibility conditions that apply to AR and MA models also apply to ARIMA models [27].

2.2.4 Parameter estimation and Goodness of fit

Here, the tools used to estimate the model parameters and, later asses the goodness of fit will be explained.

Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a method for optimizing the parameters of a statistical model. The parameters are fitted in a way that maximizes the likelihood function. This function describes the likelihood that a given set of model parameters originated the set of observations. The MLE is the method used in R for estimating ARIMA parameters [27].

Akaike Information Criterion

The Akaike information criterion (AIC) is a relative estimator of how much information is lost by a model. When searching for ARIMA parameters, the goal is to find the model with the lower AIC value, i.e, the model that lost the least amount of information [27]. This estimator also has the advantage of trading-off between the goodness of fit and the loss of degrees of freedom, which helps to avoid overfitting [32]. The AIC is defined by:

$$\text{AIC} = -2 \ln \mathcal{L}(\hat{\delta}) + 2v, \tag{2.14}$$

[2] Again, for simplicity it is assumed that $c = 0$.

where $\mathcal{L}(\hat{\delta})$ is the likelihood of the estimated model and v the total number of parameters estimated in the model [32].

Often, if the sample size is small, there is the risk that the AIC will choose more complex models and overfit the data. To counteract this, a correction was developed and AIC becomes AIC_c, defined by:

$$AIC_c = AIC + \frac{2v^2 + 2v}{n - v - 1}, \tag{2.15}$$

where n is the sample size. This correction applies and extra penalty for smaller values of n and as $n \to \infty$ AIC_c converges to AIC [33].

Ljung-Box test

ACF and PACF plots are an important tool to evaluate the goodness of fit. If a model is able to explain the data correctly the residuals will be uncorrelated and resemble white noise. This means that the ACF and PACF plots will not have significant peaks and, ideally, that residuals pass a statistical test to evaluate the hypothesis of correlation [21, 27].

The test chosen for the present work is the Ljung-Box, which is a *portmanteau* test for serial correlation. This means it tests the autocorrelations as a group instead of individually [27].

For the Ljung-Box test:

H_0: The model residuals are uncorrelated and the model does not exhibit lack of fit.

H_1: The model residuals are correlated and the model exhibits lack of fit.

This test is based on the following statistic:

$$Q* = n(n+2) \sum_{k=1}^{m} \frac{r_k^2}{n-k}, \tag{2.16}$$

where n is the number of observations, m is the maximum lag being considered and r_k is the autocorrelation at lag k. A large value of $Q*$ indicates that there is autocorrelation present. If the Q values are not significant, meaning the p-values are larger than 0.05, considering a significance level of $\alpha = 0.05$, then the null hypothesis can not be rejected. In this case, the model residuals are considered to be white-noise [29].

Root Mean Squared Error

The Root Mean Squared Error (RMSE) is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} e_i^2}. \tag{2.17}$$

Where N is the number of errors and e_i is the i^{th} error. This error is not suitable for comparisons between different series since it is scale dependent. It is useful when applying different methods to the same dataset [34, 35].

Mean Absolute Error

The Mean Absolute Error is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|e_i|. \tag{2.18}$$

It is more robust with respect to outliers than RMSE but still not suitable for comparison between different datasets since it is scale dependent.

Mean Absolute Scaled Error

The simple errors are scaled using the in-sample MAE from a naïve forecast method. This forecast method simply assumes that the future value is equal to the past value, $y_n = y_{n-1}$. The scaled error is define as:

$$\text{SE} = \frac{e_h}{\frac{1}{N-1}\sum_{i=2}^{N}|y_i - y_{i-1}|}. \tag{2.19}$$

Here the subscript h refers to the forecast set and N to in-sample data. The MASE is simply:

$$\text{MASE} = \frac{1}{h}\sum_{i=1}^{h}|e_i|. \tag{2.20}$$

This error is scale independent so it is suitable for comparing different methods. If MASE< 0 the forecast is better than the naïve method and a MASE> 1 implies the opposite [34].

Chapter 3

Model implementation

3.1 Quantile-Quantile Plots

The estimator used by R, MLE, assumes that the data came from a normally distributed population. For this reason we need a tool to perform a normality check before fitting any model.

A Quantile-Quatile (QQ) plot is a plot that compares the shape of distributions by plotting the quantiles of each one against the other. In the case of a normality check, the plot is of the estimated probability density quantiles of the sample data against a theoretical normal distribution. If the two distributions are similar, the plot will fall on a line that connects the first and third quantiles of the simulated normal distribution. Some outliers in the extremities are still acceptable [36].

3.2 ARIMA parameter estimation

The numerical implementation of these models is already present in R and the specific package uses the procedure described in [37]. Since the main difficulty in modeling ARIMA is the choice of the $(p, d, q)(P, D, Q)$ parameters, in the course of this work, an automatic search function will be used. If seasonality is allowed, the function tests the TS with a measure of seasonality according to [38]. If significant seasonality is detected the D parameter is the first to be estimated resorting to statistical testing [37]. The only orders of D allowed are 0 or 1. Then, the second parameter estimated is the nonseasonal differentiation order, $0 \leq d \leq 2$. To achieve that a unit root test is used. If the test determines a unit root is present the first differences of the data is tested and so on.

After estimating the order of D and d the orders of p, q, P and Q are determined using AIC_c. The AIC_c is not comparable between models with different orders of differentiation. This results from the likelihood function being calculated over the differentiated data [37].

By default the algorithm would fit four initial models with combinations of $0 \leq p \leq 2$, $0 \leq q \leq 2$, $P = 0, 1$ and $Q = 0, 1$ to determine the best model out of these. Next, small variations on the best model out of the previous four are considered. This stepwise process is repeated until it becomes impossible to find a model with lower AIC_c. Additionally, when estimating the parameters some approximations are also used [37].

This procedure implies that the algorithm does not consider all the possible combinations in order to resume the search. This is a very useful property if there is a large number of TS

to analyze. Unfortunately it also makes it conceivable that the model that minimizes the AIC_c will not be considered. For the purpose of this work, the approximations and the stepwise features will be disabled. Further more, the algorithm will allow larger orders of p, q, P and Q. This ensures that the maximum of possible combinations will be considered. Finally, since the TS data corresponds to hourly measurements, a 24 hours frequency will be associated to the appropriate R object. This enables the algorithm to properly test for seasonality.

3.3 ARIMA with explanatory variable

Some of the methods to extract seasonality presented in 2.1.2 involve subtracting a trend to the data. Instead of explicitly subtracting the modeled trend to the data, model the de-trended and later add the trend again, an extended ARIMA model will be used. This version of ARIMA allows for the introduction of an explanatory variable. Consider an ordinary regression model:

$$y_n = \beta_0 + \beta_1 x_{1,n} + \ldots + \beta_k x_{k,n} + \varepsilon_n, \tag{3.1}$$

where β_k are coefficients associated to the predictor variables $x_{k,n}$ and ε_n are the regression errors.

In this model, the errors become an ARIMA model, η_n. For example, if η_n is an ARMA(1,0,1) model with one explanatory variable and $\beta_0 = 0$:

$$\begin{aligned} y_n &= \beta_1 x_{1,n} + \eta_n, \\ \eta_n &= c + \phi_1 \eta_{n-1} + \theta_1 \varepsilon_{n-1} + \varepsilon_n. \end{aligned}$$

One refers to η_n as the ARIMA errors, and to ε_n as the residuals. The residuals should still be white noise and uncorrelated for the model to have performed a good fit.

The automatic search algorithm used in R is also capable of handling these types of extended ARIMA models [27, 37].

3.4 Forecast

The forecast of the ARIMA model is very straightforward. The parameters ϕ and θ are substituted with their estimated values and the subscript n is replaced with $n + h$, where h is the forecast horizon. The predictions are calculated iteratively for $h = 1, 2, \ldots, H$. For illustration purposes, consider an ARMA(1,0,1) model with no constant:

$$y_n = \phi_1 y_{n-1} + \theta_1 \varepsilon_{n-1} + \varepsilon_n. \tag{3.2}$$

Substituting n with $n + h$ with $h = 1$ yields:

$$y_{n+1} = \phi_1 y_n + \theta_1 \varepsilon_n + \varepsilon_{n+1}. \tag{3.3}$$

Where the only unknown is ε_{n+1}, which is replaced by zero, and ε_n is the last known residual from the model [21, 27].

If it is necessary to add an explanatory variable, this variable also needs to be forecasted. This procedure will be automated by R with an exponential smoothing method [39].

The other transformations like, natural logarithm, square root and differencing are already implemented in the R algorithm. Hence no such considerations are necessary [37].

3.5 Cross-Validation

When evaluating forecast performance, residual analysis is not a reliable method. In order to truly evaluate the forecast validity two methods will be employed: out-of-sample testing and TS cross-validation.

The out-of-sample testing consists simply, of holding out a subset of the data that will not be used for fitting the model. This creates a data subset that is unknown to the model, meaning that any forecasts the model makes will be genuine forecasts. Visually one can represent this method as seen in Figure 3.1.

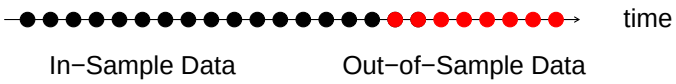

FIGURE 3.1: Out-of-sample cross-correlation diagram. Data represented as colored points as a function of time.

This method will be used as an initial look at how the models forecast unknown data since, a good fitting model is not at all indication of a good forecast. The length of the subset chosen is of 24 data points. This corresponds to about 4% of the total dataset. In the context of validation it is a small percentage. Typically, around 20% of the data is reserved but given the characteristics of TS, actually this method has several disadvantages [40, 41].

Considering only one out-of-sample dataset would yield only one forecast and only one forecast error. Because the data is autocorrelated to past values of itself, forecasting from a fixed origin could lead to distortions in the results resulting from singular events occurring at the origin [41]. Finally, the forecast error for a multi-step forecast is built by averaging point to point errors. This means that there is no representation of the evolution of the errors with the distance to the forecast origin [41].

To overcome the previous referred limitations, the classical out-of-sample cross-validation procedure needs to be modified. One possible modification is to used a rolling forecasting origin. In this method, the forecast origin is iteratively updated and new forecasts are produced starting at the new origin. A diagram for one use of this method is seen in Figure 3.2.

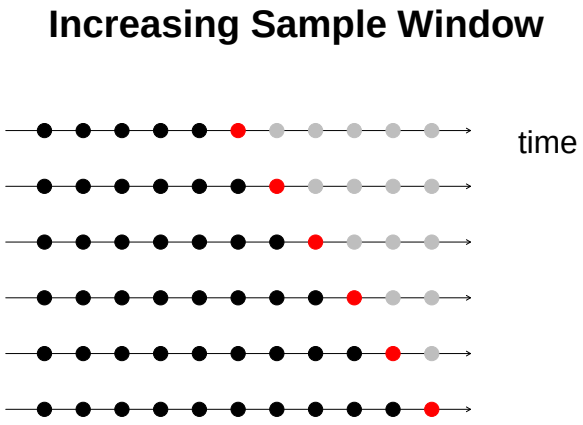

FIGURE 3.2: Diagram of a rolling forecasting origin cross-validation, with the sample window increasing. Data represented as colored points as a function of time.

Here, the forecast origin is updated by increasing the length of the data used for fitting the model. This method was implemented in this work starting with the minimum window length, which is equal to the length of the forecast. A plot of the error as a function of the window length will be produced in order to evaluate how consistent are the forecasts, considering the change in forecast origin as the length of the window increases.

Another variation on this cross-validation, represented in Figure 3.3, involves fixating the length of the sample window and just dislocating the window through the data, iteratively.

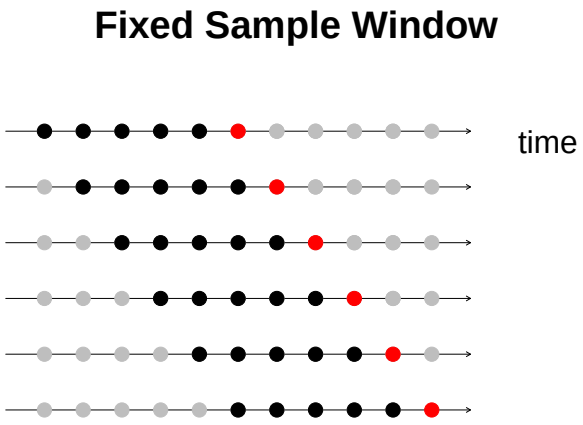

FIGURE 3.3: Diagram of a rolling forecasting origin cross-validation, with the sample window of a fixed length. Data represented as colored points as a function of time.

This method was also implemented in this work with two variations: rolling fixed window and a rolling fixed window with increasing forecast horizon. The first variation was done with a $3 \times h$ window length, where h is the forecast horizon of 24 data points. This window length is suggested in literature to ensure forecasts that are not hindered by lack of data. A plot of the error as a function of runs[1] will be produced. Here we wish to compare the results with the previous method. Hopefully it will be possible to infer if there is any difference in the errors

[1]Meaning the number of iterations of the window along the data set.

when the window length is increased versus when kept fixed. It is also possible that the forecast origin has more influence than the window length.

The second variation was performed with a window length of 400 points and forecast horizon $h = 1, \ldots, 96$. For each forecast horizon, a forecast of rolling origin with fixed window was performed and the results averaged. For this variation the errors for each different horizon will be averaged. Plotting the average error as a function of the increase in the length of the horizon will show the evolution of the error when the forecasts distance to the origin increases.

Chapter 4

Results and Discussion

The following section will be dedicated to presenting the full TS analysis, modeling and forecasting process for one pollutant, CO. Due to the extensive data generated by this type of work only the main results for the remaining pollutants will be highlighted in this chapter, the intermediate results will be presented in the Appendix. The final results for every pollutant will be discussed.

4.1 Carbon Monoxide

The first step in any type of data analysis is to plot the data and perform a visual inspection. Upon doing so, left panel of Figure 4.1, it was clear that there had been an atypical event that led to the incredibly high concentration values observed close to the 1000 hours data point. This event was the transport of particulate matter and pollutants from North Africa by the wind. It was not a equipment malfunction or human error. Nevertheless, the kind of statistical model used in this work is not capable of dealing with these extreme outliers so, they had to be removed or substituted by other values.

Using the Median Absolute Deviation (MAD) a more systematic approach to outliers was performed. Values outside an interval between plus and minus three times the value of the MAD were considered to be outliers. Some authors recommend a smaller interval [42] but a more conservative estimate was preferred. Even with a broader MAD interval, several patches of data were identified as outliers as one can see in Figure 4.2.

Dealing with outliers is not a trivial matter. On a first approach, the outliers were substituted by the mean of the data, but the intervals were so large that it created discontinuities in the TS. This problem was detected across every pollutant. Furthermore, for the purpose of comparing the forecasts performance in different pollutants, it was important to have the same amount of information for every TS. Simply removing the outliers from the data would result in a different amount of data points between pollutants and loss of autocorrelation effects.

Taking all of this into consideration it was decided to only use the longest set of data, not containing outliers, common to every pollutant. This resulted in a TS with 612 points, presented in the right side of Figure 4.1, to perform modeling and forecasting. It is not the ideal solution since the amount of data is not enough to reflect long-term seasonality but it is enough for a first study of this type of data and models.

The Normal QQ plot of the CO TS with outliers, the left panel in Figure 4.3, shows a lot of points that are not on the dotted line. This line connects the first and third quantiles and

departures from it, except at the extremes, is an indication that the sample quantiles were not generated from a normal distribution. After removing the outliers it was also verified by the Normal QQ plots, in the left panel of Figure 4.3, that it was not absurd to assume that the data was normally distributed.

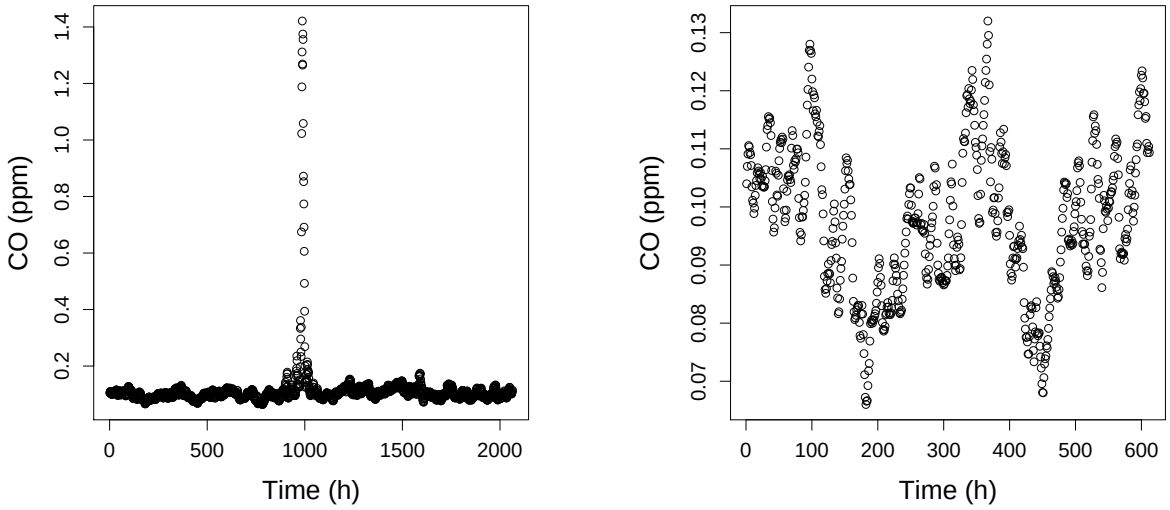

FIGURE 4.1: *Left:* CO concentration plotted over time for the complete data. *Right:* CO concentration plotted over time for the longest outlier-free data points, common to every pollutant.

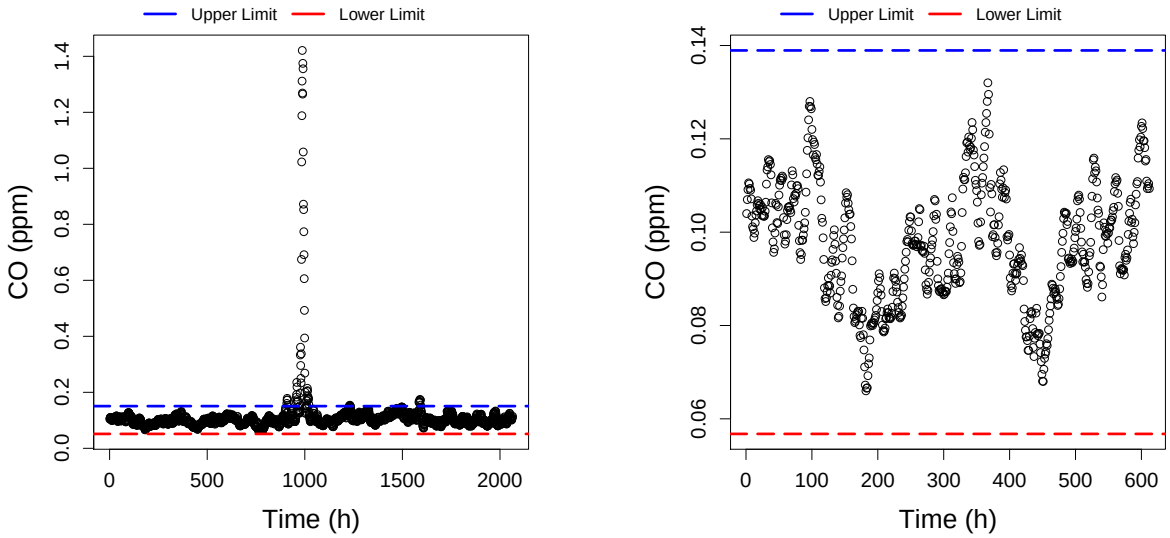

FIGURE 4.2: *Left:* CO concentration plotted over time for the complete data and MAD. The upper limit in blue and the lower limit in red. *Right:* the longest outlier-free data points for the first 612 data points and MAD. The upper limit in blue and the lower limit in red.

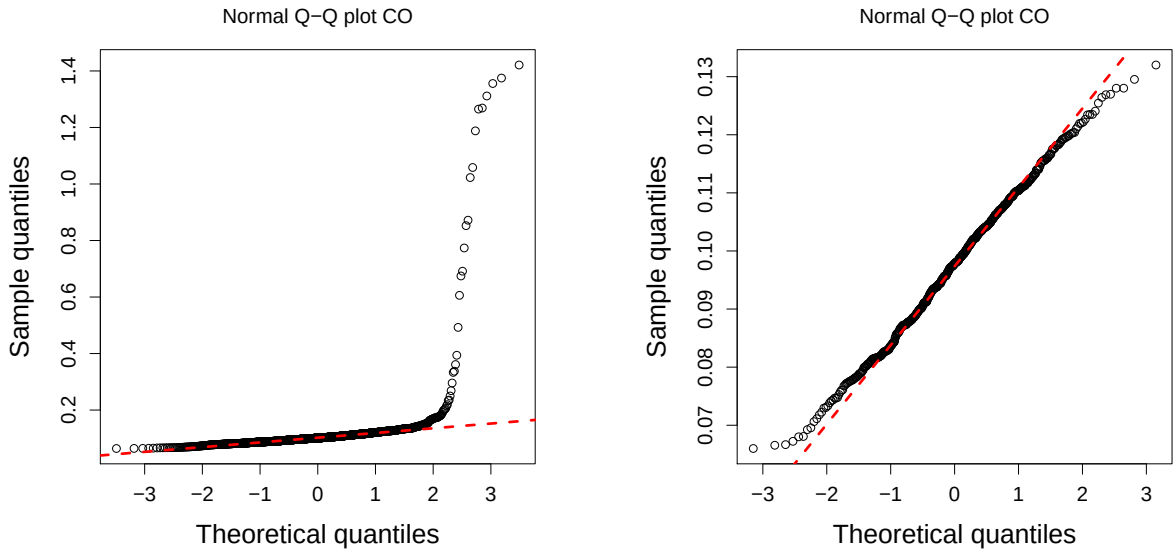

FIGURE 4.3: *Left:* QQ plot for the complete data. *Right:* QQ plot for the longest outlier-free data points.

Upon completing a more general statistical analysis, the TS analysis could start. The CO TS, represented on the left in Figure 4.1, is clearly non-stationary with a trend-cycle component and maybe some seasonal effects. Many pollutants emissions are related to human life, traffic, factories, etc, which follow patterns and seasons, that can be reflected onto the data. In an effort to stabilize the data, several transformations were tested: natural logarithm, square root, LOESS decomposition, subtraction of moving average (orders 12 and 24) and first order differencing. These are not the most contemporary methods of smoothing data but they are simple to use, understand and do work [16, 17, 27, 31]. In a way to effortlessly refer to each transformation without explicitly writing it out, the following abbreviations will be made:

CO-MA_{24} - Subtraction of a MA order 24.

CO-MA_{12} - Subtraction of a MA order 12.

ln CO - Natural logarithm.

$\sqrt{\text{CO}}$ - Square Root.

LOESS(CO) - Subtraction of the trend obtained with LOESS decomposition.

diff_1(CO) - First order differences

4.1.1 Time Series Analysis and Transformations

In the Figures 4.4, 4.5 and 4.6 the results are presented for the first three previously mentioned transformations. These produced no stabilizing results. Visually the series before and after transformation appear to be identical and the autocorrelation values in the ACF plot are significant for every lag. Finally, the last three transformations - CO-MA_{24}, CO-MA_{12} and diff_1(CO) - produced satisfactory results as shown in Figure 4.7. The visual results were further confirmed

by an ADF test. The TS CO-MA_{24}, CO-MA_{12} and $diff_1$(CO) are stationary with a 99% confidence level according to the test. All the results are presented in Table 4.1

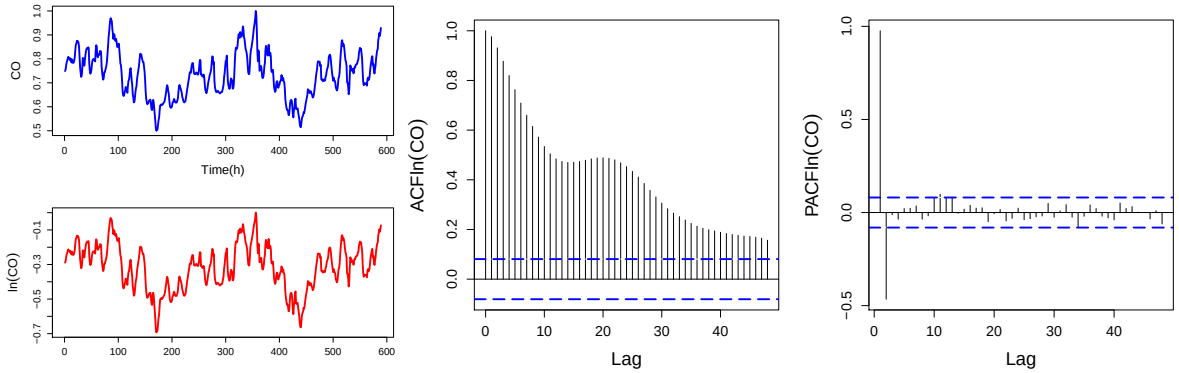

FIGURE 4.4: *Left:* Normalized CO compared to ln(CO) plotted over time. *Center:* ACF plotted over lags for ln(CO). *Right:* PACF plotted over lags for ln(CO).

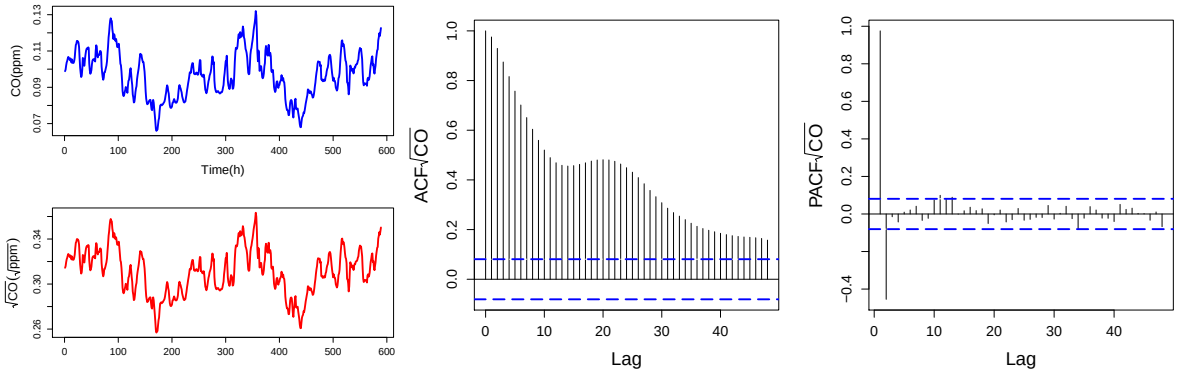

FIGURE 4.5: *Left:* CO concentrations compared to $\sqrt{CO}$ plotted over time. *Center:* ACF plotted over lags for $\sqrt{CO}$. *Right:* PACF plotted over lags for $\sqrt{CO}$.

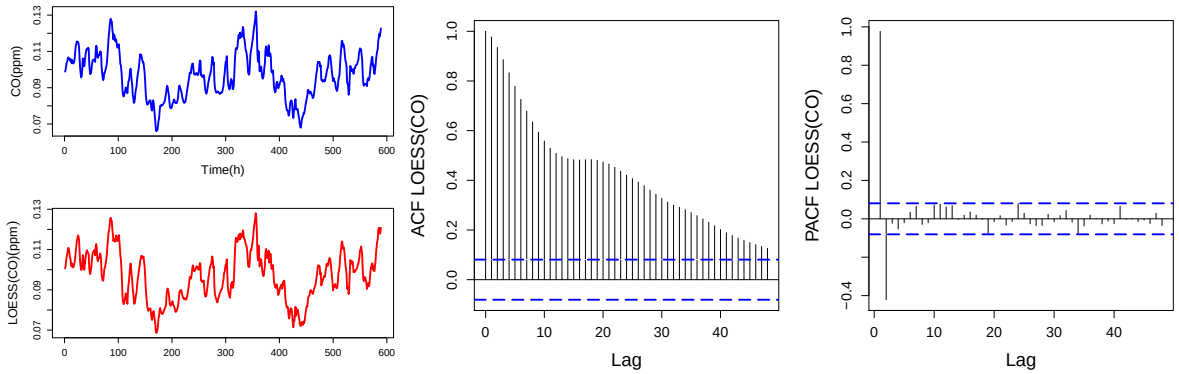

FIGURE 4.6: *Left:* CO concentrations compared to LOESS(CO) plotted over time. *Center:* ACF plotted over lags for LOESS(CO). *Right:* PACF plotted over lags for LOESS(CO).

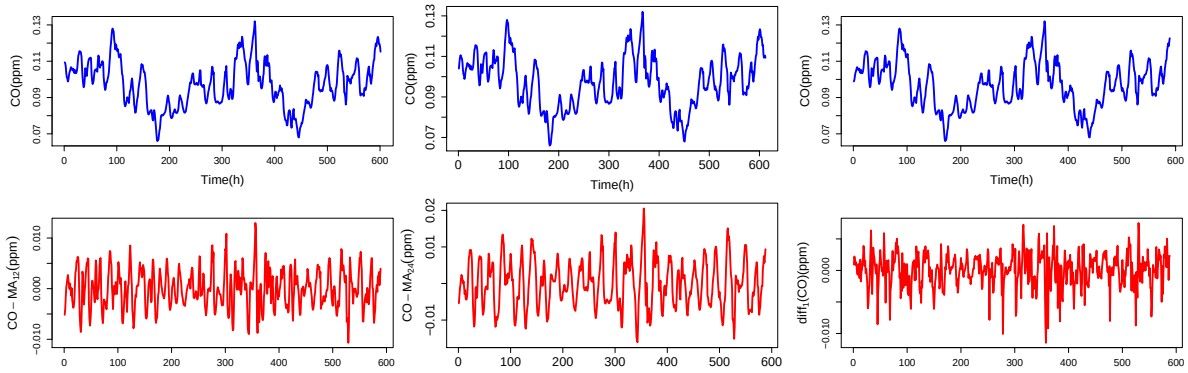

FIGURE 4.7: *Left:* CO concentrations compared to CO-MA_{12} plotted over time. *Center:* CO concentrations compared to CO-MA_{24} plotted over time. *Right:* CO concentrations compared to diff_1(CO) plotted over time.

TABLE 4.1: ADF test results and respective p-values for every transformation.

Transformation	**ADF test**	**p-value**
CO-MA_{24}	-6.93	0.01
CO-MA_{12}	-7.99	0.01
$\ln \text{CO}$	-2.39	0.41
$\sqrt{\text{CO}}$	-2.36	0.43
LOESS(CO)	-2.18	0.50
diff_1(CO)	-5.36	0.01

Determining possible model parameters is the next step in the Box-Jenkins approach and to do so, the ACF and PACF plots of the TS are closely studied. Significant lags in both ACF and PACF for every transformation suggest a mixed ARMA model.

For the CO-MA_{12} and CO-MA_{24} transformations the ACF plots, see left panel of Figures 4.8 and 4.9, show a damped periodical behavior which is typical of an AR(2) process. Combining this information with the major cut-off at lag(2) in the PACF plots, see right panel of Figures 4.8 and 4.9, seems to reinforce the estimation. This will have to be tested considering that are a few more significant lags in the PACF plot. Since there are no major cut-offs in the ACF plots there are no expectations for the value of the MA parameter could be.

For the diff_1(CO) transformation the ACF and PACF behavior is not so obvious. Both parameters will be estimated iteratively. A AR(2) behavior is expected since the ACF plot in the left panel of Figure 4.10 shows some kind of periodicity, but the cut-off on the PACF plot in the right panel of Figure 4.10 is in lag(1). Like the previous ACF plots there is not a clear cut-off and it is not possible to estimate a MA parameter only by analyzing the plots.

There appears to be some seasonality in the data since there are significant lags beyond the initial ones, but the R algorithm will also account for that and add seasonal parameters if necessary. After the fit, residual analysis will be performed.

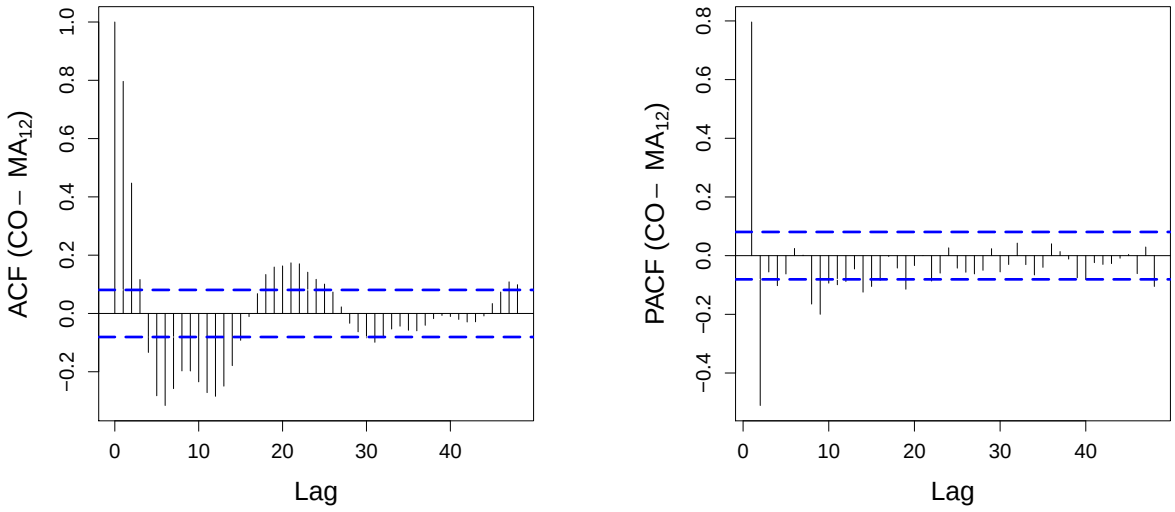

FIGURE 4.8: ACF and PACF plots for the CO-MA_{12} transformation.

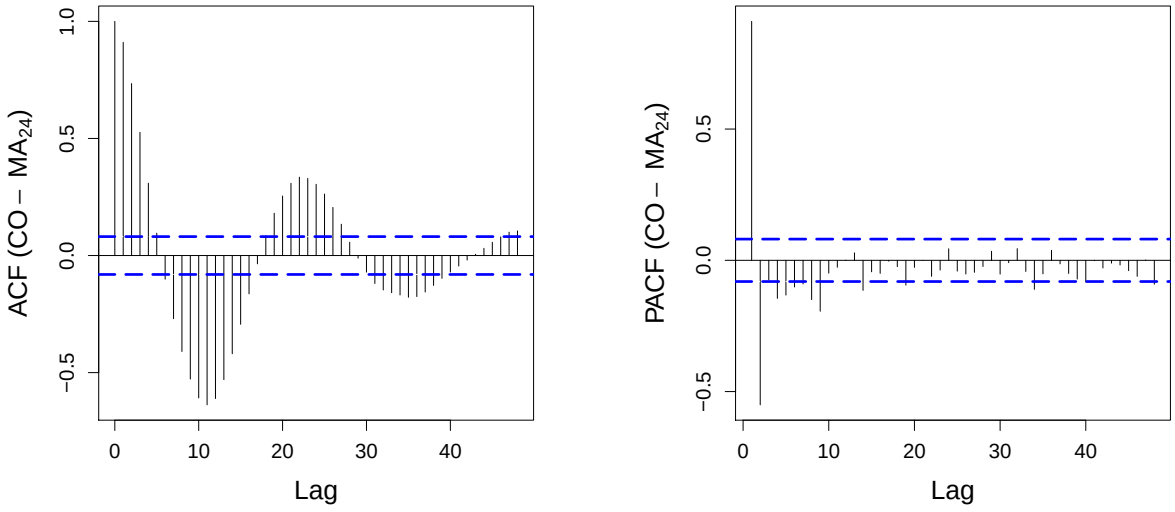

FIGURE 4.9: ACF and PACF plots for the CO-MA_{24} transformation.

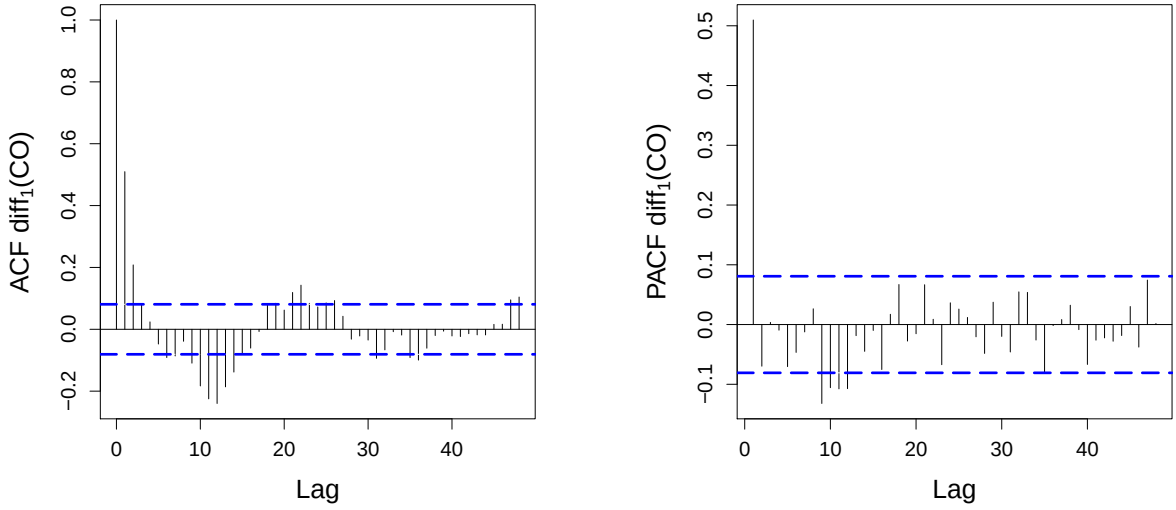

FIGURE 4.10: ACF and PACF plots for the $\text{diff}_1(\text{CO})$ transformation.

4.1.2 Fitting and Forecasting

To determine the best fitting model, 564 data points were used and 24 were saved to perform a forecast. After running the R algorithm the best models to fit each transformation and their respective forecasts, Figures 4.15, 4.11 and 4.13, were plotted against the original data. The ARIMA(p,d,q) orders the algorithm converged to were:

- ARIMA(5,0,2), for the CO-MA_{12};
- ARIMA(4,0,1), for the CO-MA_{24};
- ARIMA(5,1,1), for the $\text{diff}_1(\text{CO})$.

The model ARIMA(5,0,2) produced the best results in the fit since it had the lowest errors values, followed by ARIMA(4,0,1) and ARIMA(5,1,1) as shown in Table 4.2. To further assess the goodness of fit, residual analysis was performed for every model. Visually, the plotted residuals appear to be white noise, as one can observe in the top panels of Figures 4.16, 4.12 and 4.14. A closer inspection shows there are significant lags both in the ACF, bottom left panels of Figures 4.16, 4.12 and 4.14, and PACF plots of every model residuals, bottom right panels of Figures 4.16, 4.12 and 4.14. Every model, ARIMA(5,0,2), ARIMA(4,0,1) and ARIMA(5,1,1), passed the Ljung-Box test for autocorrelation. This implies that the models were not able to fully capture the information in the data. They can still be used to forecast but the prediction intervals might not be very accurate and the errors might be underestimated [27].

Forecasting is a very different matter and a good fitting model does not equate to a good forecast. This was not the case, as shown by the results in Table 4.3. The best fitting model had the best forecast performance followed by ARIMA(5,1,1) and lastly ARIMA(4,0,1) with the worst results. To further test the robustness of these results cross-validation was performed and is presented in the next subsection.

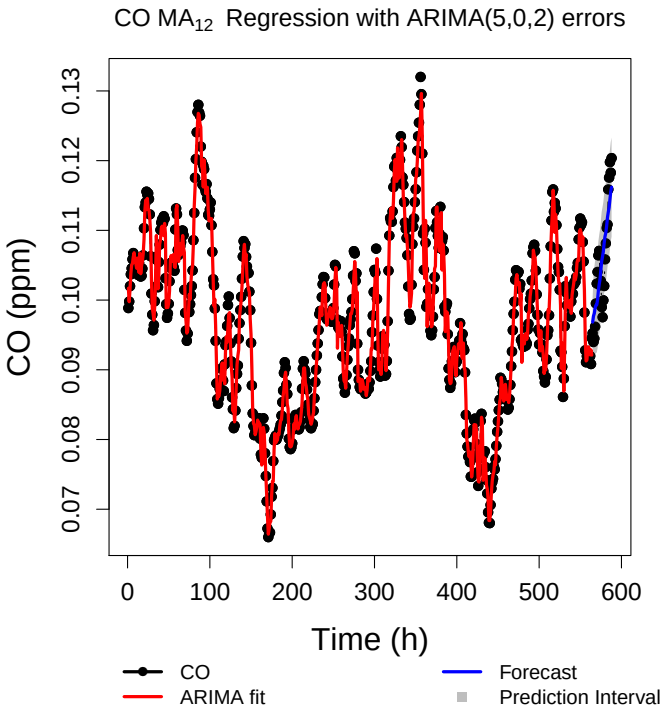

FIGURE 4.11: Best fitting model determined by R for the CO-MA_{12} transformation in red, forecast of 24 hours in blue and the 95% prediction interval for the forecast.

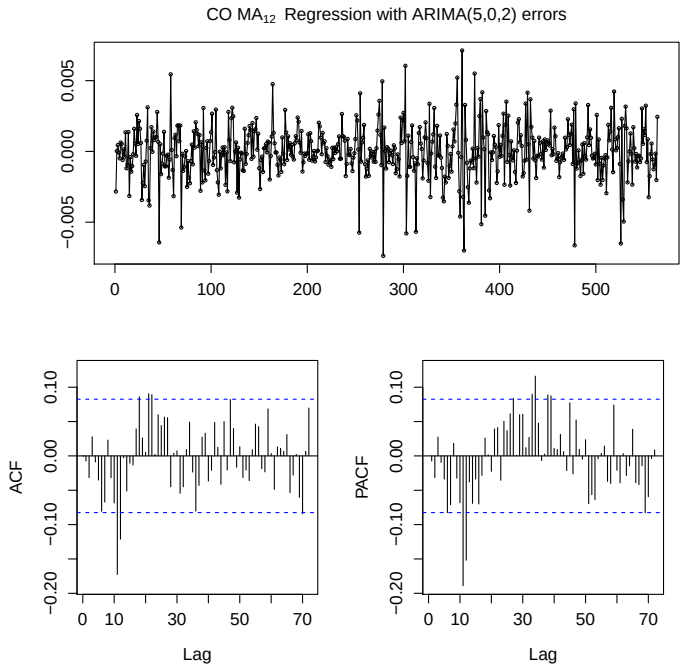

FIGURE 4.12: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

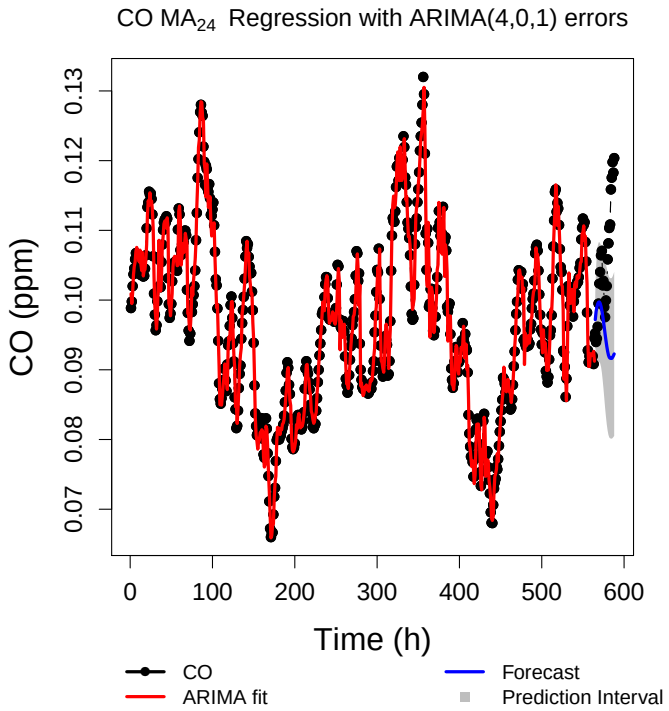

FIGURE 4.13: Best fitting model determined by R for the CO-MA_{24} transformation in red, forecast of 24 hours in blue and the 95% prediction interval for the forecast.

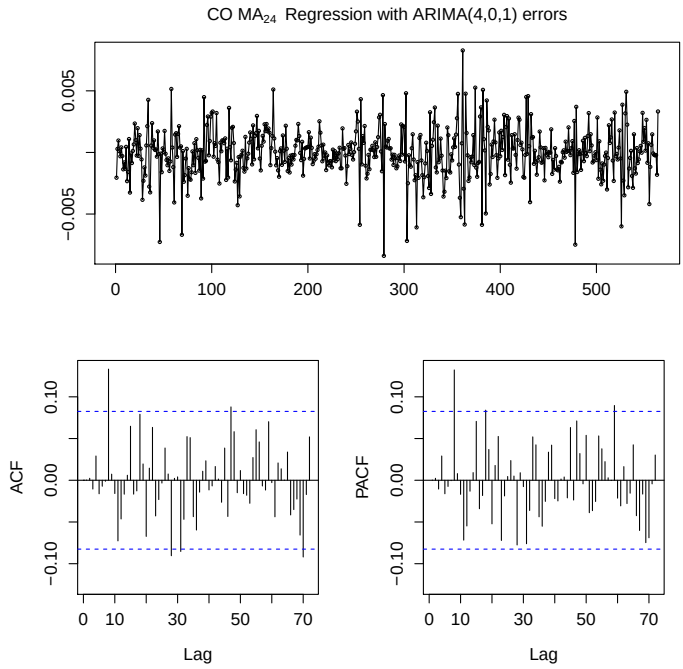

FIGURE 4.14: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

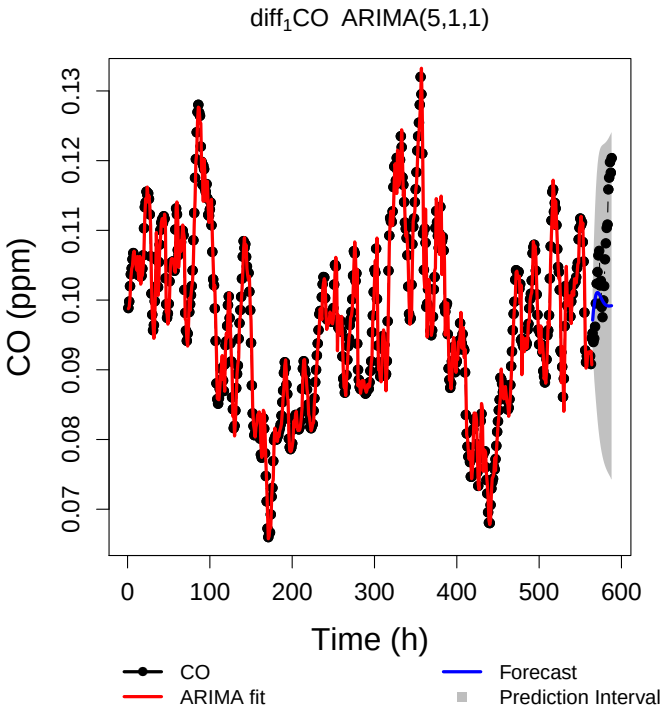

FIGURE 4.15: Best fitting model determined by R for the $\text{diff}_1(\text{CO})$ transformation in red, forecast of 24 hours in blue and the 95% prediction interval for the forecast.

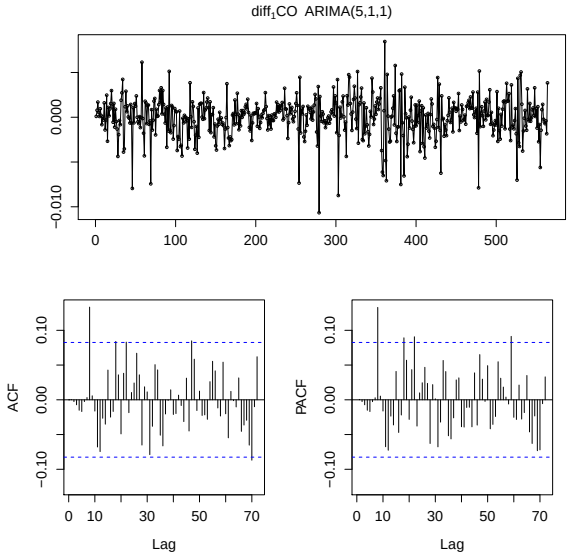

FIGURE 4.16: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

TABLE 4.2: Values of the RMSE, AIC, BIC and Ljung-box test of the residuals of every model.

	RMSE [10^{-3}ppm]	**MASE**	**Ljung-Box**	**p-value**
CO-MA$_{12}$	1.84	0.673	37.4	0.01
CO-MA$_{24}$	1.96	0.717	11.1	0.01
diff$_1$(CO)	2.20	0.810	10.7	0.03

TABLE 4.3: Values of the RMSE, MAE and MASE for the forecasts produced by each model.

	RMSE [10^{-3}ppm]	**MAE** [10^{-3} ppm]	**MASE**
CO-MA$_{12}$	4.16	3.56	1.81
CO-MA$_{24}$	14.5	11.1	5.67
diff$_1$(CO)	10.0	7.36	3.76

4.1.3 Time Series Cross-Validation

After finding the best models, the next step is to test the robustness of the forecasts. Since ARIMA models depend on past values, they can prove to only be effective in some section of the TS. To avoid this, forecasts with rolling origin were performed with two variants: increasing data window and a fixed data window.

For the increasing data window, the ARIMA models determined previously, are applied to the data window without parameter[1] re-estimation. A 24 hour forecast is performed and the error measured. The data window is increased by 1 data point and the process is repeated until there are no more data points available.

The second method has a data window with a fixed length, 96 data points. The models determined are still applied to the data window without parameter re-estimation and a 24 hour forecast performed. Next, the window is dislocated 1 point ahead and the process repeats itself until no more data points are available. The expectation, for the fixed window is that the error values, if the models indeed produce robust forecasts, will have small oscillations around a mean value but not an overall increasing or decreasing trend.

In Figure 4.17, the evolution of RMSE, MAE and MASE were tracked when performing a rolling origin forecast with an increasing window. The errors in CO-MA$_{24}$ appear be more stable after a certain window size, but CO-MA$_{12}$ and diff$_1$(CO) have very volatile behaviors besides larger error values. In Figure 4.18, where the rolling forecast was performed with a fixed window, it seems that the same behavior is observed. CO-MA$_{24}$ has smaller errors values, which was unexpected according to the previous forecast results.

Finally, to check how the errors evolved when the length of the forecasts increased, a variation of the rolling forecast with fixed window was performed. The size of the window was 400 data points to try eliminate errors from insufficient data. This time the errors resulting of the the rolling forecast are averaged and the process is repeated for a forecast 1 data point longer.

The average RMSE,MAE and MASE were plotted as a function of the forecast length in hours is shown in Figure 4.19. The error for the different models evolves differently. For shorter forecasts, until approximately 6 data points, the errors of CO-MA$_{24}$ and CO-MA$_{12}$ are very similar. After approximately 60 data points, the error for CO-MA$_{24}$ continues to grow linearly and for diff$_1$(CO) starts to stabilize. This makes diff$_1$(CO) the best model for longer forecasts and CO-MA$_{24}$ the best model for the forecast length of 24 hours. This information was not possible to observe in previous results, Figures 4.17 and 4.18, since the forecasts were of the length chosen to optimize in the course of this work, 24 hours. It also shows how important it is

[1]Here, parameter refers to the (p,d,q) ARIMA parameters and not to the model coefficients ϕ and θ referred in previous chapters.

to test the forecast through out the whole TS. Judging only by the information of the previous subsection, the user would wrongly assume the best model for forecasting to be CO-MA_{12} when it is in fact CO-MA_{24}.

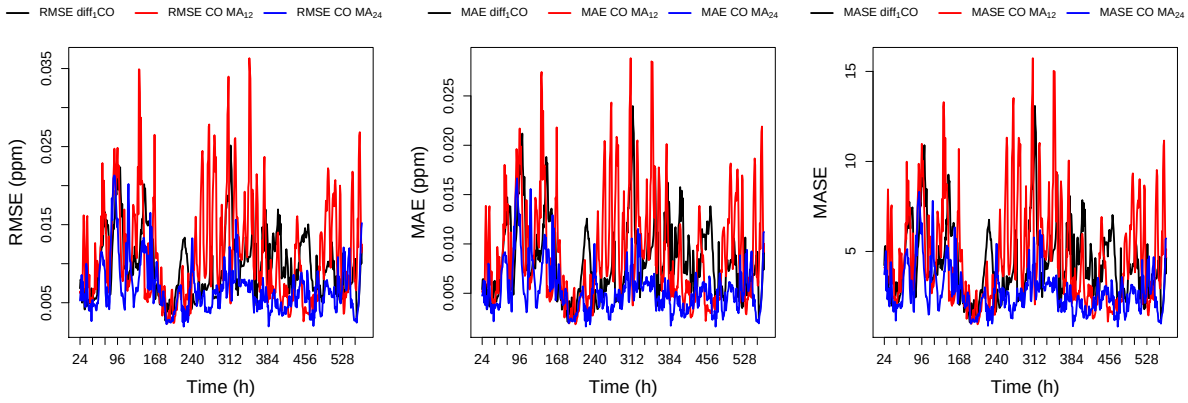

FIGURE 4.17: *Left:* RMSE of different models plotted as a function of data points used by the model. *Center:* MAE of different models plotted as a function of data points used by the model. *Right:* MASE of different models plotted as a function of data points used by the model.

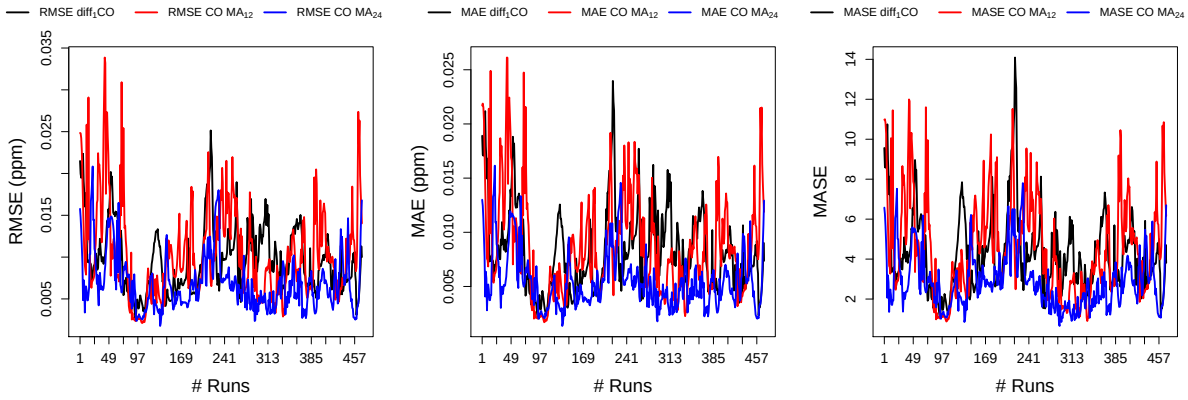

FIGURE 4.18: *Left:* RMSE of different models plotted as a function of the number of fixed windows. *Center:* MAE of different models plotted as a function of fixed windows. *Right:* MASE of different models plotted as a function of fixed windows.

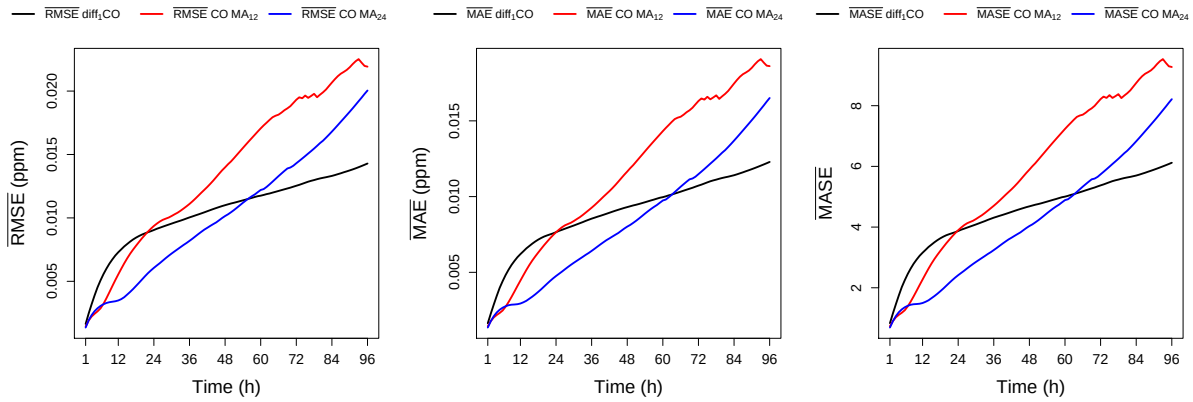

FIGURE 4.19: RMSE, MAE and MASE of different models plotted as a function of the forecast length.

For the remaining pollutants, the TS Cross validation will be discussed in Section 4.7 for a more effortless comparison between results.

4.2 Nitrogen Dioxide

The same procedure of data analysis was performed for NO_2. The outlier free data is presented in Figure 4.20. As seen and mentioned in the previous section the amount of information produced is too massive to be conveniently presented in the body of this thesis. Every plot not presented here is available in Appendix A.

Similar to the CO results, after the transformations the only ones that yielded stationary TS where $NO_2 - MA_{12}$, $NO_2 - MA_{24}$ and $\text{diff}_1(NO_2)$. The ADF test results for every transformation are presented in Table 4.4.

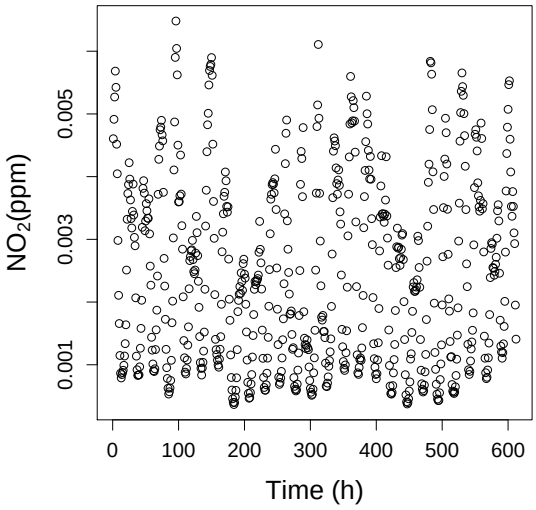

FIGURE 4.20: NO_2 concentration plotted over time for the longest outlier-free data points, common to every pollutant.

TABLE 4.4: ADF test results and respective p-values for every transformation.

Transformation	ADF test	p-value
NO_2-MA_{24}	-8.64	0.01
NO_2-MA_{12}	-10.4	0.01
$\ln NO_2$	-3.12	0.11
$\sqrt{NO_2}$	-3.27	0.08
LOESS(NO_2)	-3.28	0.08
$diff_1(NO_2)$	-6.44	0.01

The models determined by R as the best for the transformations used are:

- ARIMA(2,0,3), for the $NO_2 - MA_{12}$;
- ARIMA(4,0,0), for the $NO_2 - MA_{24}$;
- ARIMA(2,1,5), for the $diff_1(NO_2)$.

None of the models failed the Ljung-Box test and is evident from the residual analysis in Figures 4.22, 4.24 and 4.26 that there is still autocorrelation present. In regards to fit performance Table 4.5 shows that the best model is ARIMA(2,0,3) for $NO_2 - MA_{12}$, followed by ARIMA(2,1,5) for $diff_1(NO_2)$ and lastly ARIMA(4,0,0) for $NO_2 - MA_{24}$.

The forecast performance is very similar for each model as seen in Table 4.6. The model for $diff_1(NO_2)$ has a slightly better performance.

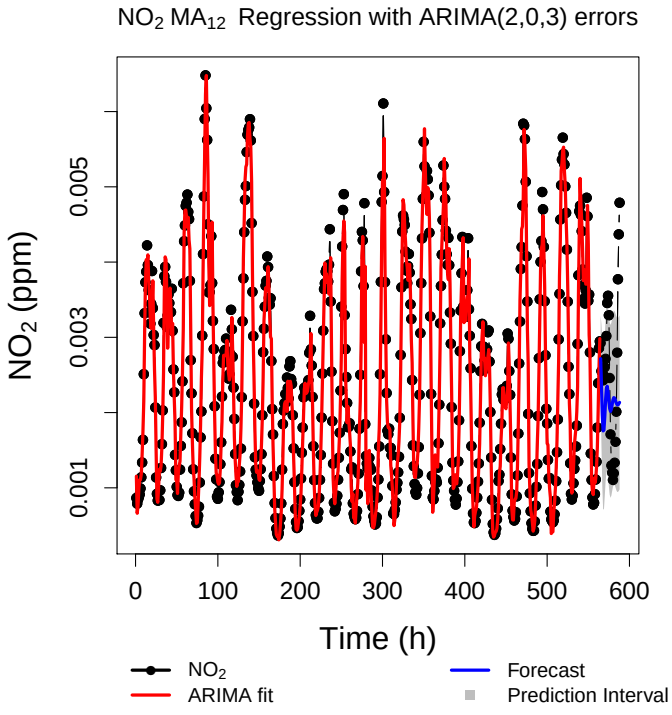

FIGURE 4.21: Best fitting model determined by R for the NO_2-MA_{12} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

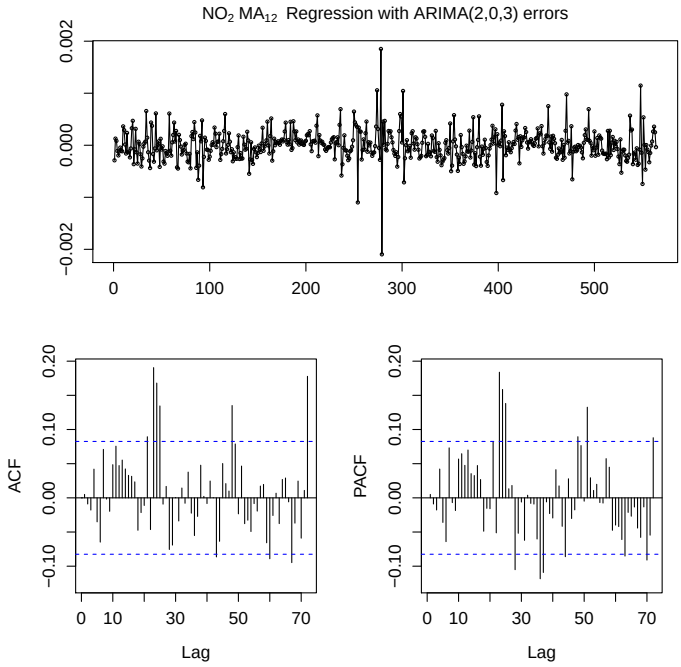

FIGURE 4.22: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

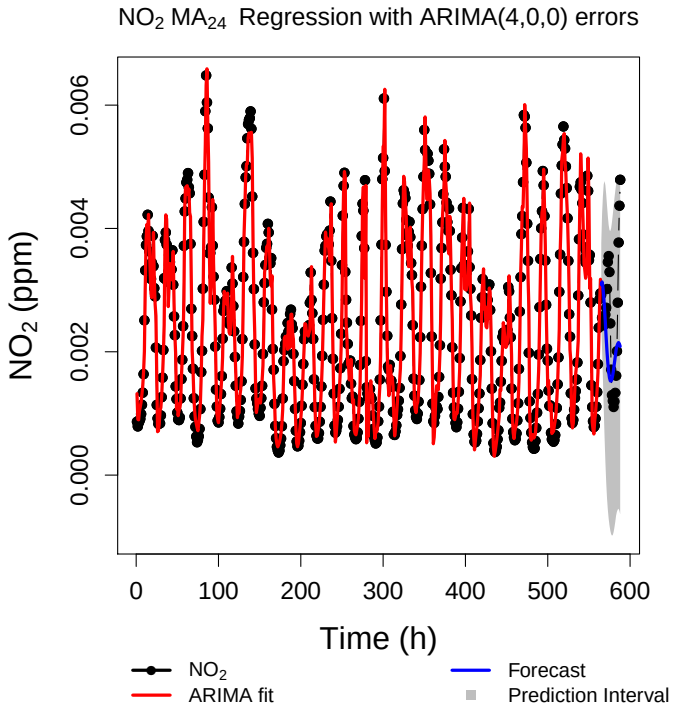

FIGURE 4.23: Best fitting model determined by R for the NO_2-MA_{24} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

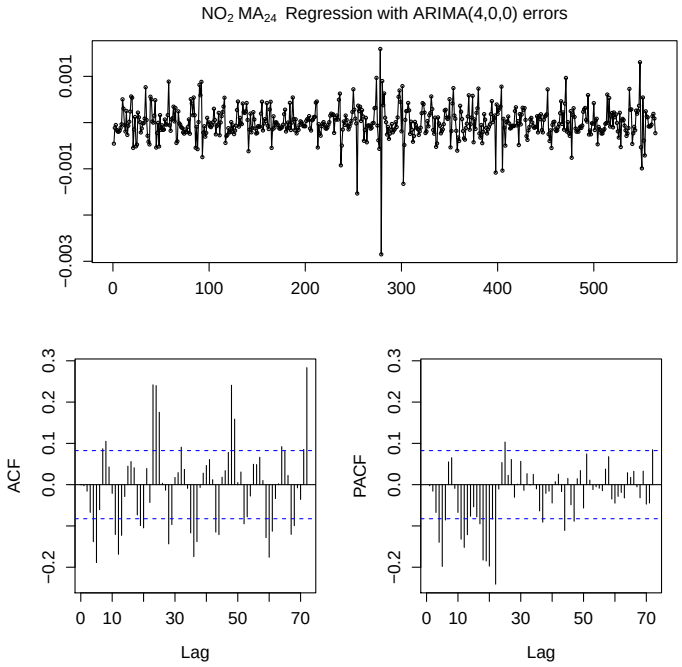

FIGURE 4.24: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

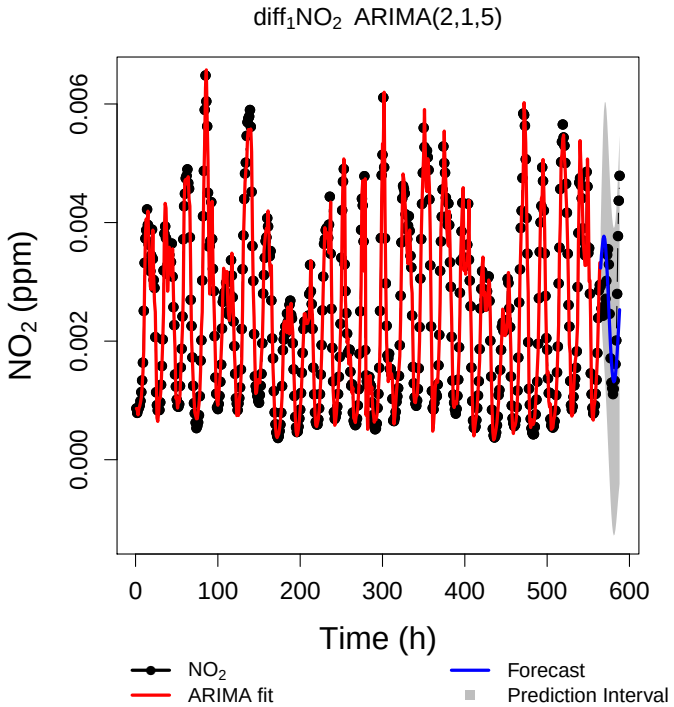

FIGURE 4.25: Best fitting model determined by R for the $\text{diff}_1(NO_2)$ transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

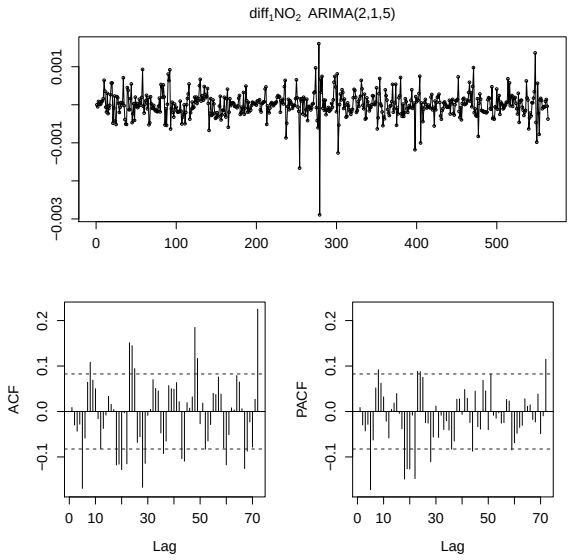

FIGURE 4.26: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

TABLE 4.5: Values of the RMSE, AIC, BIC and Ljung-box test of the residuals of every model.

	RMSE [10^{-4}ppm]	**MASE**	**Ljung-Box**	**p-value**
NO_2-MA_{12}	2.89	0.563	8.81	0.03
NO_2-MA_{24}	3.40	0.650	48.0	0.01
$diff_1(NO_2)$	3.38	0.631	33.7	0.01

TABLE 4.6: Values of the RMSE, MAE and MASE for the forecasts produced by each model.

	RMSE [10^{-3}ppm]	**MAE** [10^{-4} ppm]	**MASE**
NO_2-MA_{12}	1.06	8.75	2.44
NO_2-MA_{24}	1.10	8.42	2.35
$diff_1(NO_2)$	0.977	7.60	2.12

4.3 Sulfur Dioxide

After cutting the SO_2 TS to 612 data points there were still a few outliers. This was always a possibility since the MAD intervals for the full length data are not the same as can be seen in left and right panels of Figure A.15, respectively. The variation was not very extreme and so it was decided not to further treat the data. The SO_2 TS is presented in Figure 4.27.

After the transformations performed, the stationarity of the data was tested and the results presented in Table 4.7. Every transformation is stationary above the 95% confidence level but only the moving averages and first order differentiation are stationary above the 99% confidence level. Subsequently these were the transformations chosen to perform model fitting and forecasting.

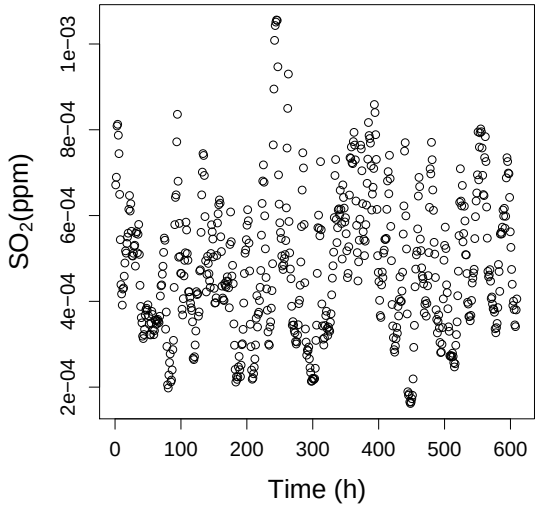

FIGURE 4.27: SO_2 concentration plotted over time for the longest outlier-free data points, common to every pollutant.

TABLE 4.7: ADF test results and respective p-values for every transformation.

Transformation	ADF test	p-value
SO_2-MA_{24}	-7.44	0.01
SO_2-MA_{12}	-8.64	0.01
$\ln SO_2$	-3.67	0.03
$\sqrt{SO_2}$	-3.64	0.03
LOESS(SO_2)	-3.60	0.03
$\text{diff}_1(SO_2)$	-6.46	0.01

The best models, for each transformation, determined by the R algorithm were:

- ARIMA(5,0,0), for the $SO_2 - MA_{12}$;
- ARIMA(5,0,0), for the $SO_2 - MA_{24}$;
- ARIMA(3,1,3), for the $\text{diff}_1(SO_2)$.

The best performing model regarding only the fit was ARIMA(5,0,0) for $SO_2 - MA_{12}$, see Figure 4.28, but this also had the lowest p-value in the Ljung-Box test, meaning that the residuals are correlated. Indeed, there are several significant peaks in the left and right bottom panels in Figure 4.29. The second best model is ARIMA(5,0,0) for $SO_2 - MA_{24}$, see Figure 4.30. Unfortunately this model also passed the Ljung-Box test and it is also possible to see a few significant peaks in the left and right bottom panels of Figure 4.31. The worst fitting model was ARIMA(3,1,3), see Figure 4.32. This model is the only one that failed the Ljung-Box test, as seen in Table 4.8. The ADF and PACF plots of the residuals, see left and bottom panels of Figure 4.33, only have significant peaks after lag 20.

The forecast performance is much different. The best performing model was for $\text{diff}_1(SO_2)$. It had the lower values for RMSE, MAE and MASE. The results for the forecast performance are shown in Table 4.8.

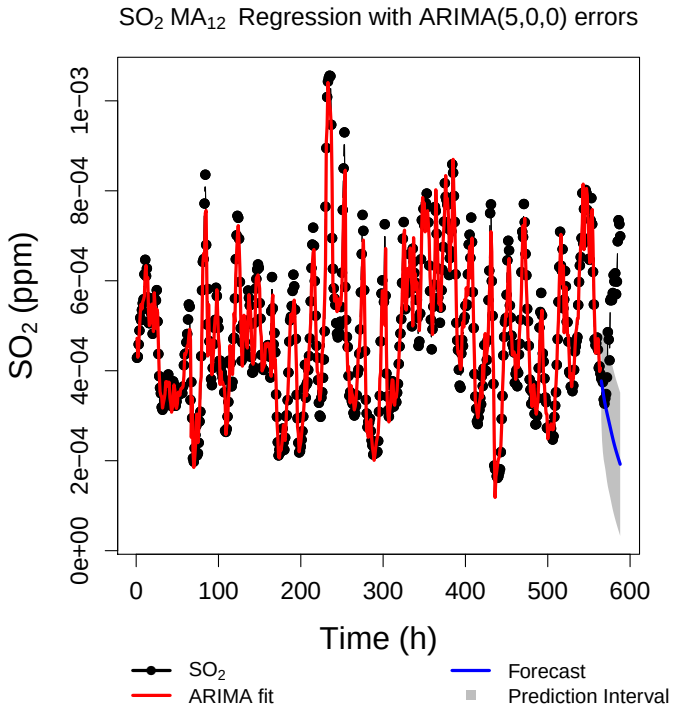

FIGURE 4.28: Best fitting model determined by R for the SO_2-MA_{12} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

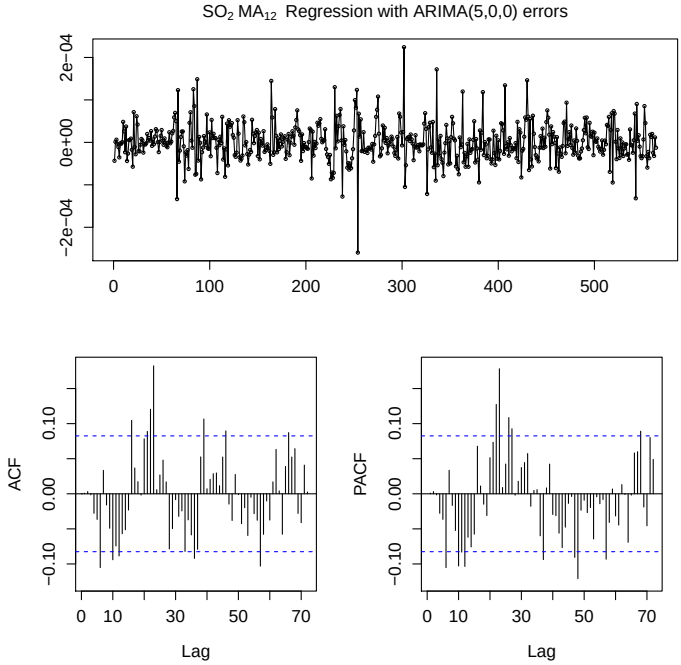

FIGURE 4.29: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

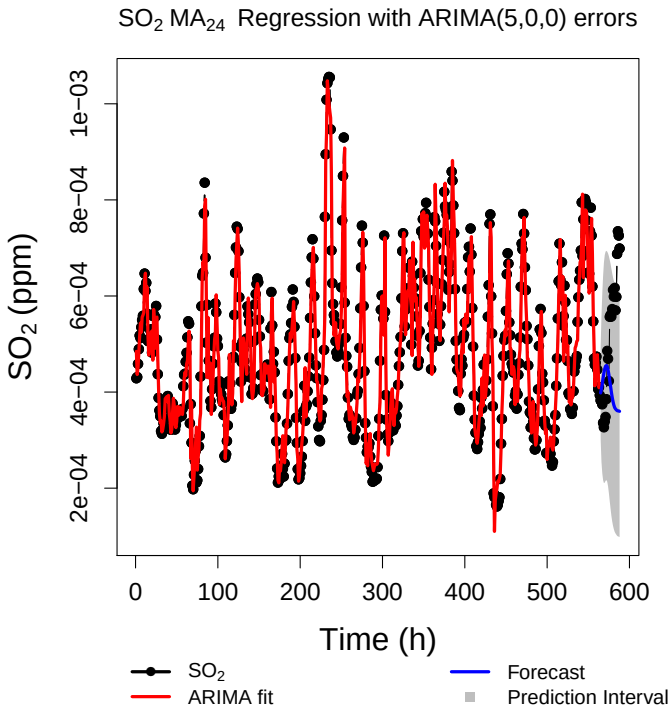

FIGURE 4.30: Best fitting model determined by R for the SO_2-MA_{24} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

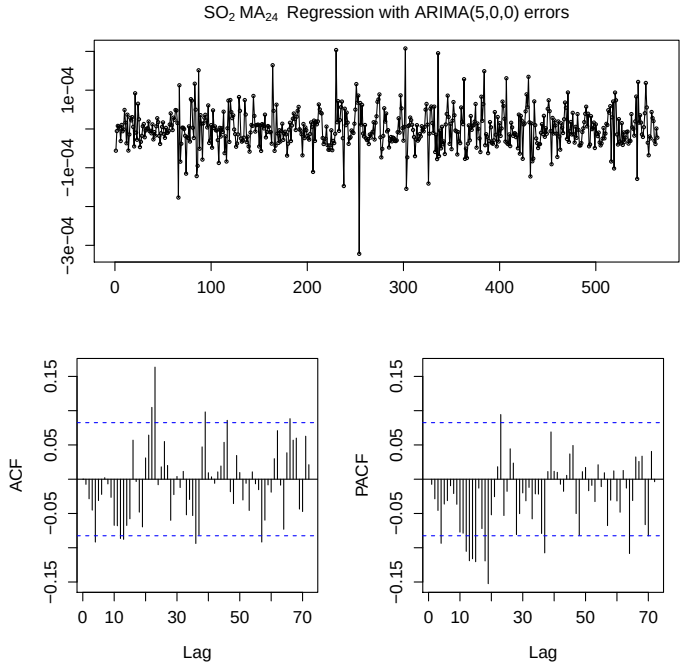

FIGURE 4.31: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

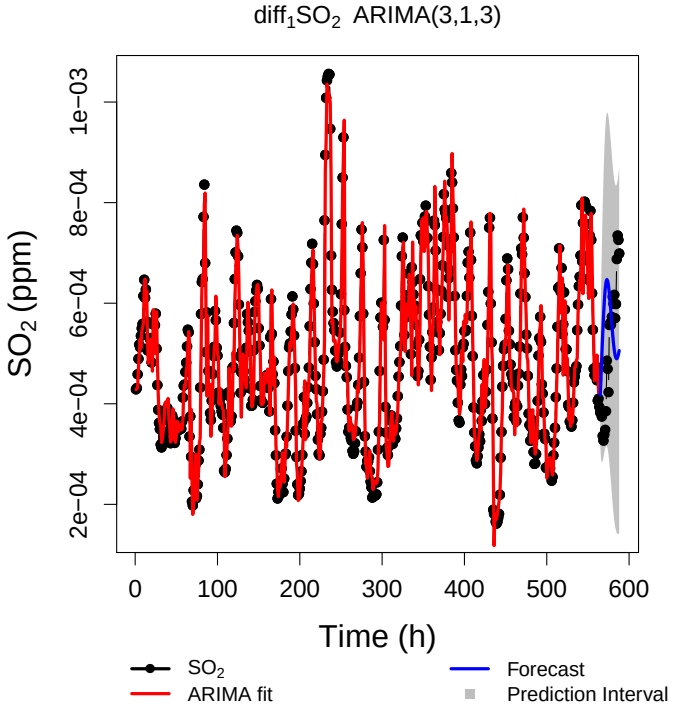

FIGURE 4.32: Best fitting model determined by R for the $\text{diff}_1(SO_2)$ transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

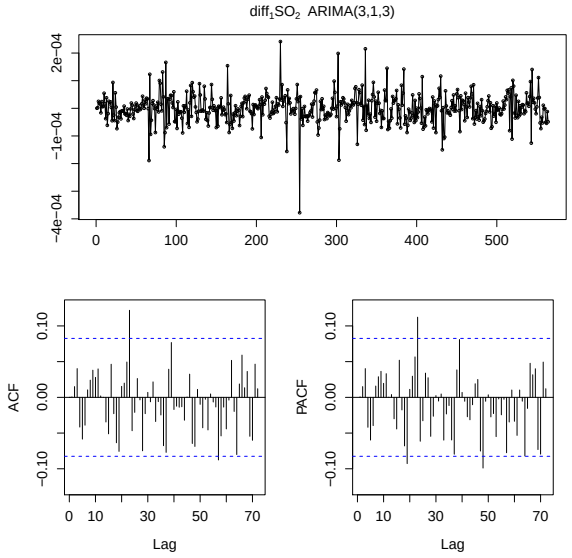

FIGURE 4.33: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

TABLE 4.8: Values of the RMSE, AIC, BIC and Ljung-box test of the residuals of every model.

	RMSE [10^{-5}ppm]	**MASE**	**Ljung-Box**	**p-value**
SO_2-MA_{12}	4.41	0.731	14.8	0.01
SO_2-MA_{24}	4.88	0.805	10.3	0.04
$\text{diff}_1(SO_2)$	5.16	0.837	6.54	0.16

TABLE 4.9: Values of the RMSE, MAE and MASE for the forecasts produced by each model.

	RMSE [10^{-4}ppm]	**MAE** [10^{-4} ppm]	**MASE**
SO_2-MA_{12}	3.02	2.41	5.58
SO_2-MA_{24}	1.95	1.59	3.68
$diff_1(SO_2)$	1.68	1.45	3.36

4.4 Ozone

After analyzing outlier free O_3 TS, see Figure 4.34, different transformations were applied to stabilize the data. Results of the stationarity test are presented in Table 4.10. The transformations that passed the test, $O_3 - MA_{12}$, $O_3 - MA_{24}$, $diff_1(O_3)$ were used to estimate ARIMA models.

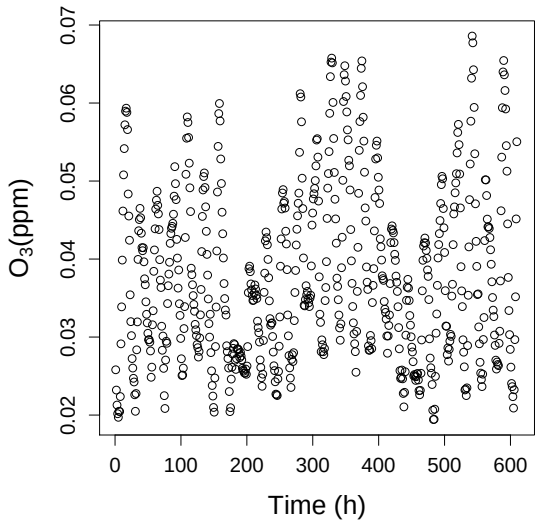

FIGURE 4.34: O_3 concentration plotted over time for the longest outlier-free data points, common to every pollutant.

TABLE 4.10: ADF test results and respective p-values for every transformation.

Transformation	**ADF test**	**p-value**
O_3-MA_{24}	-8.43	0.01
O_3-MA_{12}	-11.5	0.01
$\ln O_3$	-2.45	0.39
$\sqrt{O_3}$	-2.44	0.39
LOESS(O_3)	-2.34	0.44
$diff_1(O_3)$	-7.10	0.01

The best models, for each transformation, determined by the R algorithm were:

- ARIMA(5,0,1), for the $O_3 - MA_{12}$;
- ARIMA(5,0,0), for the $O_3 - MA_{24}$;

- ARIMA(2,1,2), for the $\text{diff}_1(O_3)$.

The best performing model regarding fitting the data is ARIMA(5,0,1), followed by ARIMA(5,0,0) and ARIMA(2,1,2). As seen on Table 4.11, all of the models passed the Ljung-Box test and it is also evident that the residuals are still autocorrelated when looking at Figures 4.36, 4.38 and 4.40. Although the residuals appear to be white-noise, there are still many significant peaks in the ACF and PACF plot of every model residuals.

For the forecast the results are presented in Table 4.12. The performances are quite different from the fit. The best forecast was done by the model ARIMA(2,1,2), see Figure 4.39. The second best result was for model ARIMA(5,0,0), see Figure 4.37 and the worst forecast was performed by model ARIMA(5,0,1), see Figure 4.35.

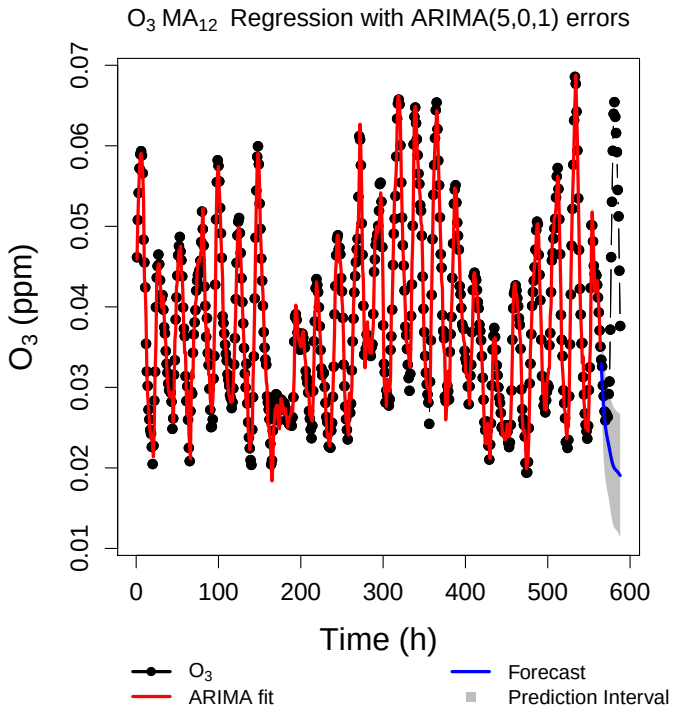

FIGURE 4.35: Best fitting model determined by R for the O_3-MA_{12} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

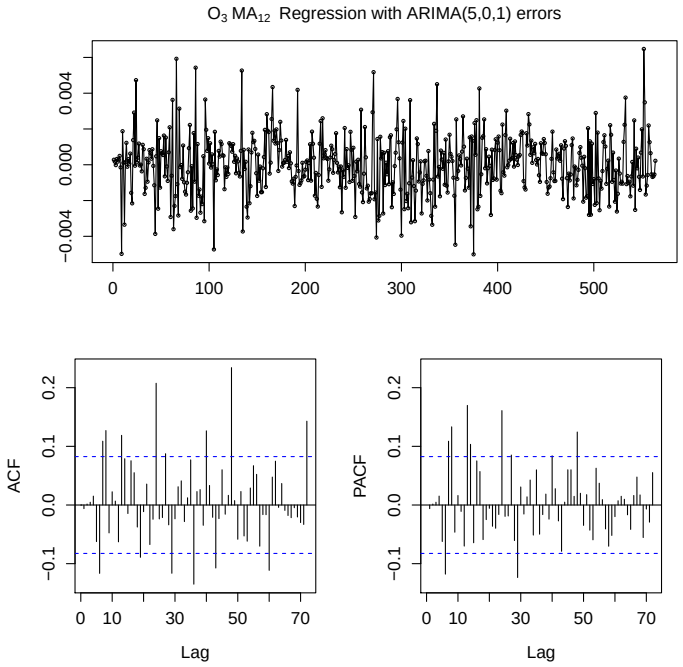

FIGURE 4.36: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

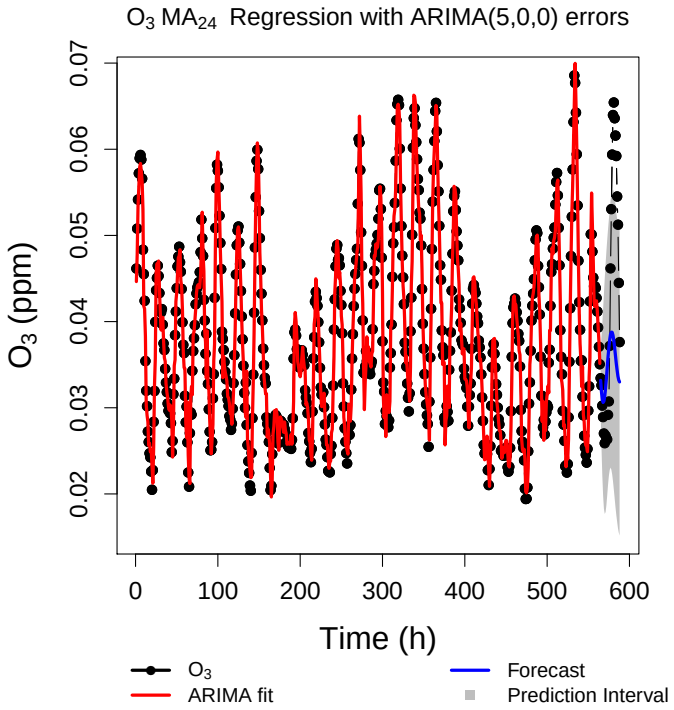

FIGURE 4.37: Best fitting model determined by R for the O_3-MA_{24} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

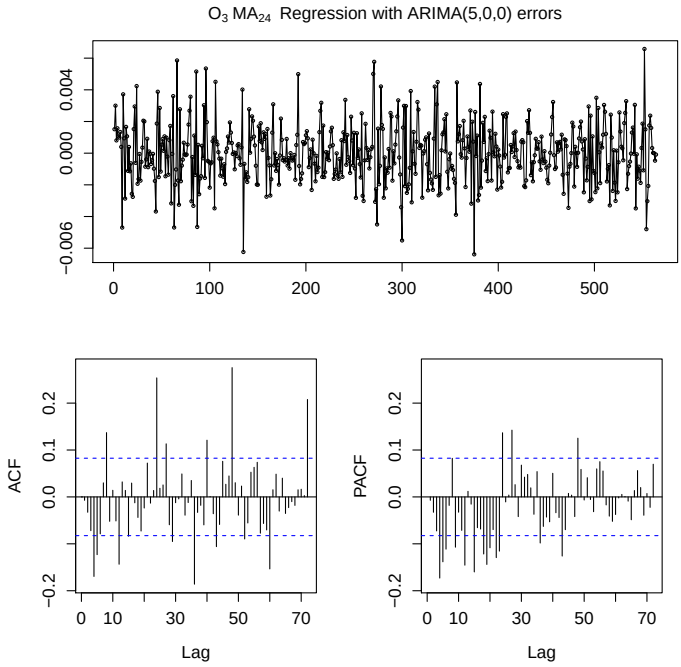

FIGURE 4.38: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

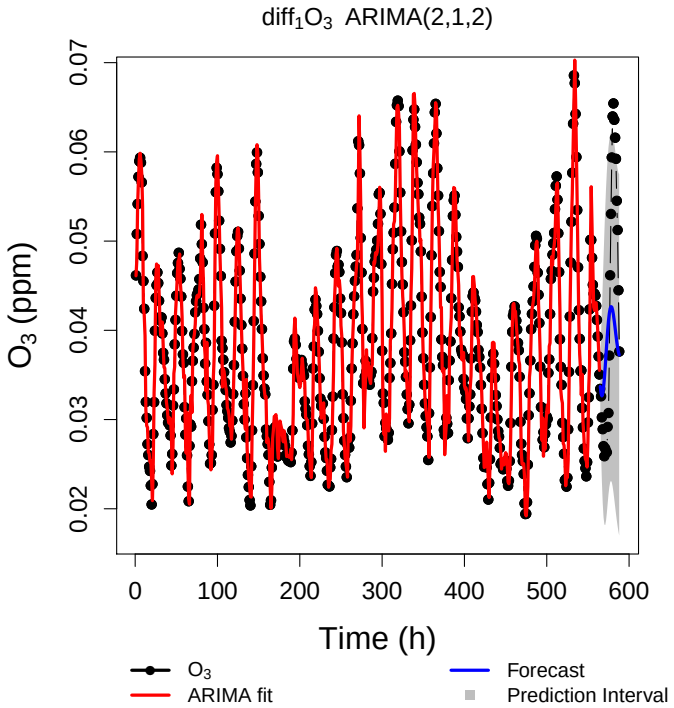

FIGURE 4.39: Best fitting model determined by R for the $\text{diff}_1(O_3)$ transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

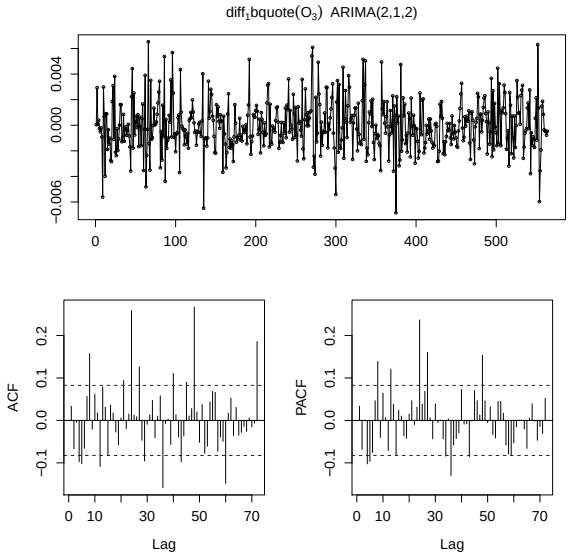

FIGURE 4.40: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

TABLE 4.11: Values of the RMSE, AIC, BIC and Ljung-box test of the residuals of every model.

	RMSE $[10^{-3}$ppm]	**MASE**	**Ljung-Box**	**p-value**
O_3-MA_{12}	1.63	0.520	27.8	0.01
O_3-MA_{24}	1.85	0.592	45.1	0.01
$diff_1(O_3)$	1.93	0.615	35.6	0.01

TABLE 4.12: Values of the RMSE, MAE and MASE for the forecasts produced by each model.

	RMSE $[10^{-2}$ppm]	**MAE** $[10^{-2}$ ppm]	**MASE**
O_3-MA_{12}	2.58	1.94	8.10
O_3-MA_{24}	1.47	1.14	4.74
$diff_1(O_3)$	1.29	1.07	4.46

4.5 PM_{10} - Particulate matter

After transforming the outlier free PM_{10} TS, see Figure 4.5, stationarity was tested. The full results are presented in Table4.13. The only transformations capable of making the TS stationary were $PM_{10} - MA_{12}$, $PM_{10} - MA_{24}$ and $diff_1(PM_{10})$. These were the ones used to estimate the ARIMA models parameters.

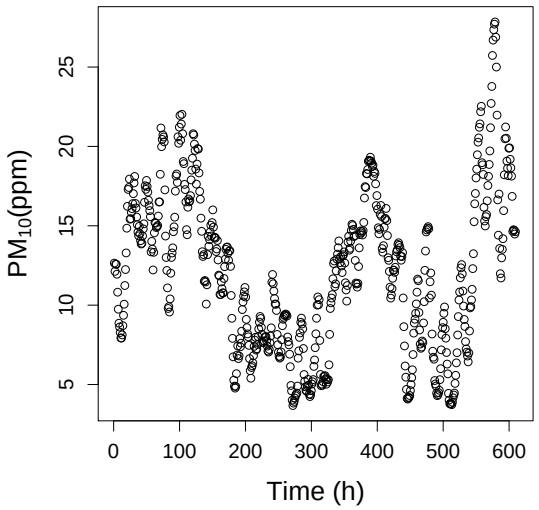

FIGURE 4.41: PM_{10} concentration plotted over time for the longest outlier-free data points, common to every pollutant.

TABLE 4.13: ADF test results and respective p-values for every transformation.

Transformation	**ADF test**	**p-value**
PM_{10}-MA_{24}	-6.58	0.01
PM_{10}-MA_{12}	-8.71	0.01
$\ln PM_10$	-2.31	0.45
$\sqrt{PM_{10}}$	-2.25	0.47
LOESS(PM_{10})	-2.05	0.56
$diff_1(PM_{10})$	-4.87	0.01

The best models, for each transformation, determined by the R algorithm were:

- ARIMA(3,0,3), for the $PM_{10} - MA_{12}$;
- ARIMA(2,0,3), for the $PM_{10} - MA_{24}$;
- ARIMA(5,1,1), for the $diff_1(PM_{10})$.

As can be seen in Table 4.14, the best performing model is ARIMA(3,0,3) with the lowest RMSE and MASE. It was also the only model that passed the Ljung-Box test, meaning the residuals are correlated, as can also be seen in Figure 4.43. The second best model for fitting is ARIMA(2,0,3) and this model has uncorrelated residuals according to the Ljung-Box test. In Figure 4.45, the ACF plot, left bottom panel, shows some almost significant lags up until lag 30. The PACF plot, bottom right panel in Figure 4.45, shows the same kind of behavior with significant lags only showing after lag 50. The worst performing model was ARIMA(5,1,1) although it also failed the Ljung-Box test. The Figure 4.47 shows significant peaks in the ACF and PACF plots, left and right bottom panels, around lag 20.

The forecast performances are shown in Table 4.15. The model ARIMA(3,0,3) had an almost completely useless forecast, that can be seen in Figure 4.42. ARIMA(2,0,3), see Figure 4.44,

was the best forecasting model and ARIMA(5,1,1), see Figure 4.46, performed the second best forecast.

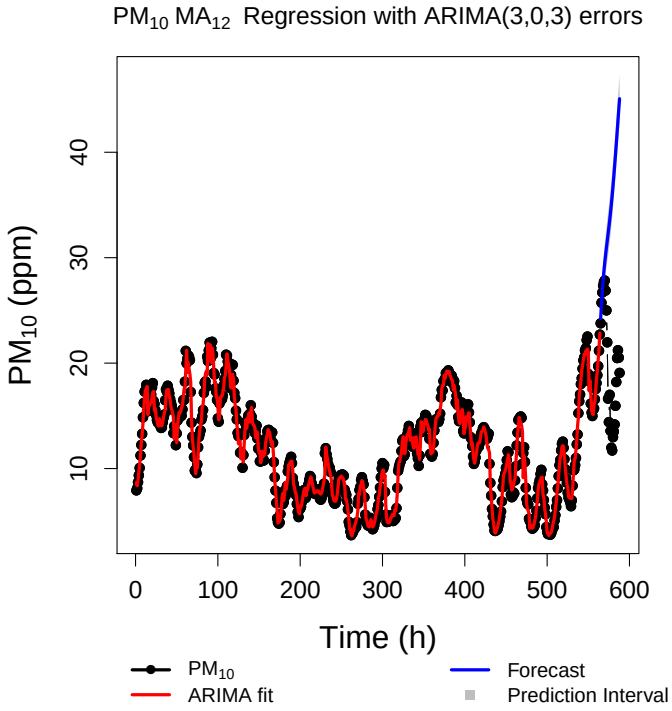

FIGURE 4.42: Best fitting model determined by R for the PM_{10}-MA_{12} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

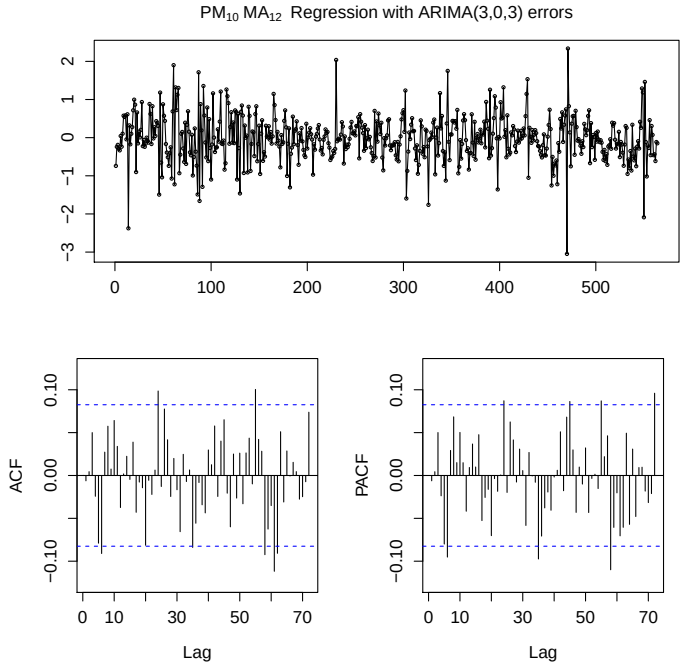

FIGURE 4.43: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

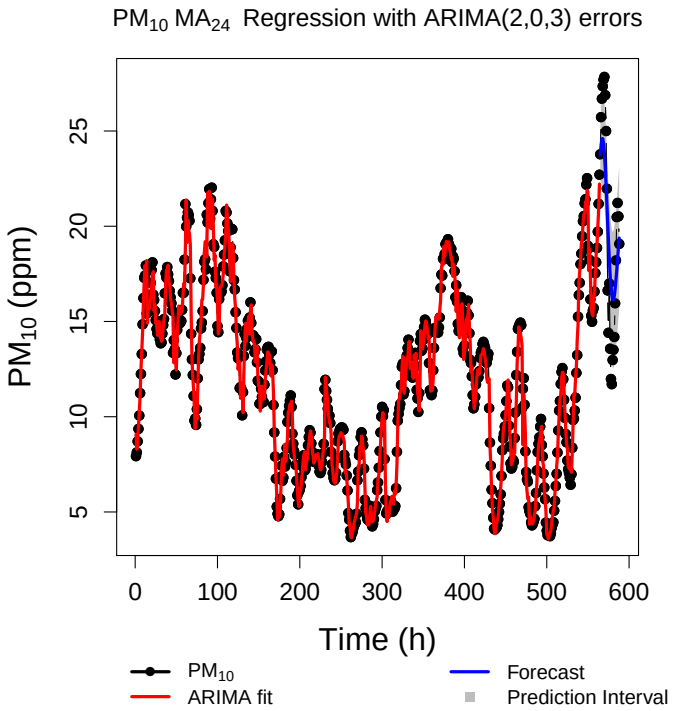

FIGURE 4.44: Best fitting model determined by R for the PM_{10}-MA_{24} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

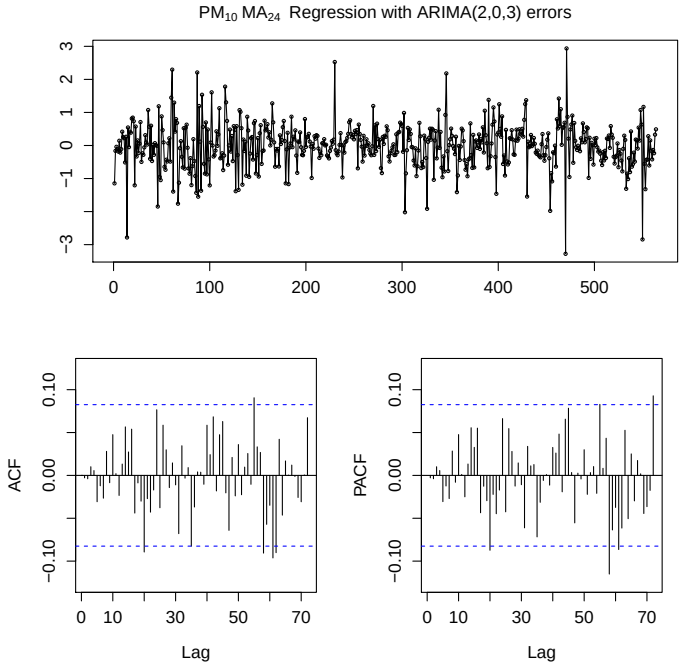

FIGURE 4.45: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

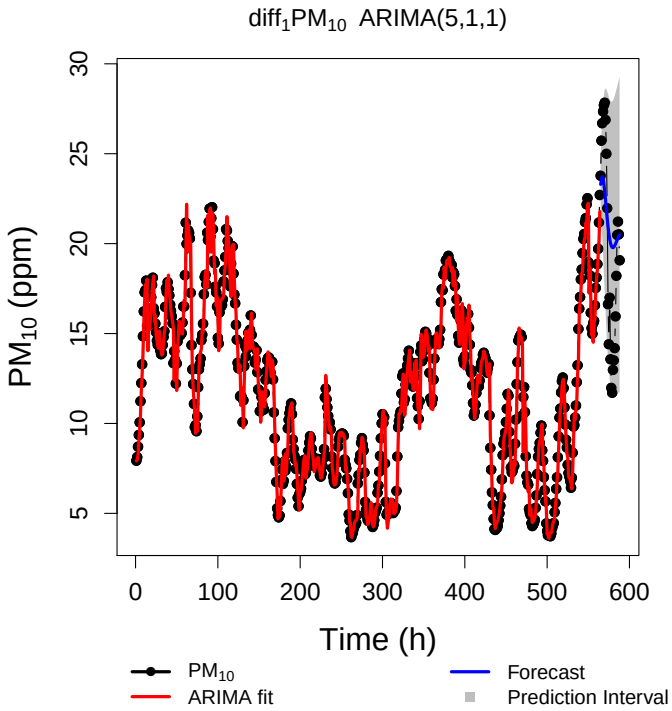

FIGURE 4.46: Best fitting model determined by R for the $\text{diff}_1(\text{PM}_{10})$ transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

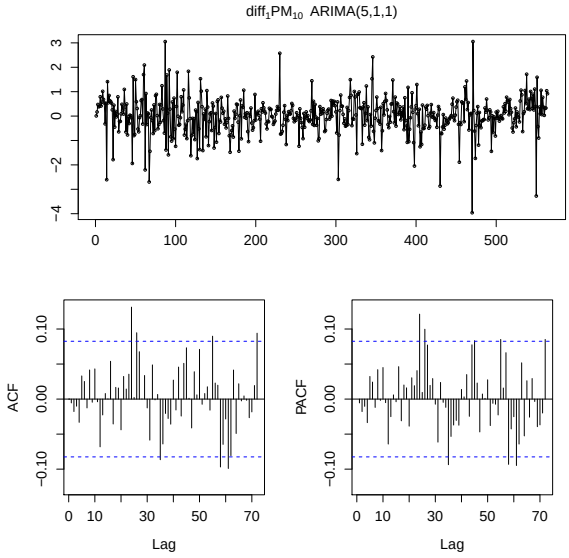

FIGURE 4.47: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

TABLE 4.14: Values of the RMSE, AIC, BIC and Ljung-box test of the residuals of every model.

	RMSE [ppm]	**MASE**	**Ljung-Box**	**p-value**
PM_{10}-MA_{12}	0.592	0.674	14.8	0.01
PM_{10}-MA_{24}	0.664	0.740	2.91	0.40
$\text{diff}_1(\text{PM}_{10})$	0.752	0.818	4.03	0.40

TABLE 4.15: Values of the RMSE, MAE and MASE for the forecasts produced by each model.

	RMSE [ppm]	**MAE** [ppm]	**MASE**
PM_{10}-MA_{12}	17.7	17.7	22.6
PM_{10}-MA_{24}	2.91	2.56	3.94
$diff_1(PM_{10})$	4.47	3.77	5.80

4.6 $PM_{2.5}$ - Particulate matter

After cutting the TS to 612 data points, see Figure 4.48, several transformations were tested in order to coerce the data to be stationary. After performing the ADF test, see Table 4.16, the only transformations that passed were $PM_{2.5} - MA_{12}$, $PM_{2.5} - MA_{24}$ and $\text{diff}_1(PM_{2.5})$. These were used to estimate the ARIMA model parameters.

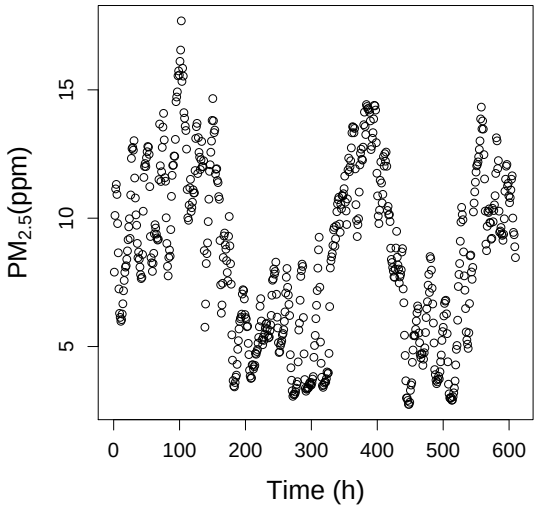

FIGURE 4.48: $PM_{2.5}$ concentration plotted over time for the longest outlier-free data points, common to every pollutant.

TABLE 4.16: ADF test results and respective p-values for every transformation.

Transformation	**ADF test**	**p-value**
$PM_{2.5}$-MA_{24}	-7.10	0.01
$PM_{2.5}$-MA_{12}	-9.19	0.01
$\ln PM_{2.5}$	-2.19	0.50
$\sqrt{PM_{2.5}}$	-2.10	0.54
LOESS($PM_{2.5}$)	-1.99	0.58
$\text{diff}_1(PM_{2.5})$	-4.98	0.01

The best models, for each transformation, determined by the R algorithm were:

- ARIMA(4,0,5), for the $PM_{2.5} - MA_{12}$;

- ARIMA(2,0,5), for the $PM_{2.5} - MA_{24}$;
- ARIMA(2,1,3), for the $\text{diff}_1(PM_{2.5})$.

The models ARIMA(4,0,5) and ARIMA(2,0,5), Figures 4.49 and 4.51 respectively, have very similar performances regarding the fit of the data. The worst performing model is ARIMA(2,1,3),see Figure 4.53. All the results are presented in Table 4.17. Every model passed the Ljung-Box test meaning there is still correlation in the residuals. In Figures 4.52 and 4.54 there are less significant peaks in the ACF and PACF plots, bottom left and right panels respectively, than in Figure 4.50. This accounts for the lower p-value in the Ljung-Box test for model ARIMA(4,0,5) compared to the other models.

For the forecast all of the models have similar performances with model ARIMA(4,0,5) being a slightly better as seen in Table 4.18.

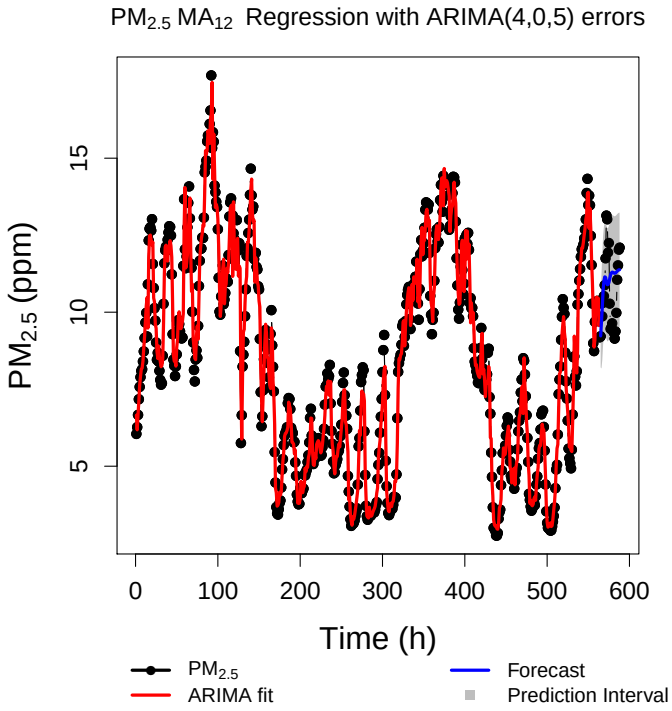

FIGURE 4.49: Best fitting model determined by R for the $PM_{2.5}$-MA_{12} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

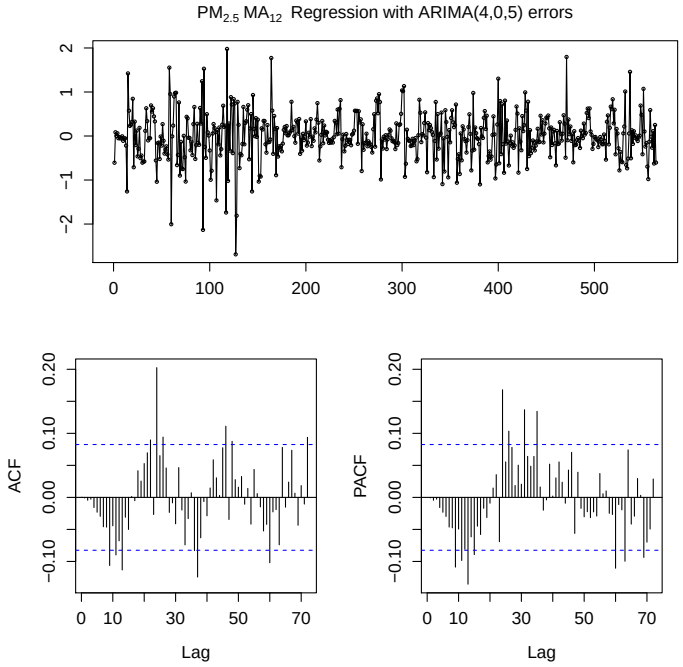

FIGURE 4.50: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

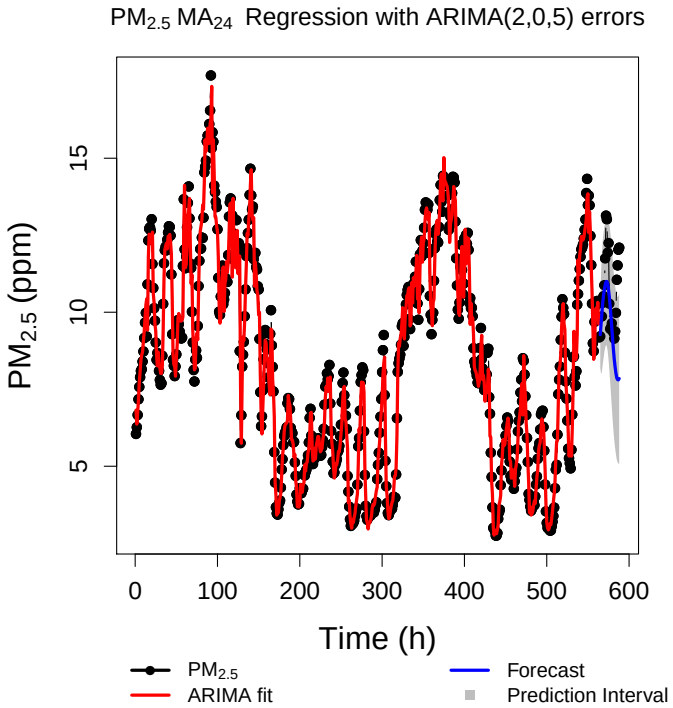

FIGURE 4.51: Best fitting model determined by R for the $PM_{2.5}$-MA_{24} transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

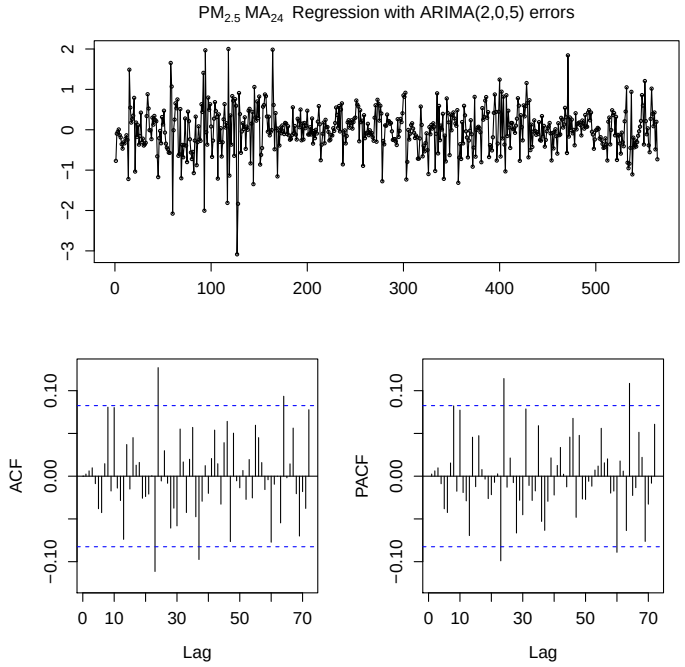

FIGURE 4.52: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

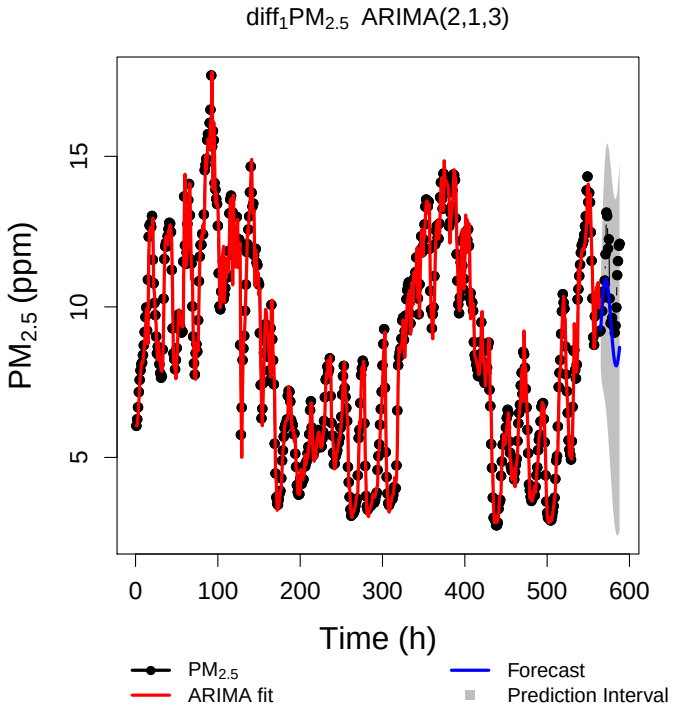

FIGURE 4.53: Best fitting model determined by R for the $\text{diff}_1(\text{PM}_{2.5})$ transformation in red, forecast of 72 hours in blue and the 95% prediction interval for the forecast.

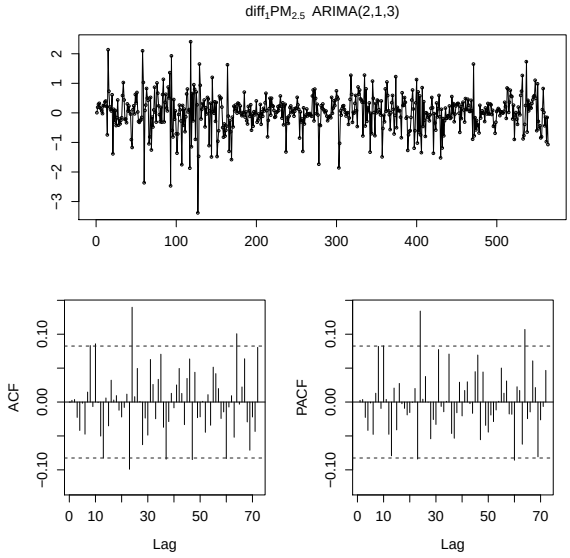

FIGURE 4.54: Plot of the model residuals as a function of time on the top. ACF and PACF plots of the residuals on the bottom right and bottom left, respectively.

TABLE 4.17: Values of the RMSE, AIC, BIC and Ljung-box test of the residuals of every model.

	RMSE [ppm]	**MASE**	**Ljung-Box**	**p-value**
$PM_{2.5}$-MA_{12}	0.517	0.724	26.3	0.01
$PM_{2.5}$-MA_{24}	0.541	0.770	10.3	0.02
$diff_1(PM_{2.5})$	0.607	0.824	11.0	0.05

TABLE 4.18: Values of the RMSE, MAE and MASE for the forecasts produced by each model.

	RMSE [ppm]	**MAE** [10^{-3} ppm]	**MASE**
$PM_{2.5}$-MA_{12}	1.31	1.10	2.15
$PM_{2.5}$-MA_{24}	1.82	1.25	2.45
$diff_1(PM_{2.5})$	1.71	1.27	2.48

4.7 Time Series Cross-Validation

In this section the results of cross-validation for every pollutant will be analyzed and compared. Only the MASE plots are presented here since it is the only error measure that does not depend on the scale. This means it is appropriate to compare results between pollutants. The cross-validation methodology performed is exactly identical to the one used for CO and consequently, the full procedure will not be explained here.

In the NO_2 rolling origin forecast with a fixed window, left panel of Figure 4.55, the less erratic behavior and lower error values come from $\text{diff}_1(NO_2)$, which was also the best performing model in Table 4.6. This carried on when a increasing window was used, center panel of Figure 4.55, and when the performance with increasing forecast length was evaluated, right panel in Figure 4.55.

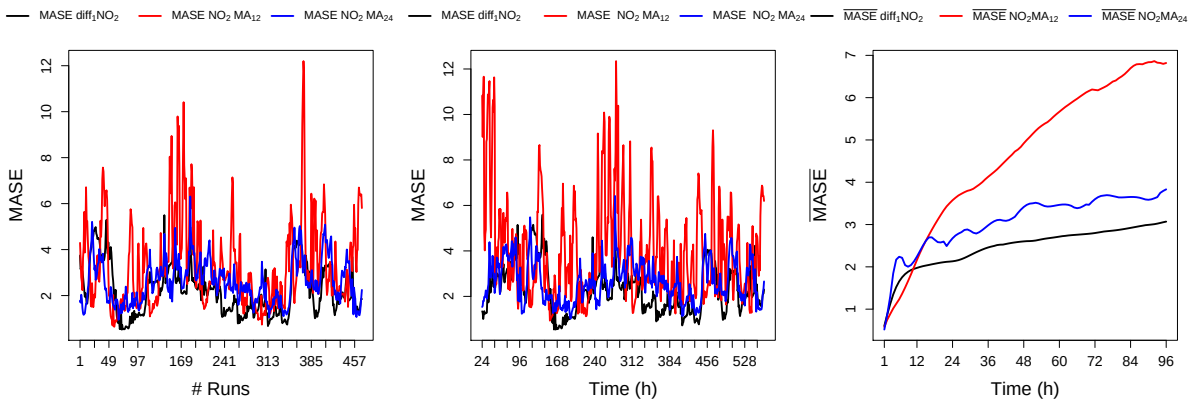

FIGURE 4.55: *Left:* MASE of different models plotted as a function of fixed windows. *Center:* MASE of different models plotted as a function of data points used by the model. *Right:* Average MASE of different models plotted as a function of the forecast length.

For the SO_2 cross-validation the results from the rolling forecast with fixed window, left panel in Figure 4.56, and the rolling forecast with increasing window, center panel in Figure 4.56, show that the models for $SO_2 - MA_{24}$ and $\text{diff}_1(SO_2)$ do not differ much. Only in the right panel of Figure 4.56, the average MASE as a function of the forecast length, it is possible to see that for forecasts length 24 hours the model $SO_2 - MA_{24}$ outperforms $\text{diff}_1(SO_2)$, which was the best performing model in Table 4.9. There is also the possibility that the error for $SO_2 - MA_{24}$ is being underestimated , since the model residuals were correlated.

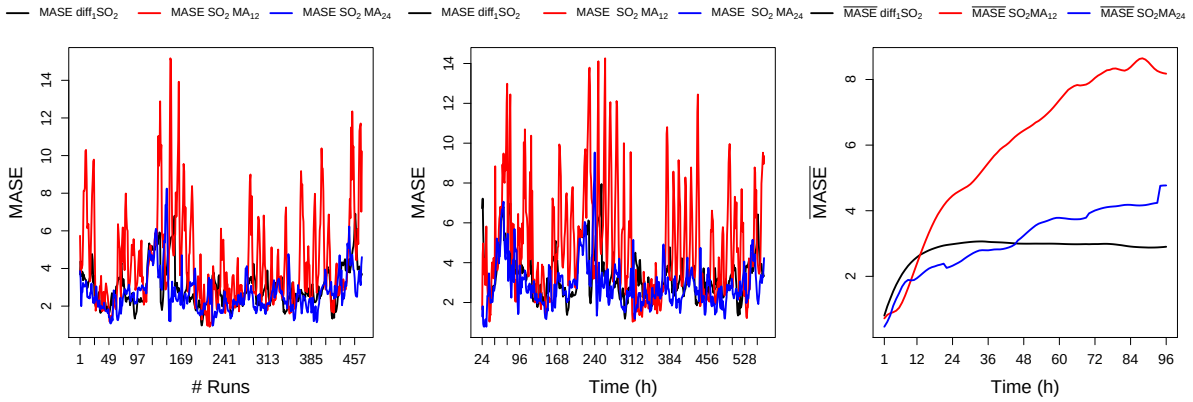

FIGURE 4.56: *Left:* MASE of different models plotted as a function of fixed windows. *Center:* MASE of different models plotted as a function of data points used by the model. *Right:* Average MASE of different models plotted as a function of the forecast length.

The models obtain for O_3 had very correlated residuals so the expectations for forecast performance were low. In the left panel of Figure 4.57, the forecast performance with a fixed window seems to be better for $O_3 - MA_{24}$. The forecast with increasing window, center panel of Figure 4.57, showed the same performance. Between 384 and 456 data points used by the model the MASE values of $O_3 - MA_{24}$ keep constant, where for the other two models increase. When the forecast length was varied, right panel in Figure 4.57, the results are very interesting. The

errors for $O_3 - MA_{24}$ and $O_3 - MA_{12}$ are very irregular. For forecasts with length of 24 data points $O_3 - MA_{24}$ has a lower error value.

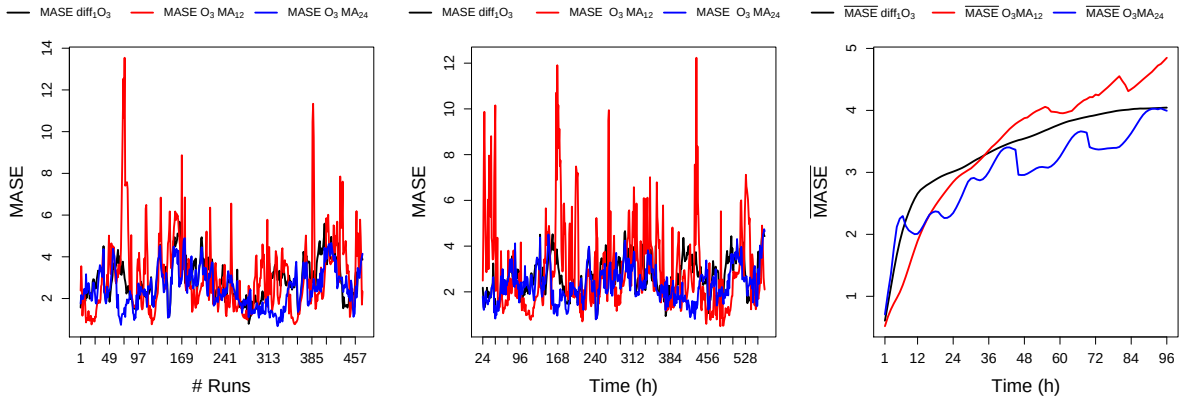

FIGURE 4.57: *Left:* MASE of different models plotted as a function of fixed windows. *Center:* MASE of different models plotted as a function of data points used by the model. *Right:* Average MASE of different models plotted as a function of the forecast length.

The MASE values in the left and center panels in Figure 4.58 show that the previously best performing model in Table 4.15, $PM_{2.5} - MA_{24}$ was also the best performing model in cross-validation since it had the lowest error values. This performance remained the same when the forecast length varied, in Figure 4.58.

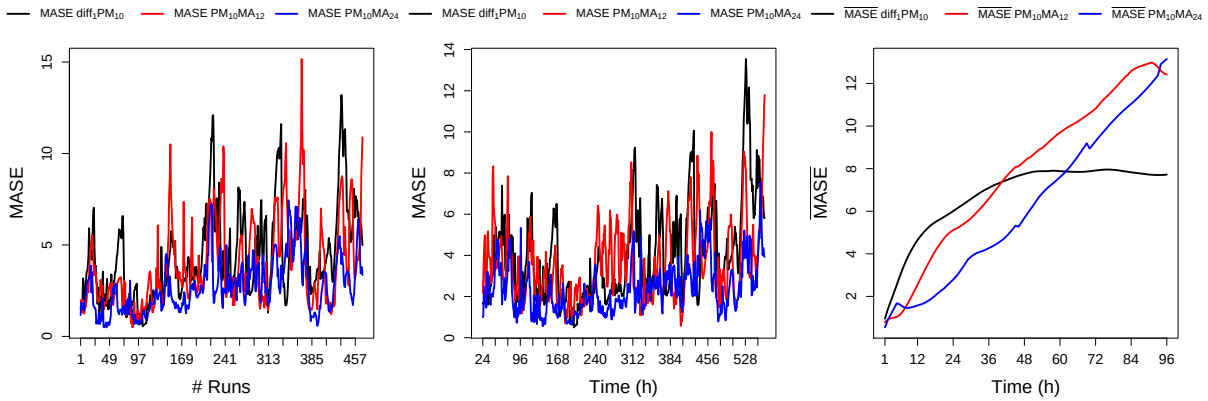

FIGURE 4.58: *Left:* MASE of different models plotted as a function of fixed windows. *Center:* MASE of different models plotted as a function of data points used by the model. *Right:* Average MASE of different models plotted as a function of the forecast length.

The results of the previous forecasts in Table 4.18 showed that all the models had very close performances with the one for $PM_{2.5} - MA_{12}$ performing slightly better. The cross-validation showed that actually the model for $PM_{2.5} - MA_{24}$ was the better for forecast. The left and center panels in Figure 4.59 show smaller values of MASE for $PM_{2.5} - MA_{24}$ than for both $PM_{2.5} - MA_{12}$ and $\text{diff}_1(PM_{2.5})$, although all the models show an erratic behavior. For the performance with increasing forecast length, once again the model for $PM_{2.5} - MA_{24}$ outperforms the other models and with an average MASE lower that the other models with a 24 data points forecast.

In the previous analysis concludes that the best results were obtained for $PM_{10} - MA_{24}$. It was the only model with uncorrelated residuals that had a MASE between 2 and 3. The MASE for $diff_1(NO_2)$ is lower, but the model had correlated residuals so this could be an underestimated value.

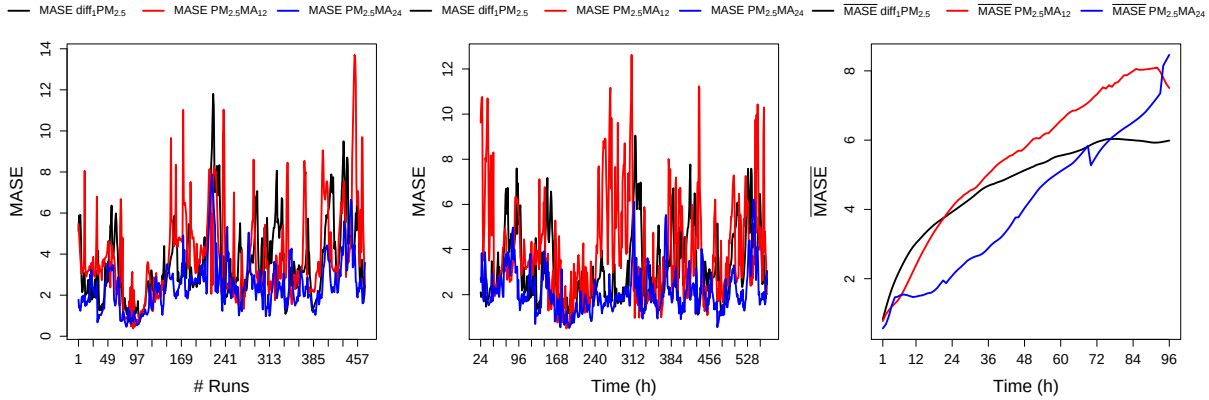

FIGURE 4.59: *Left:* MASE of different models plotted as a function of fixed windows. *Center:* MASE of different models plotted as a function of data points used by the model. *Right:* Average MASE of different models plotted as a function of the forecast length.

It was unexpected that the R algorithm did not pick any seasonal models but possibly there was not enough data to do so. Indeed a good fit does not equate to a good forecast. It is difficult to compare results to other similar studies since the data sets used are different and not available to the public. Also the decision of using MASE and not median absolute percentage error, which is a more popular error measurement, made it impossible to find comparable results. This type of measurement does not depend on scale but can lead to bias when the data values are close to zero so was excluded from the analysis.

Our results were compared to the results of the M3-Competition, a forecasting competition where different methods are applied to several TS. The best performing methods had MASE values between 2 and 3 [34, 43], of which was achieved in this work. But it would be unwise to claim that the models in this work perform as well. Because of the correlation in the residuals of many models, the errors could be underestimated. Overall the results were good, considering the modeling conditions were not ideal however, there is definitely room for improvement.

Chapter 5

Conclusions and Further Work

In the present work the Box-Jenkins ARIMA methodology was studied in order to produce models capable of good forecasts for several pollutants. After analyzing the data, it was detected the existence of extreme outliers in every TS. Since they imparted such acute changes on the mean values of the data and spanned across many data points it was chosen to not deal with the outliers and simply to cut down all the series to a common size were no outliers were detected. This meant a trade-off between the number of data points available and a easier to forecast series. The choice was made and the resulting TS re-analyzed. The TS were not stationary and so, several transformations were used to try and stabilize the data. In every pollutant only three transformation produced stationarity: subtraction of a moving average of orders 12 and 24 and first order differentiation. It was also expected that LOESS produced satisfactory results but it was not the case.

After obtaining the stationary series the ARIMA models parameters were estimated. This was done by a R algorithm. It iteratively considered several ARIMA models of different orders, optimized the coefficients and then chose the best ones according to the AIC. The models residuals where analyzed for autocorrelation and many of them still had correlation between them. This meant there was information in the data that the models did not capture. These models were also used for an initial forecast of 24 hours and their performance measured with RMSE, MAE and MASE.

Cross-validation was the next step. Since ARIMA models depend on past values they sometimes produce a good forecast merely because the future data resembles the past data and they are not robust. To assess this, multiple methods of forecast with a rolling origin were used: fixed window, increasing window and a fixed window with an increase in forecast length. The best models showed a less volatile behavior with the fixed and increasing window along with the smaller error values in all the methods. For the fixed window with increasing forecast length the best model was considered the one with lower error value at 24 data points. This was the forecast length that interested us for this work. It was interesting to see how the error evolved with the forecast length since the models changed behavior. The best model depended on the length of the forecast considered.

The results produced were overall good since a forecast of 24 data points is reasonably long, considering the results from M3-competition [34, 43]. Still, the forecasts errors might be underestimated due to autocorrelation in the models residuals so, there is much room for improvement. Finding real data without outliers is practically impossible so in future work

it would be necessary either to, find a method flexible enough to be able to forecast such a change in the data or, a way to deal with the outliers. In addition, this work was made resorting only to information about pollutant concentrations but there are several explanatory variables that could have been taken into consideration. Data like temperature and UV index would be particularly important. Also more elaborate smoothing techniques that were not considered could be used.

Finally the complexity of the model itself could be increased. A dynamic harmonic regression could be implemented were the seasonal pattern is modeled by Fourier terms, or multiple seasonality patterns could be defined for the data. All these are possible options for improvement, even before reaching for truly more complex models like ANN or Fuzzy Logic.

In conclusion, the world of forecasting is vast and the options to perform the same task are endless. Ultimately it is the scientist responsibility to make decisions based on knowledge and experience. This work delved only on the tip of the iceberg and recognizes that there is much more to learn and to experiment with.

Appendix A

Intermediate Plots

A.1 NO_2 - Nitrogen Dioxide

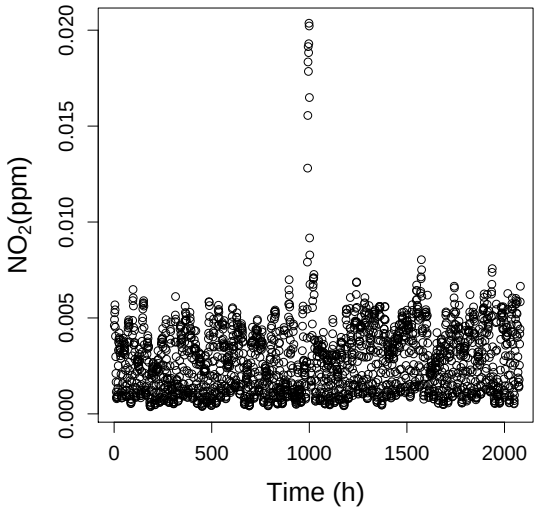

FIGURE A.1: NO_2 concentration plotted over time for the complete data.

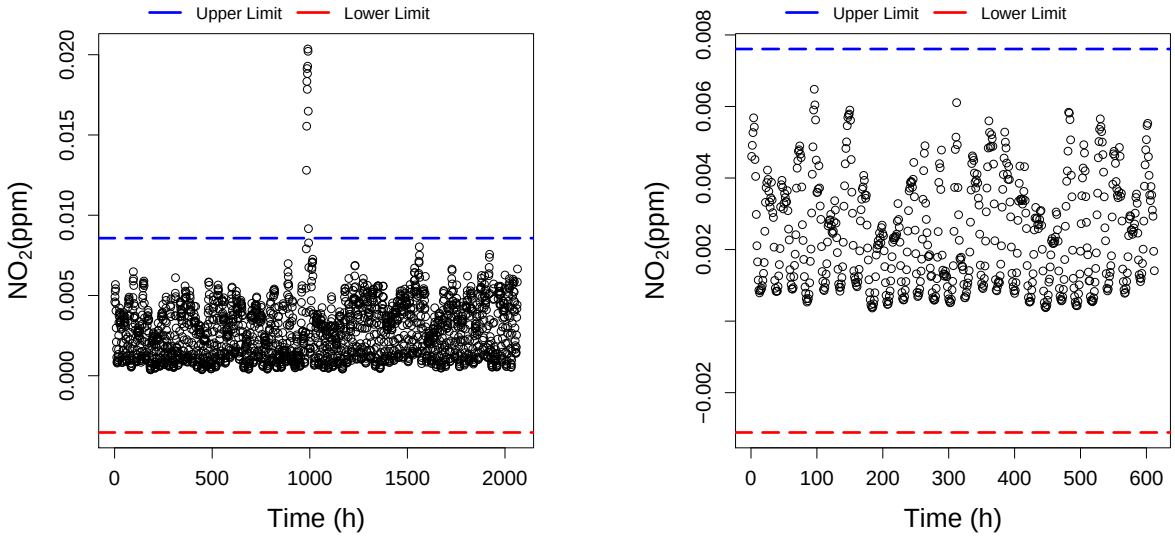

FIGURE A.2: *Left:* NO_2 concentration plotted over time for the complete data and MAD. The upper limit in blue and the lower limit in red. *Right:* The longest outlier-free data points for the first 612 data points and MAD. The upper limit in blue and the lower limit in red.

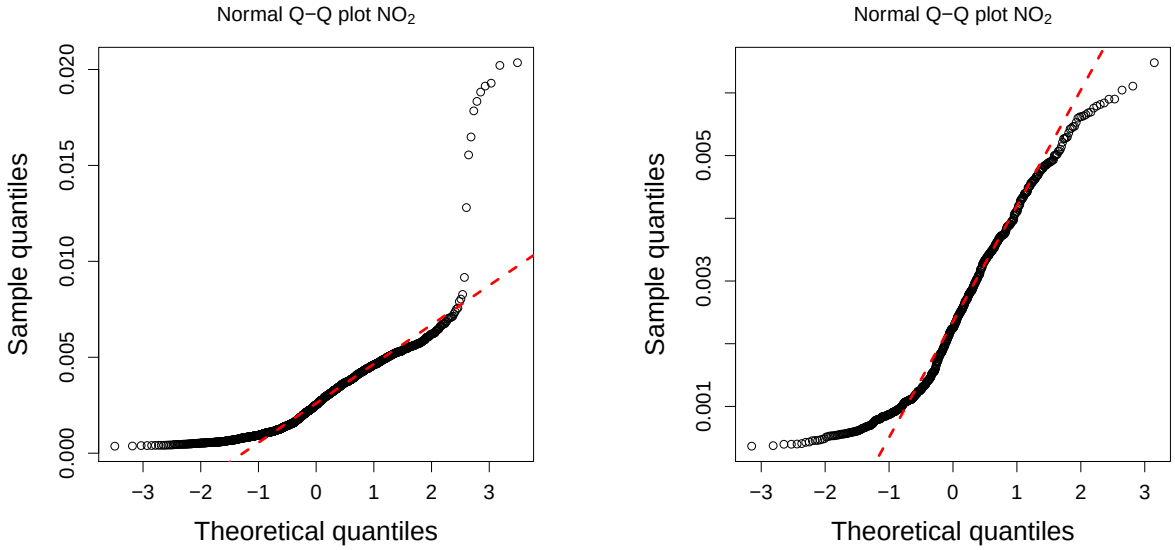

FIGURE A.3: *Left:* QQ plot for the complete data. *Right:* QQ plot for the longest outlier-free data points.

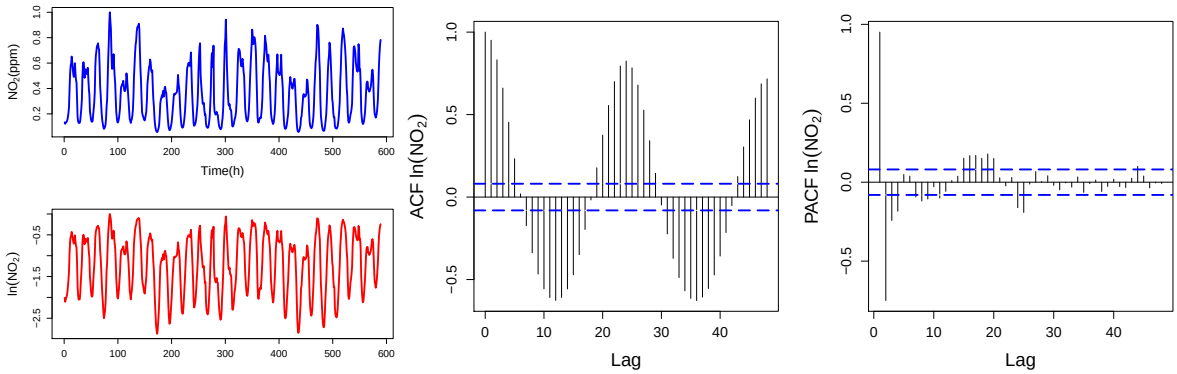

FIGURE A.4: *Left:* Normalized NO_2 compared to $\ln(NO_2)$ plotted over time. *Center:* ACF plotted over lags for $\ln(NO_2)$. *Right:* PACF plotted over lags for $\ln(NO_2)$.

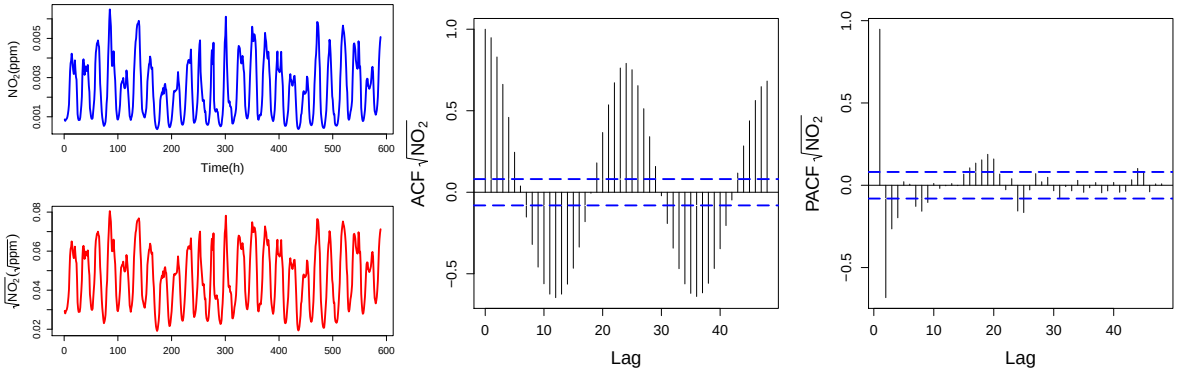

FIGURE A.5: *Left:* NO_2 concentrations compared to $\sqrt{NO_2}$ plotted over time. *Center:* ACF plotted over lags for $\sqrt{NO_2}$. *Right:* PACF plotted over lags for $\sqrt{NO_2}$.

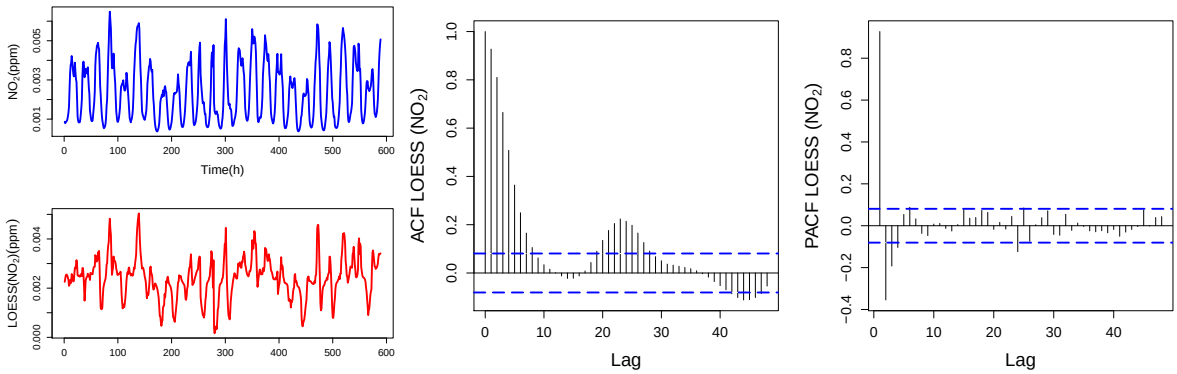

FIGURE A.6: *Left:* NO_2 concentrations compared to LOESS(NO_2) plotted over time. *Center:* ACF plotted over lags for LOESS(NO_2). *Right:* PACF plotted over lags for LOESS(NO_2).

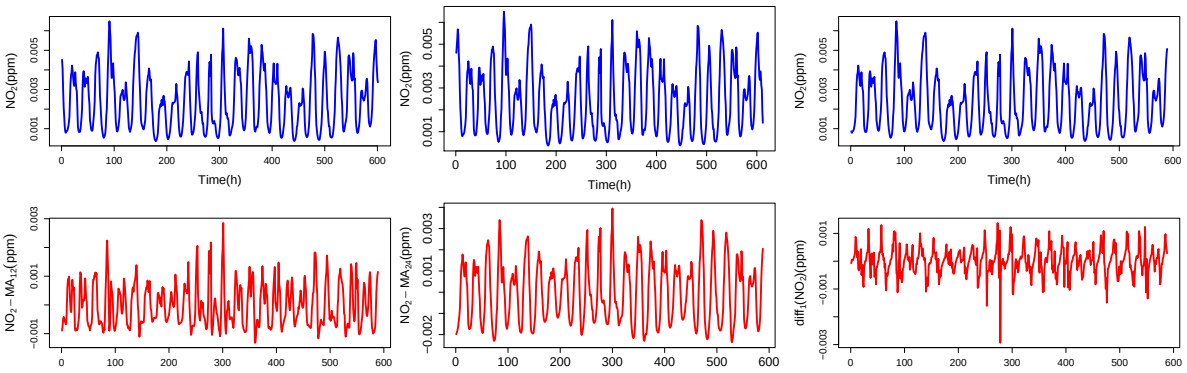

FIGURE A.7: *Left:* NO_2 concentrations compared to NO_2-MA_{12} plotted over time. *Center:* NO_2 concentrations compared to NO_2-MA_{24} plotted over time. *Right:* NO_2 concentrations compared to $\text{diff}_1(NO_2)$ plotted over time.

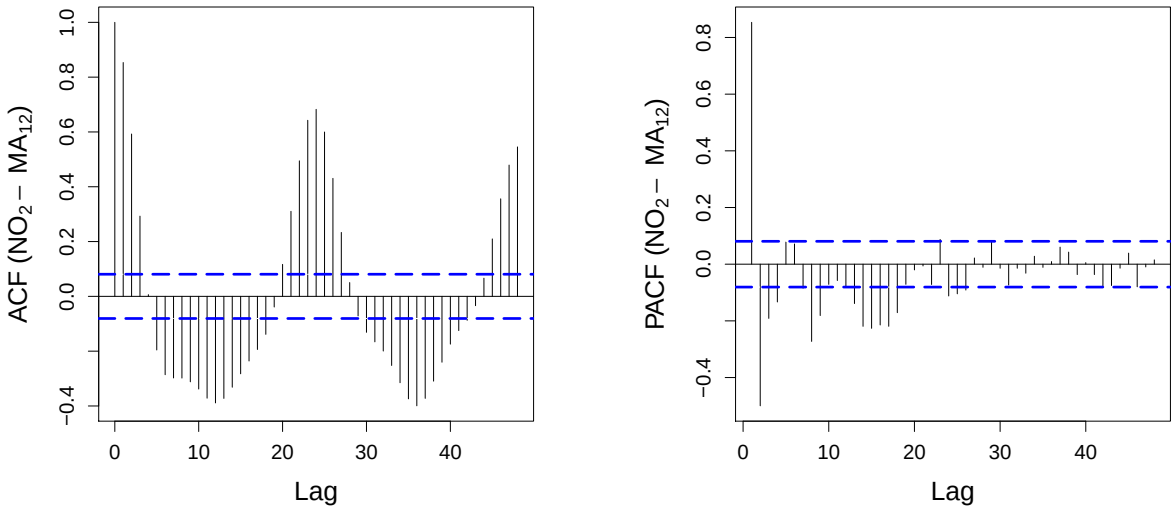

FIGURE A.8: ACF and PACF plots for the NO_2-MA_{12} transformation.

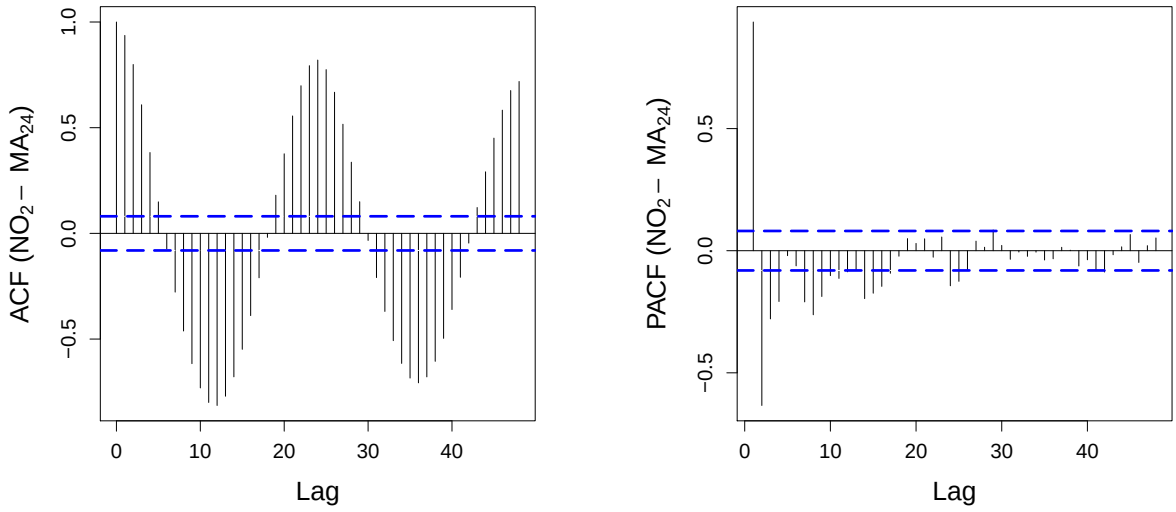

FIGURE A.9: ACF and PACF plots for the NO_2-MA_{24} transformation.

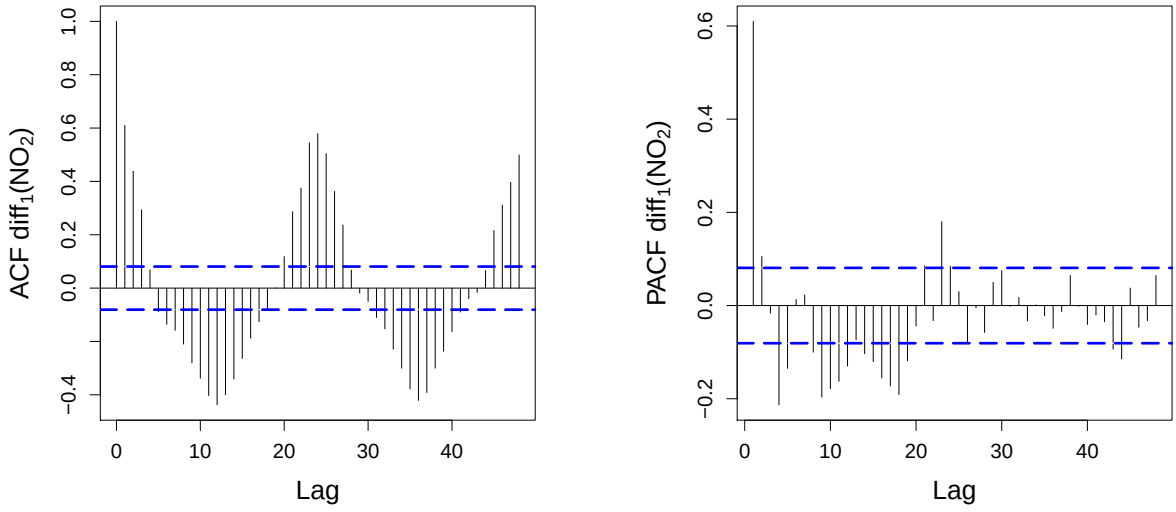

FIGURE A.10: ACF and PACF plots for the diff$_1$(NO_2) transformation.

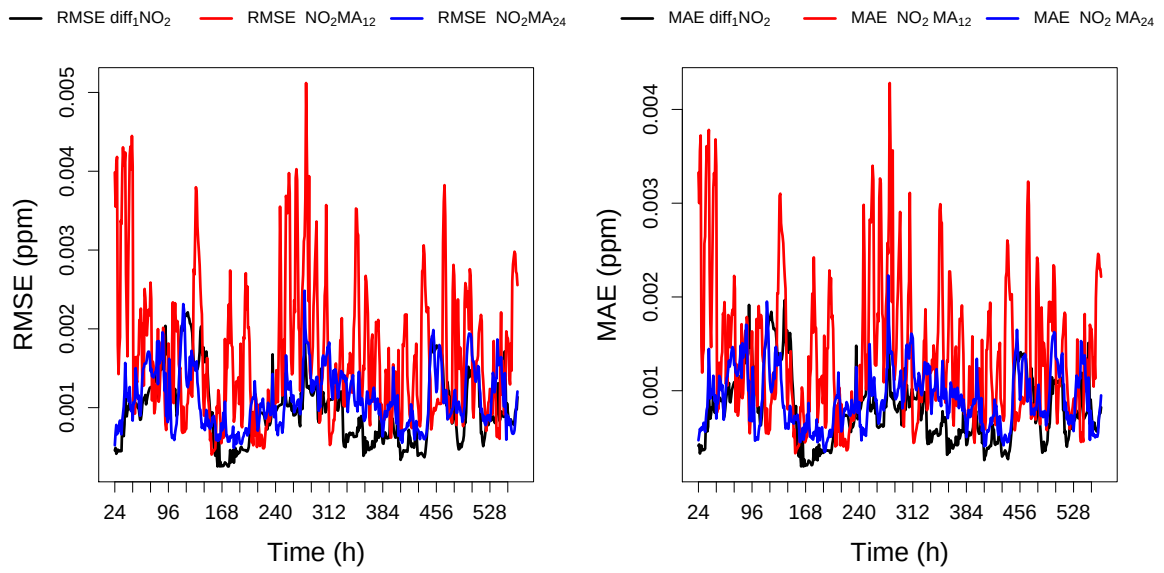

FIGURE A.11: *Left:* RMSE of different models plotted as a function of data points used by the model. *Rigth:* MAE of different models plotted as a function of data points used by the model.

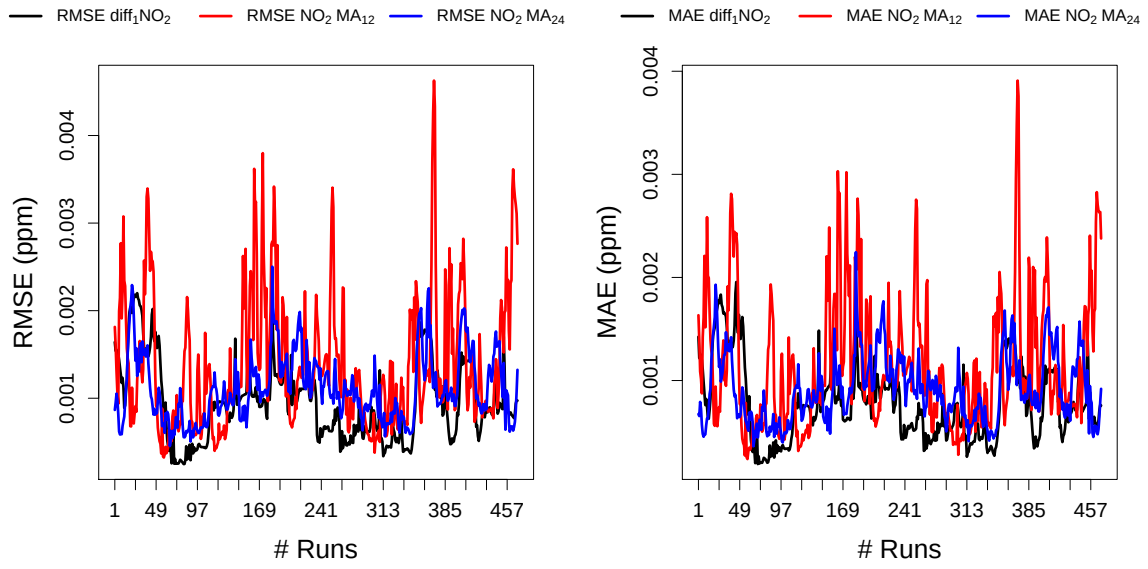

FIGURE A.12: *Left:* RMSE of different models plotted as a function of the number of fixed windows. *Rigth:* MAE of different models plotted as a function of fixed windows.

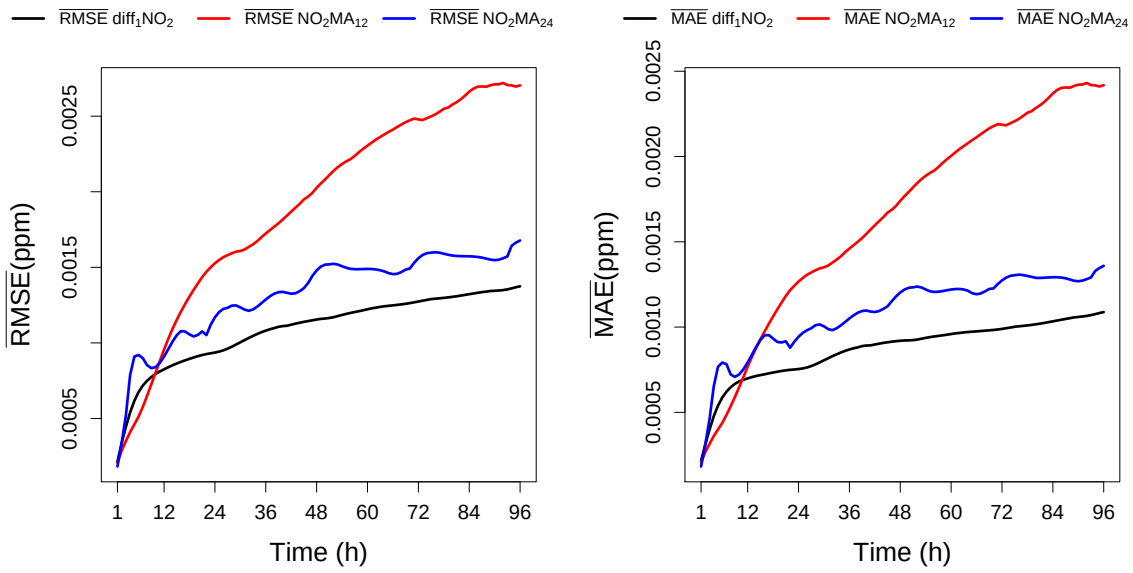

FIGURE A.13: *Left:* Average RMSE of different models plotted as a function of the forecast length. *Rigth:* Average MAE of different models plotted as a function of the forecast length.

A.2 SO_2 - Sulfur Dioxide

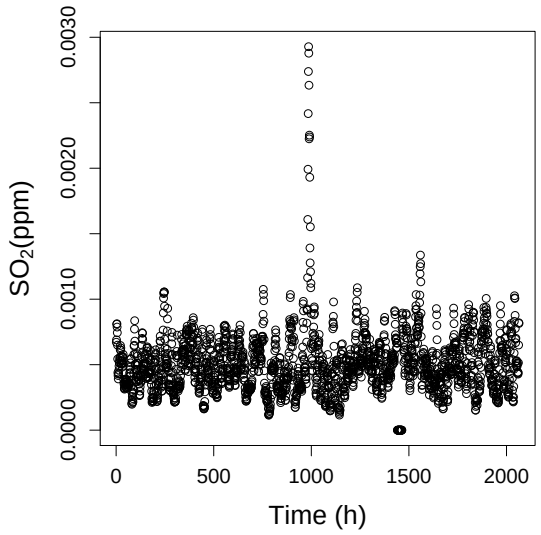

FIGURE A.14: SO_2 concentration plotted over time for the complete data.

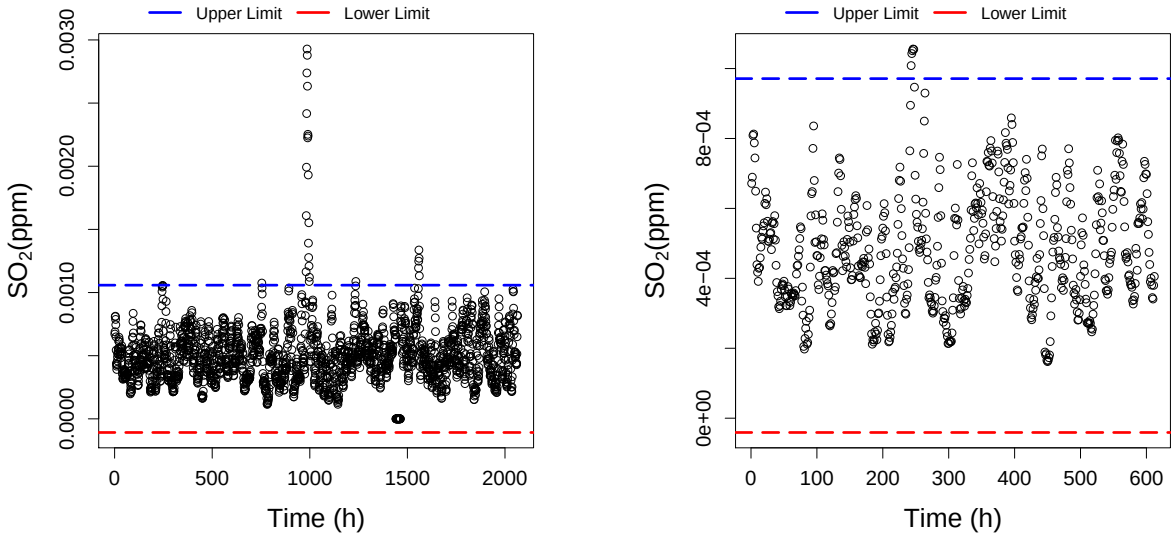

FIGURE A.15: *Left:* SO_2 concentration plotted over time for the complete data and MAD. The upper limit in blue and the lower limit in red. *Right:* The longest outlier-free data points for the first 612 data points and MAD. The upper limit in blue and the lower limit in red.

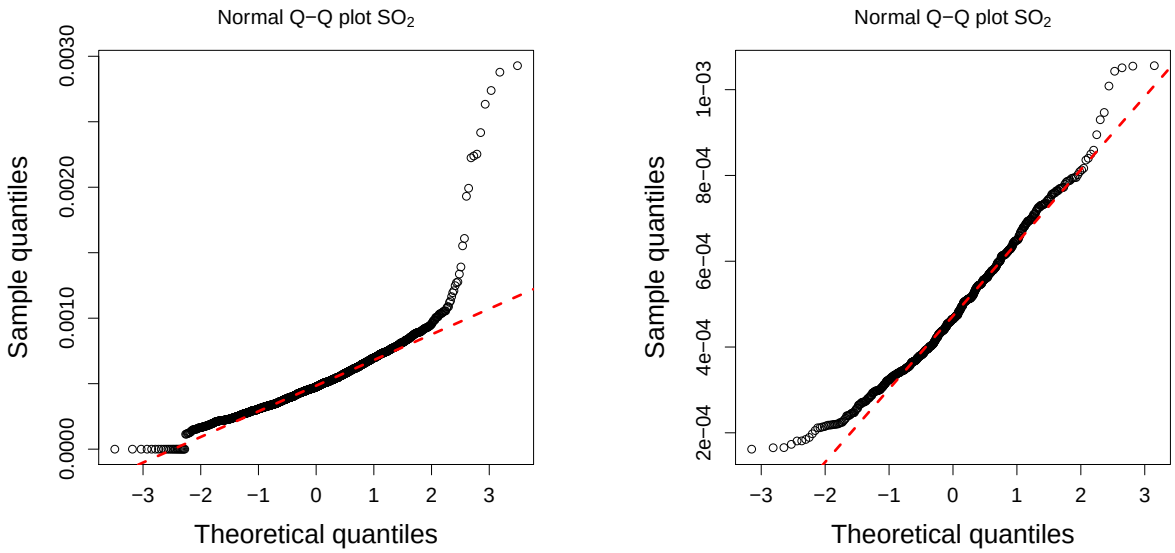

FIGURE A.16: *Left:* QQ plot for the complete data. *Right:* QQ plot for the longest outlier-free data points.

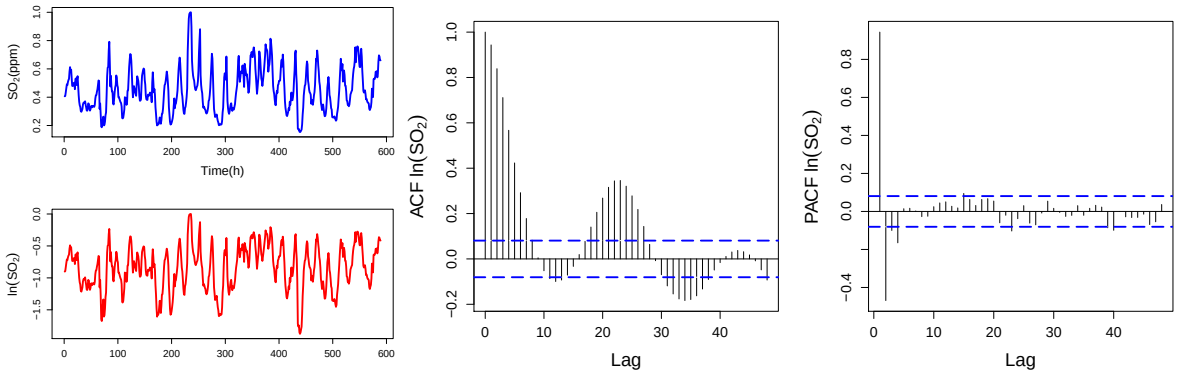

FIGURE A.17: *Left:* Normalized SO_2 compared to $\ln(SO_2)$ plotted over time. *Center:* ACF plotted over lags for $\ln(SO_2)$. *Right:* PACF plotted over lags for $\ln(SO_2)$.

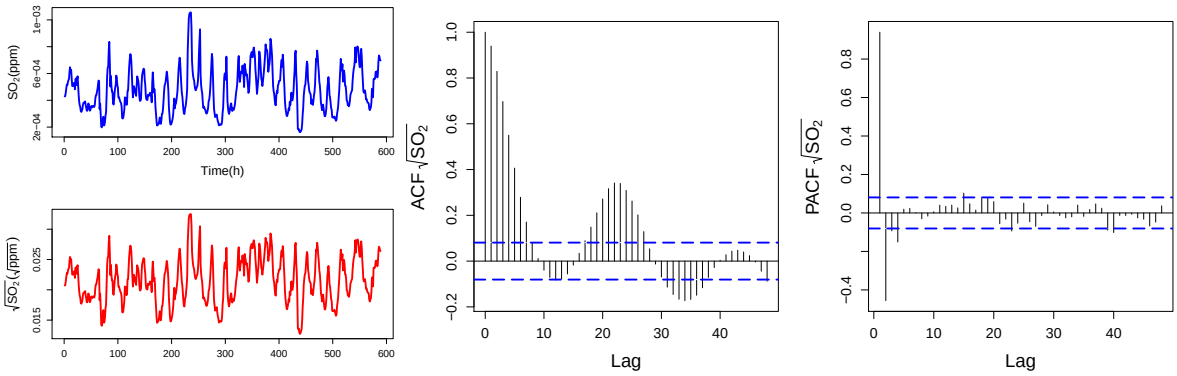

FIGURE A.18: *Left:* SO_2 concentrations compared to $\sqrt{SO_2}$ plotted over time. *Center:* ACF plotted over lags for $\sqrt{SO_2}$. *Right:* PACF plotted over lags for $\sqrt{SO_2}$.

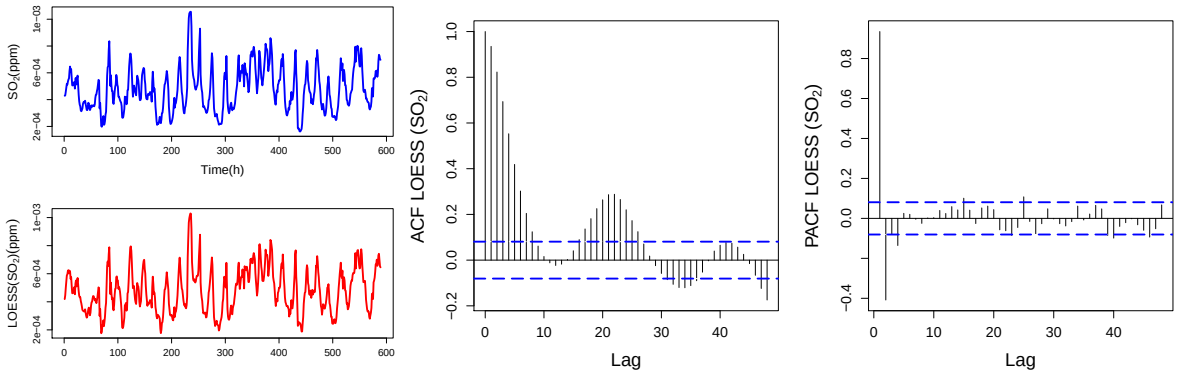

FIGURE A.19: *Left:* SO_2 concentrations compared to LOESS(SO_2) plotted over time. *Center:* ACF plotted over lags for LOESS(SO_2). *Right:* PACF plotted over lags for LOESS(SO_2).

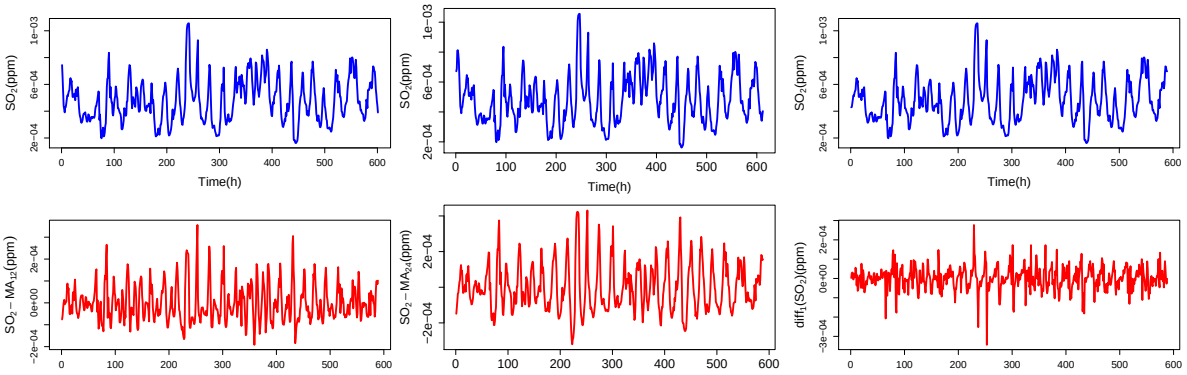

FIGURE A.20: *Left:* SO_2 concentrations compared to SO_2-MA_{12} plotted over time. *Center:* SO_2 concentrations compared to SO_2-MA_{24} plotted over time. *Right:* SO_2 concentrations compared to $\text{diff}_1(SO_2)$ plotted over time.

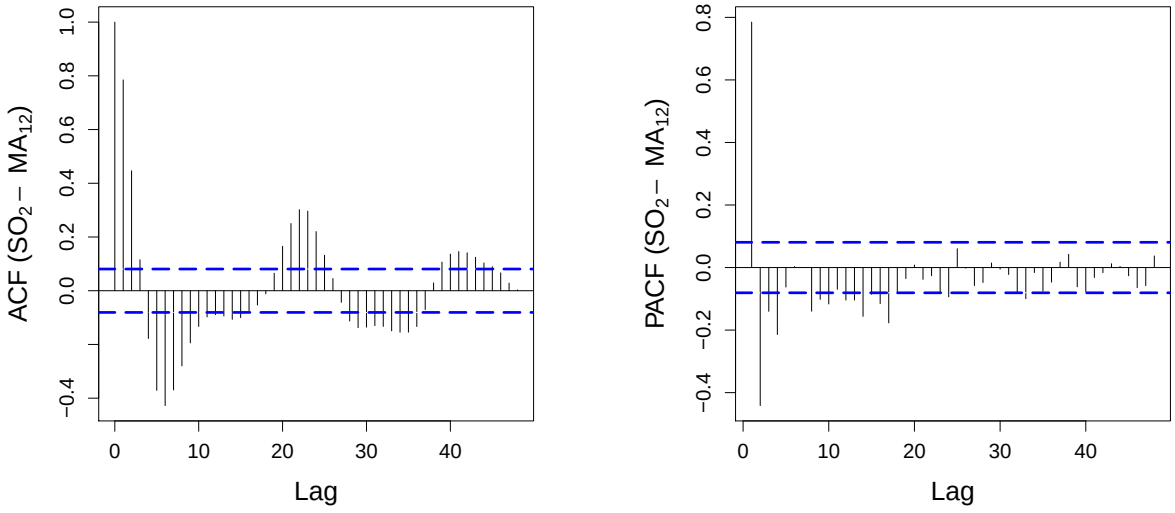

FIGURE A.21: ACF and PACF plots for the SO_2-MA_{12} transformation.

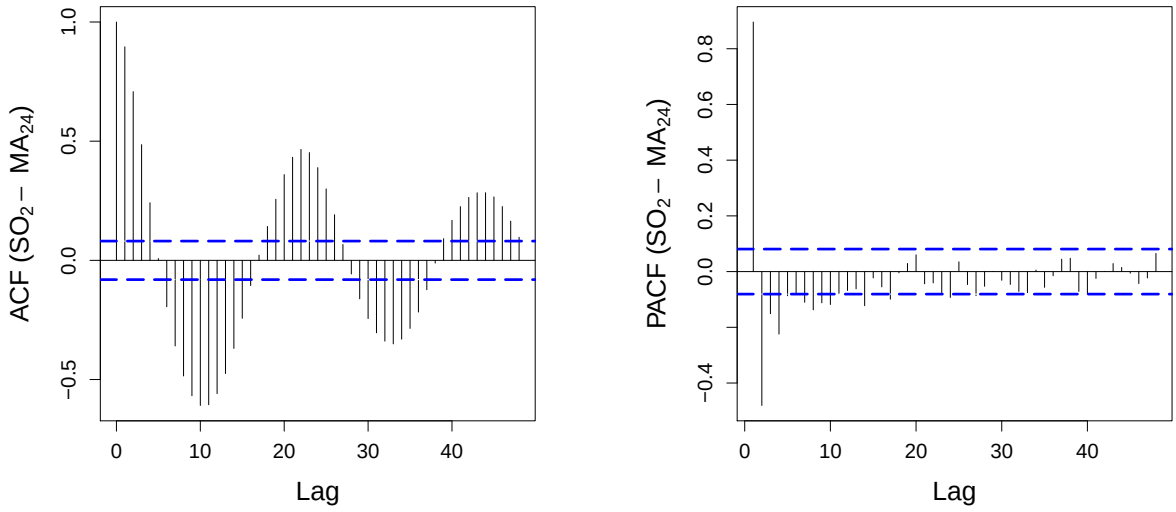

FIGURE A.22: ACF and PACF plots for the SO_2-MA_{24} transformation.

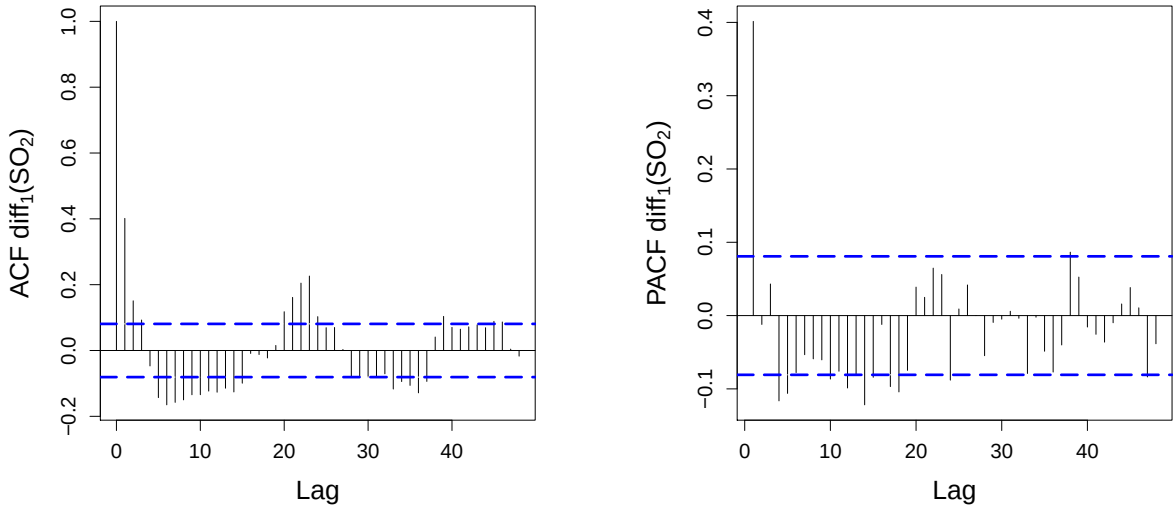

FIGURE A.23: ACF and PACF plots for the $diff_1(SO_2)$ transformation.

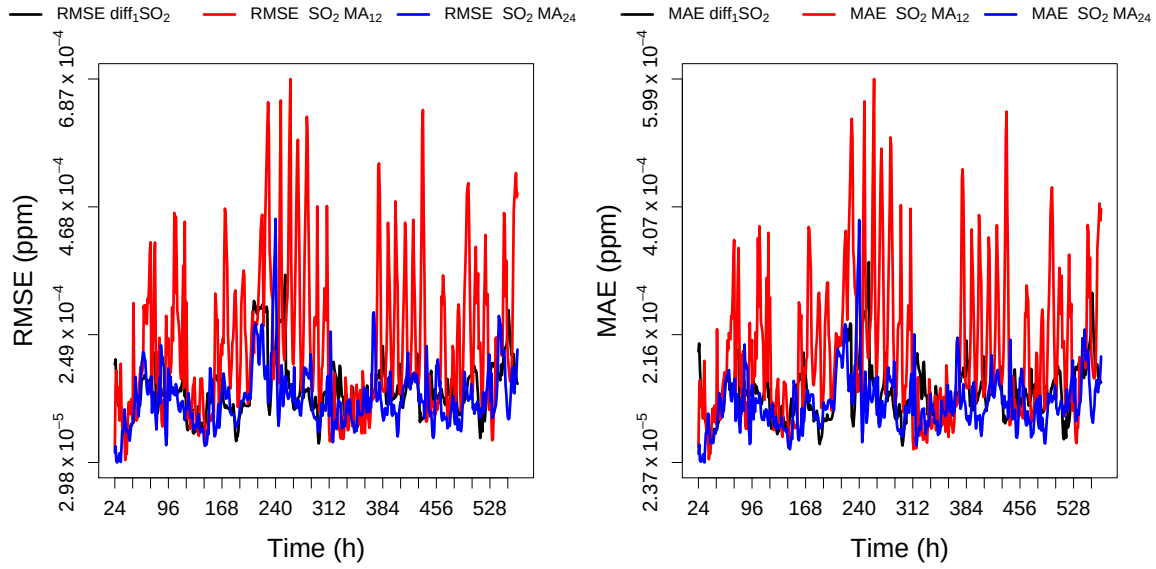

FIGURE A.24: *Left:* RMSE of different models plotted as a function of data points used by the model. *Rigth:* MAE of different models plotted as a function of data points used by the model.

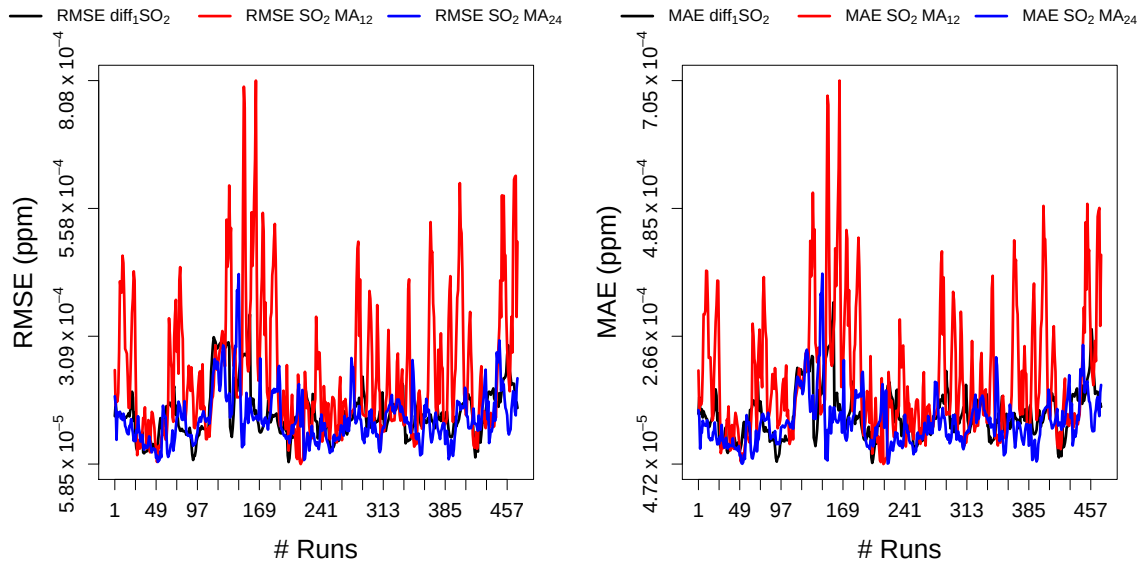

FIGURE A.25: *Left:* RMSE of different models plotted as a function of the number of fixed windows. *Rigth:* MAE of different models plotted as a function of fixed windows.

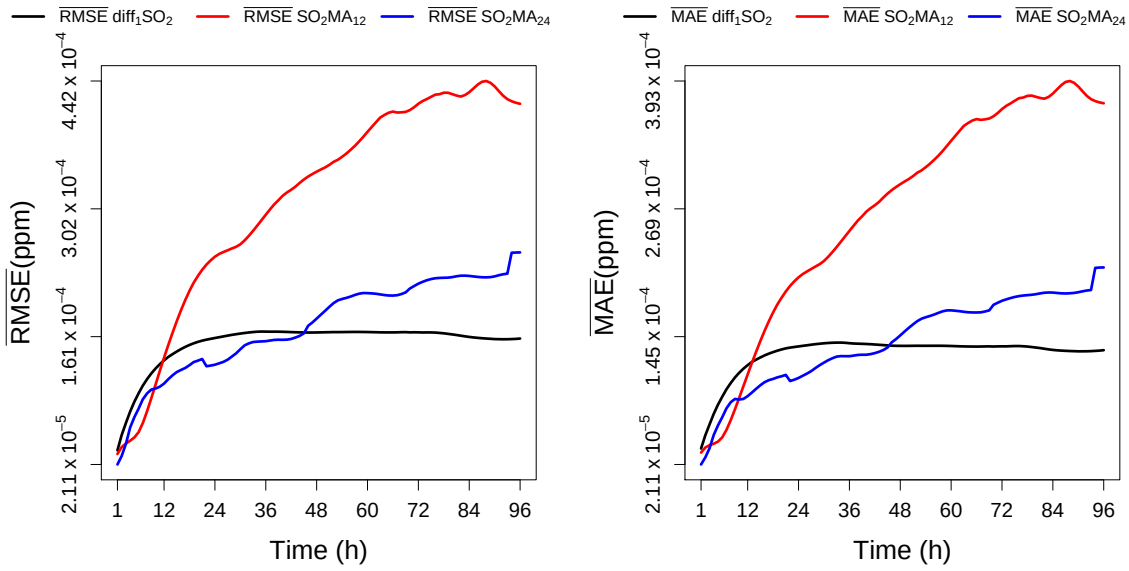

FIGURE A.26: *Left:* Average RMSE of different models plotted as a function of the forecast length. *Rigth:* Average MAE of different models plotted as a function of the forecast length.

A.3 O_3 - Ozone

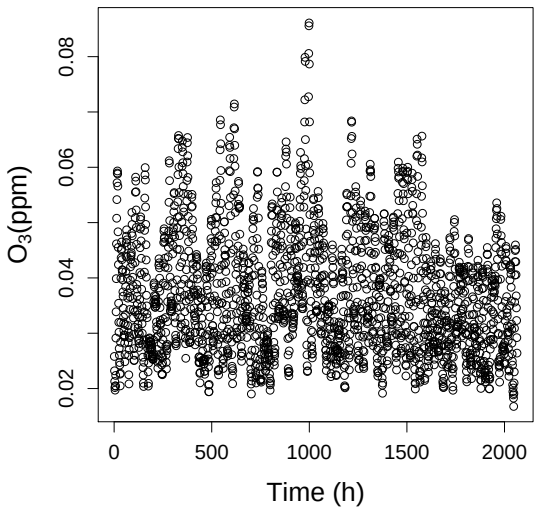

FIGURE A.27: O_3 concentration plotted over time for the complete data.

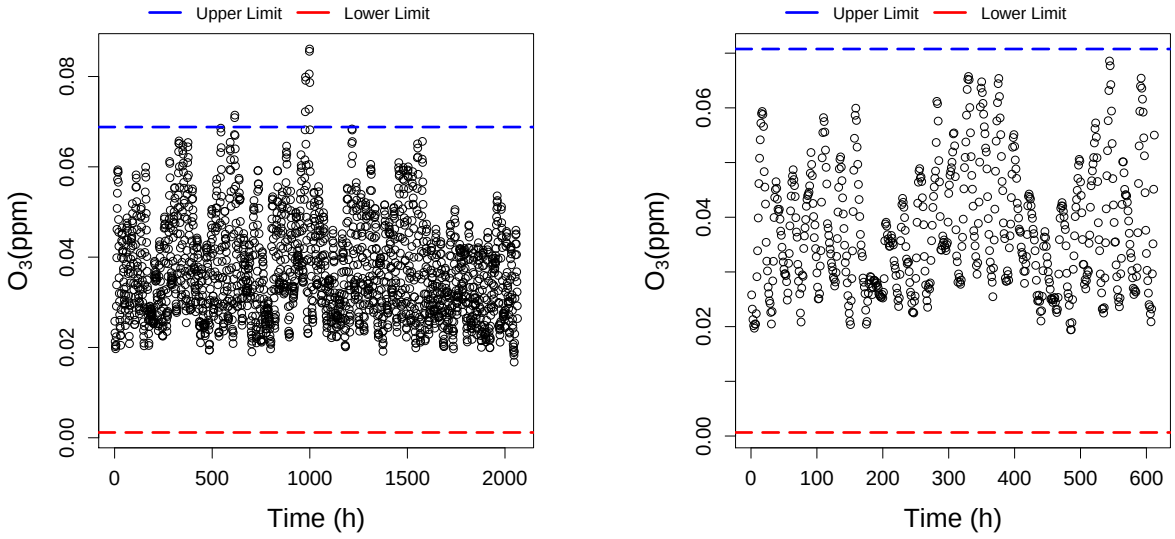

FIGURE A.28: *Left:* O_3 concentration plotted over time for the complete data and MAD. The upper limit in blue and the lower limit in red. *Right:* The longest outlier-free data points for the first 612 data points and MAD. The upper limit in blue and the lower limit in red.

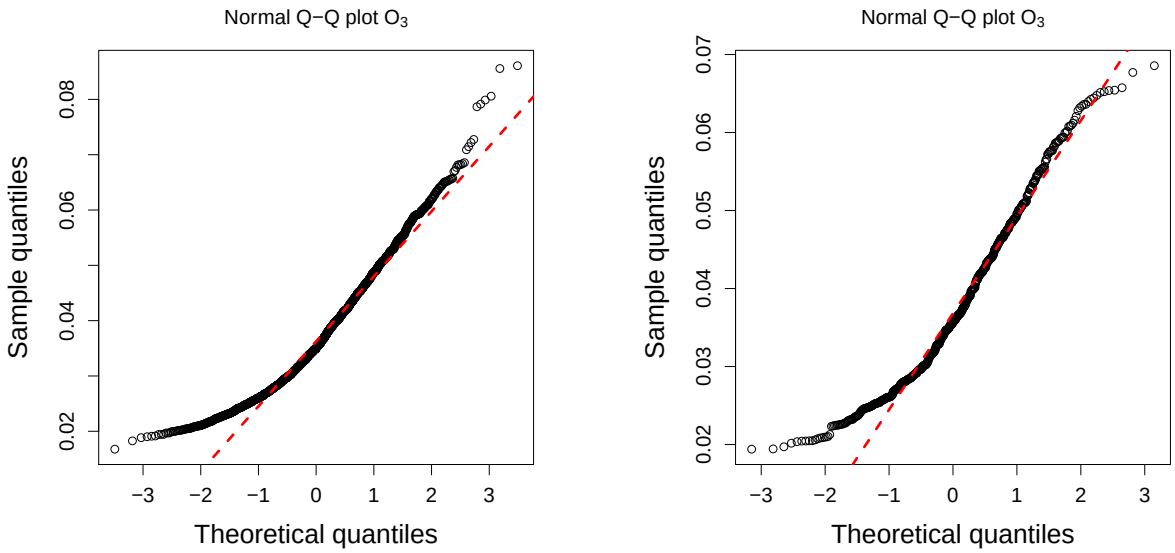

FIGURE A.29: *Left:* QQ plot for the complete data. *Right:* QQ plot for the longest outlier-free data points.

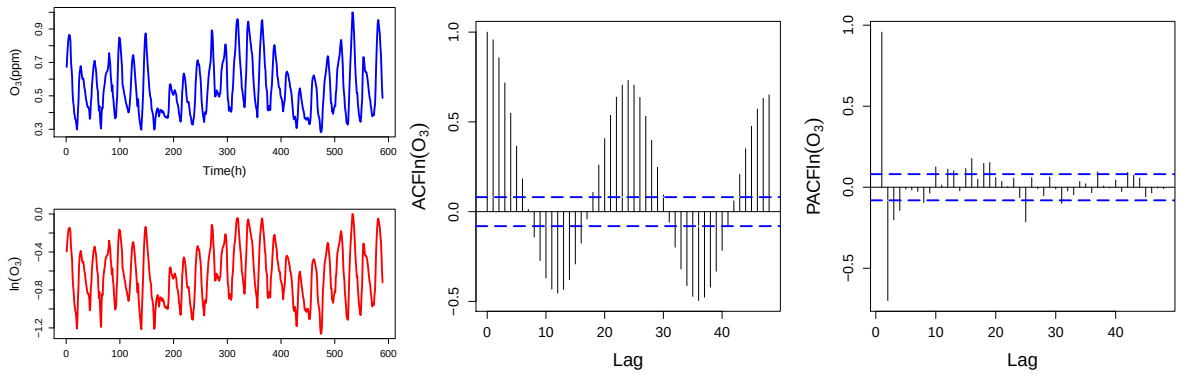

FIGURE A.30: *Left:* Normalized O_3 compared to $\ln(O_3)$ plotted over time. *Center:* ACF plotted over lags for $\ln(O_3)$. *Right:* PACF plotted over lags for $\ln(O_3)$.

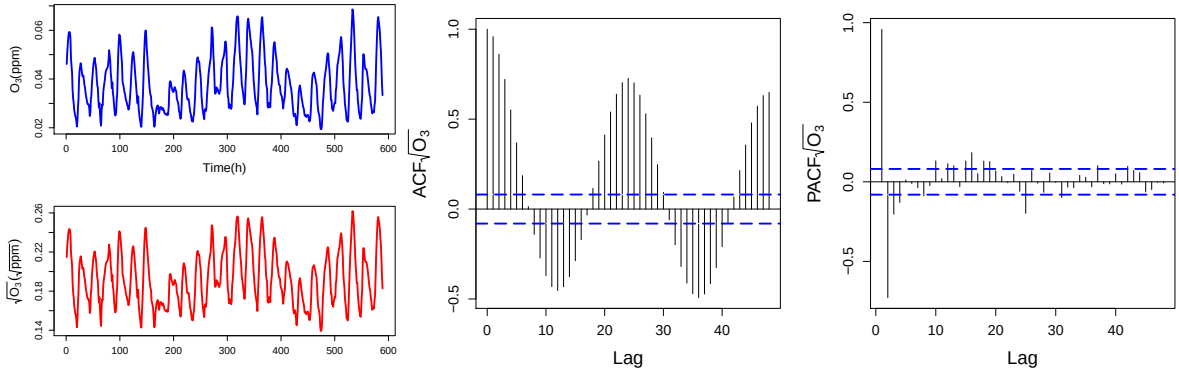

FIGURE A.31: *Left:* O_3 concentrations compared to $\sqrt{O_3}$ plotted over time. *Center:* ACF plotted over lags for $\sqrt{O_3}$. *Right:* PACF plotted over lags for $\sqrt{O_3}$.

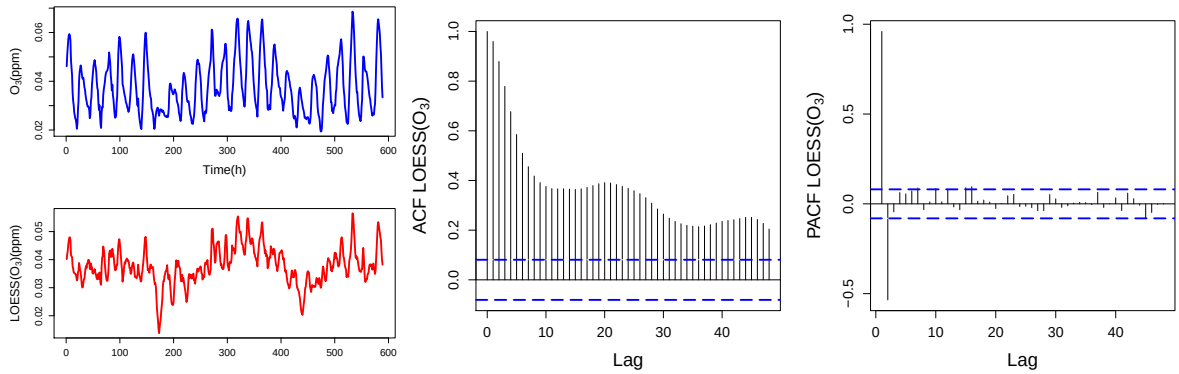

FIGURE A.32: *Left:* O_3 concentrations compared to LOESS(O_3) plotted over time. *Center:* ACF plotted over lags for LOESS(O_3). *Right:* PACF plotted over lags for LOESS(O_3).

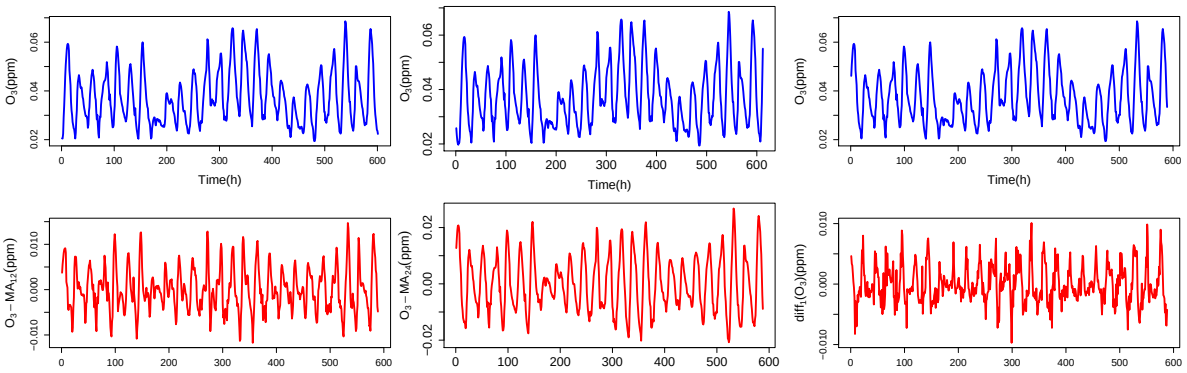

FIGURE A.33: *Left:* O_3 concentrations compared to O_3-MA_{12} plotted over time. *Center:* O_3 concentrations compared to O_3-MA_{24} plotted over time. *Right:* O_3 concentrations compared to $diff_1(O_3)$ plotted over time.

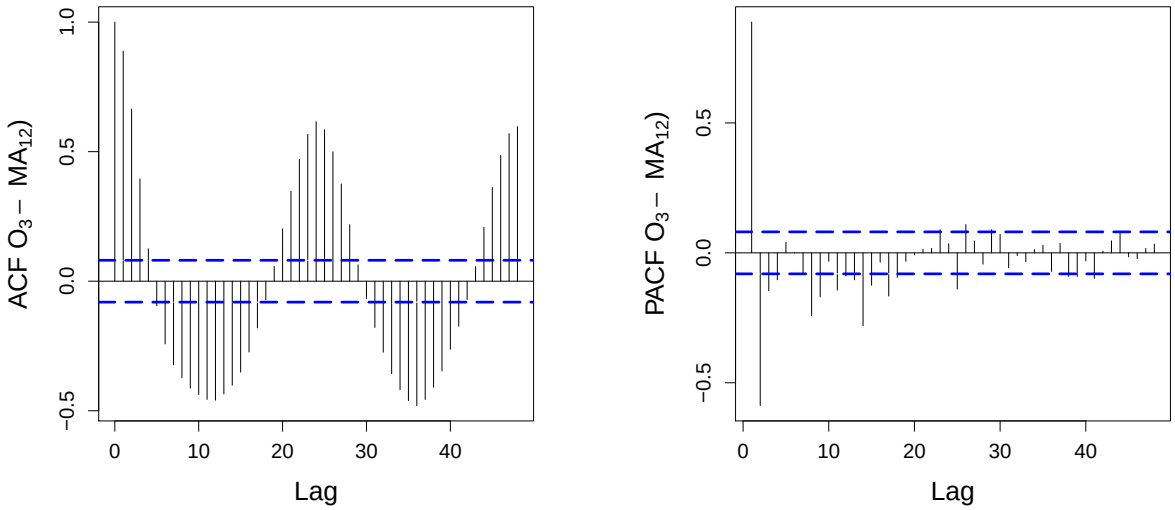

FIGURE A.34: ACF and PACF plots for the O_3-MA_{12} transformation.

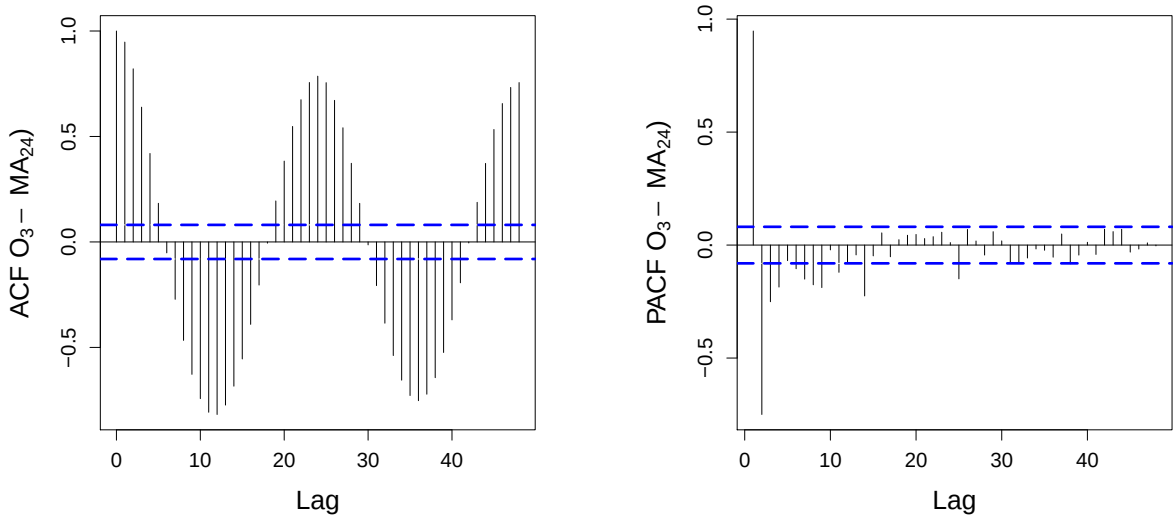

FIGURE A.35: ACF and PACF plots for the O_3-MA_{24} transformation.

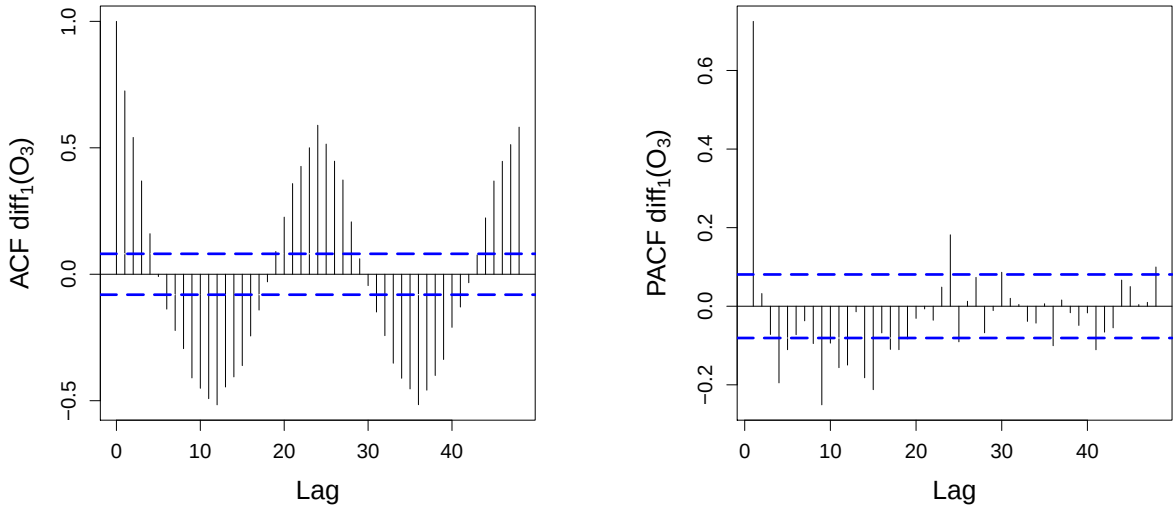

FIGURE A.36: ACF and PACF plots for the $diff_1(O_3)$ transformation.

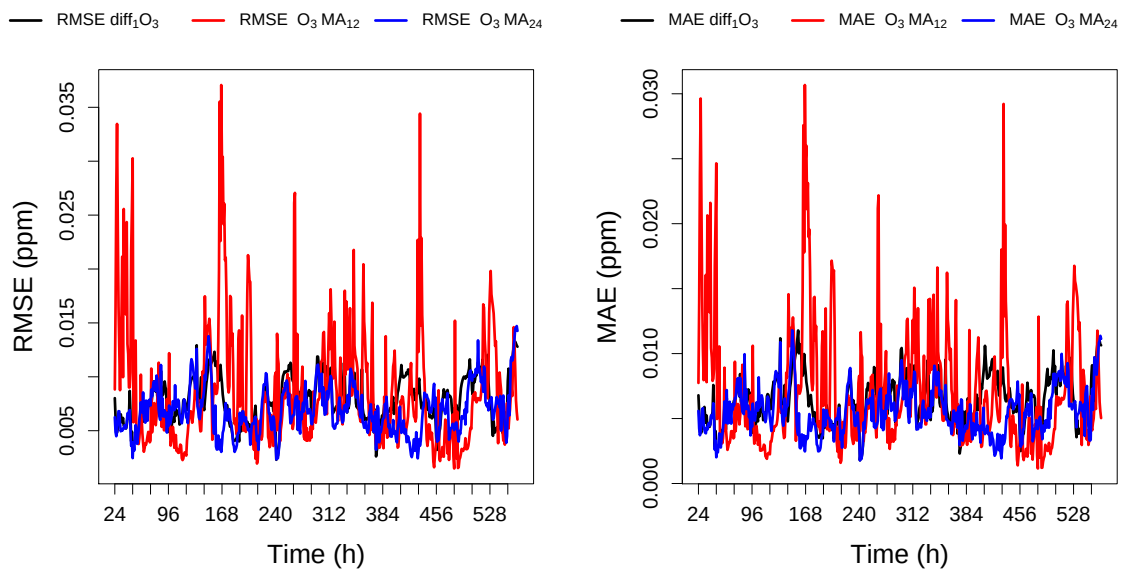

FIGURE A.37: *Left:* RMSE of different models plotted as a function of data points used by the model. *Rigth:* MAE of different models plotted as a function of data points used by the model.

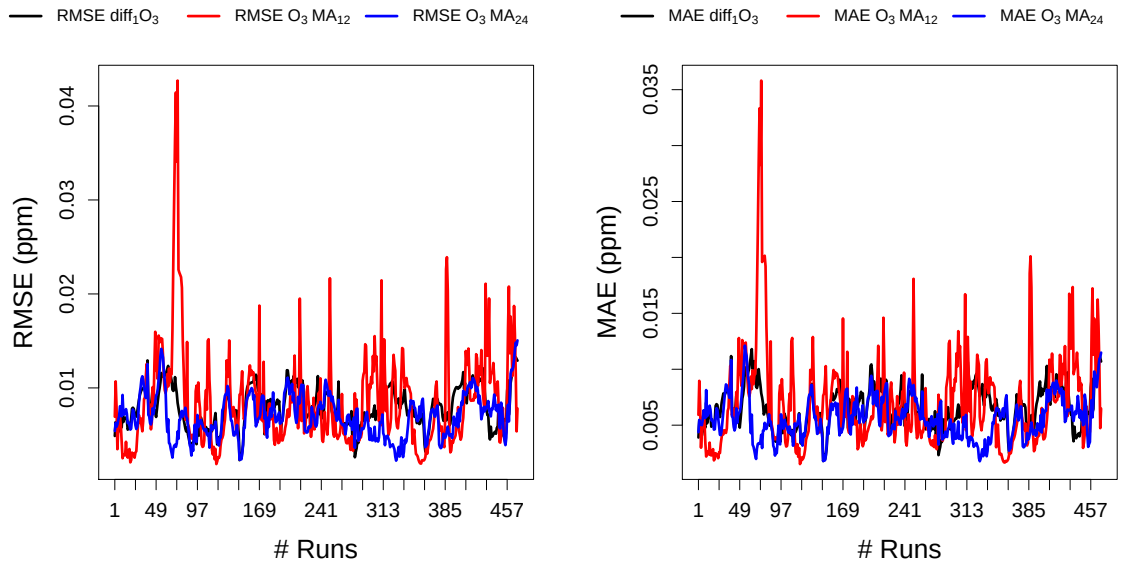

FIGURE A.38: *Left:* RMSE of different models plotted as a function of the number of fixed windows. *Rigth:* MAE of different models plotted as a function of fixed windows.

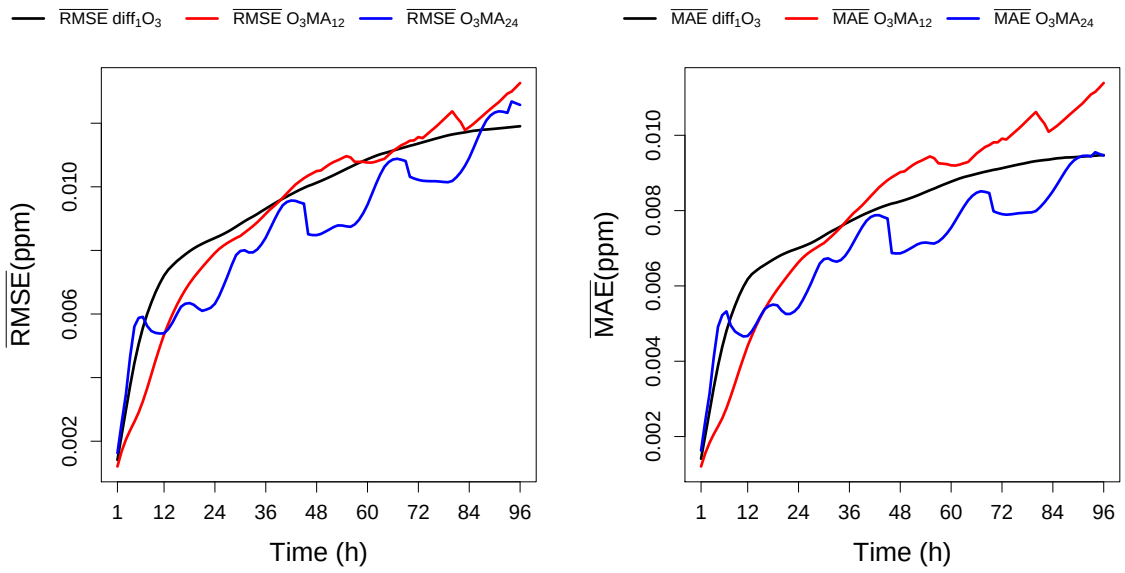

FIGURE A.39: *Left:* Average RMSE of different models plotted as a function of the forecast length. *Rigth:* Average MAE of different models plotted as a function of the forecast length.

A.4 PM_{10} - Particulate matter

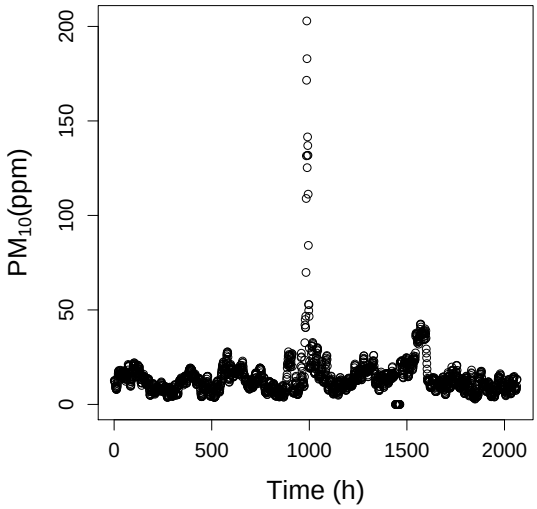

FIGURE A.40: PM_{10} concentration plotted over time for the complete data.

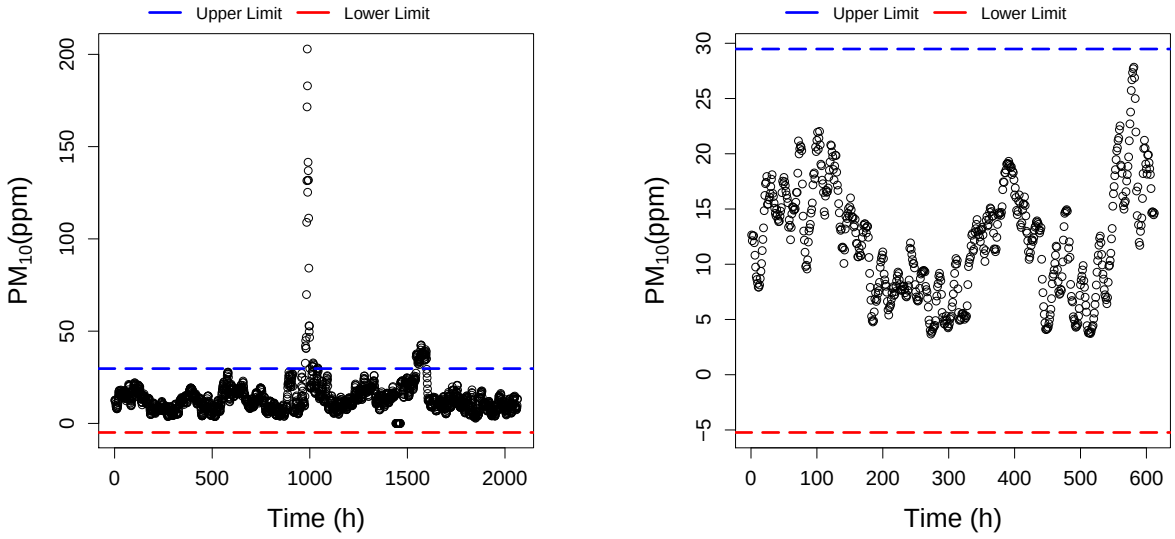

FIGURE A.41: *Left:* PM_{10} concentration plotted over time for the complete data and MAD. The upper limit in blue and the lower limit in red. *Right:* The longest outlier-free data points for the first 612 data points and MAD. The upper limit in blue and the lower limit in red.

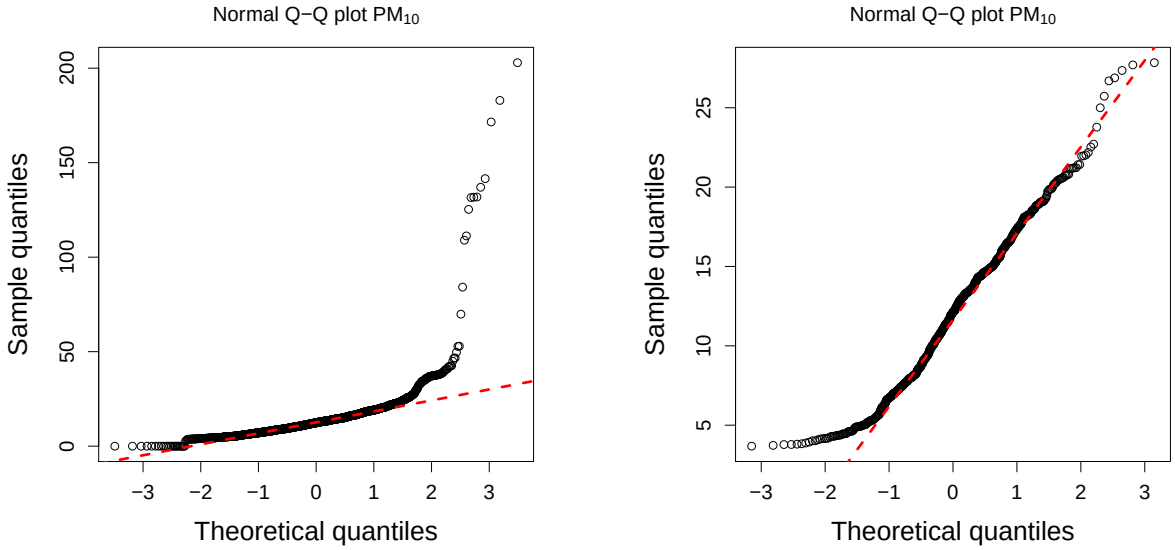

FIGURE A.42: *Left:* QQ plot for the complete data. *Right:* QQ plot for the longest outlier-free data points.

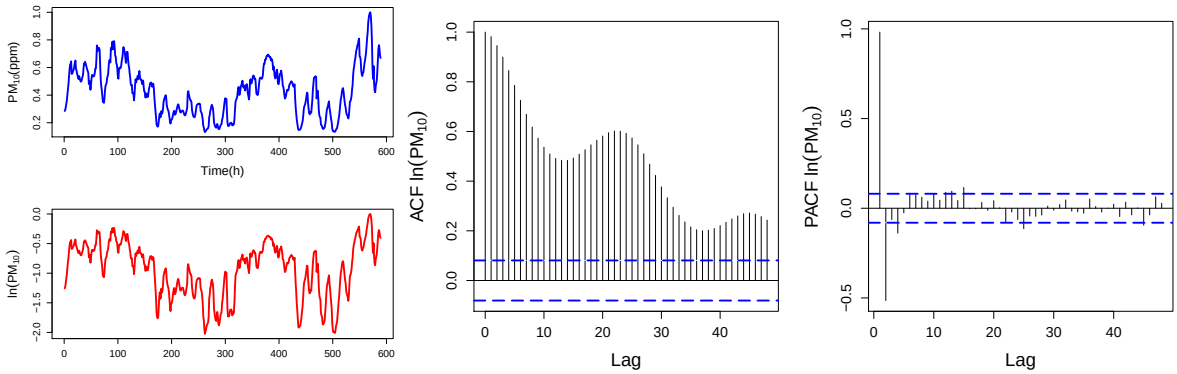

FIGURE A.43: *Left:* Normalized PM_{10} compared to $\ln(PM_{10})$ plotted over time. *Center:* ACF plotted over lags for $\ln(PM_{10})$. *Right:* PACF plotted over lags for $\ln(PM_{10})$.

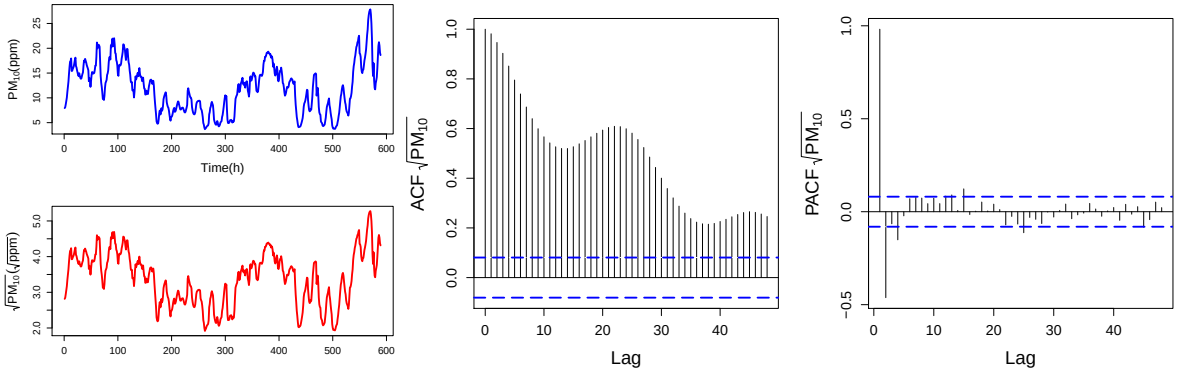

FIGURE A.44: *Left:* PM_{10} concentrations compared to $\sqrt{PM_{10}}$ plotted over time. *Center:* ACF plotted over lags for $\sqrt{PM_{10}}$. *Right:* PACF plotted over lags for $\sqrt{PM_{10}}$.

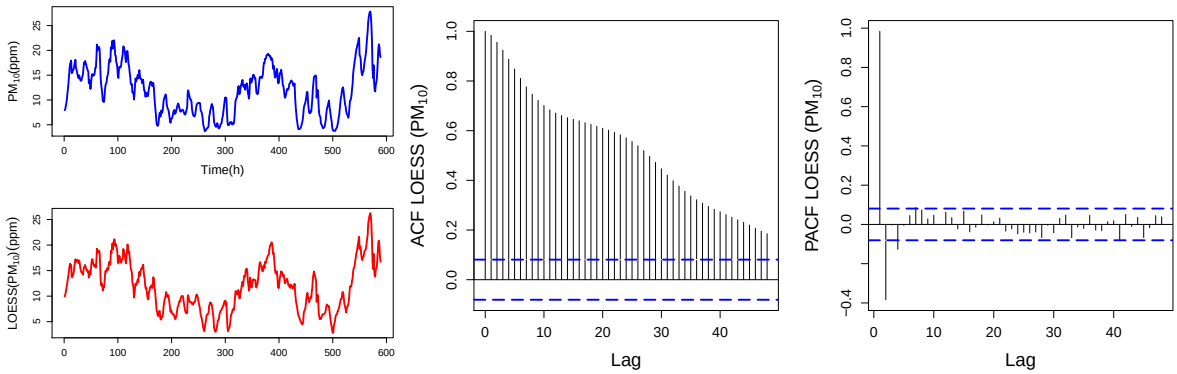

FIGURE A.45: *Left:* PM_{10} concentrations compared to LOESS(PM_{10}) plotted over time. *Center:* ACF plotted over lags for LOESS(PM_{10}). *Right:* PACF plotted over lags for LOESS(PM_{10}).

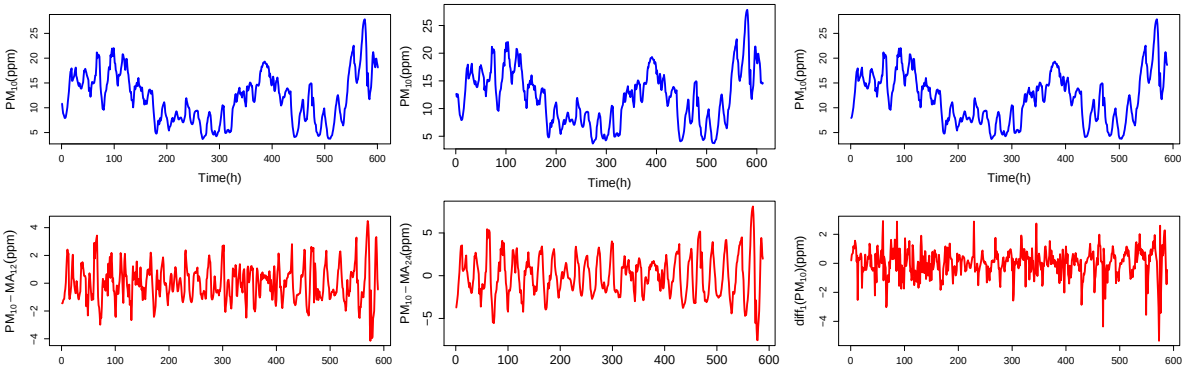

FIGURE A.46: *Left:* PM_{10} concentrations compared to PM_{10}-MA_{12} plotted over time. *Center:* PM_{10} concentrations compared to PM_{10}-MA_{24} plotted over time. *Right:* PM_{10} concentrations compared to $diff_1(PM_{10})$ plotted over time.

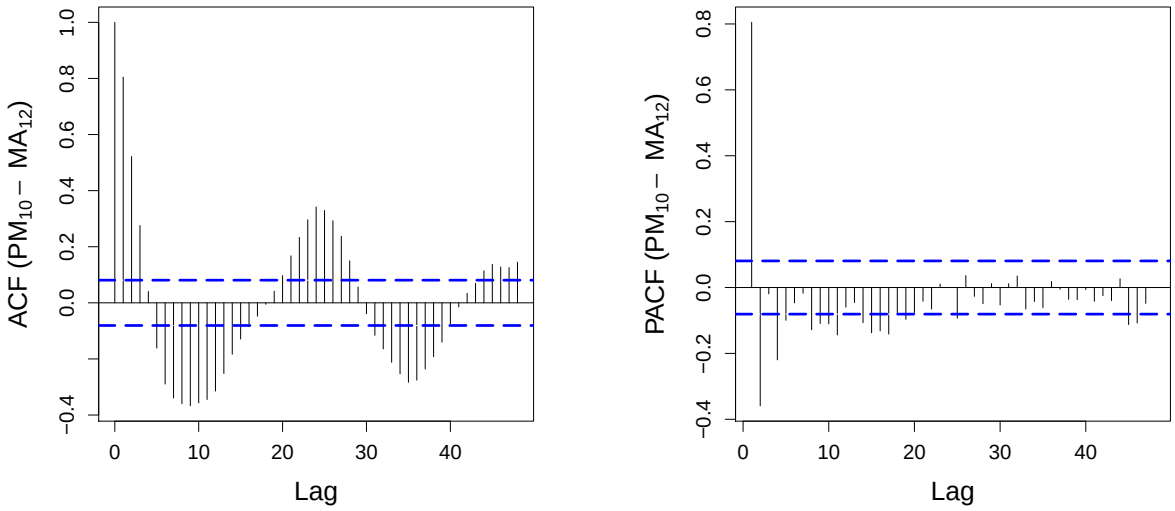

FIGURE A.47: ACF and PACF plots for the PM_{10}-MA_{12} transformation.

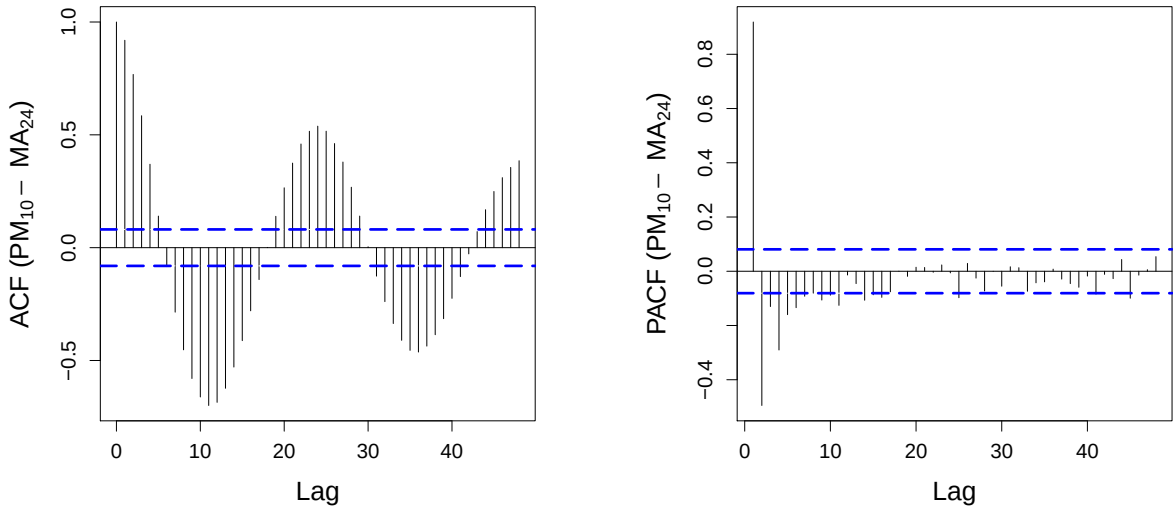

FIGURE A.48: ACF and PACF plots for the PM_{10}-MA_{24} transformation.

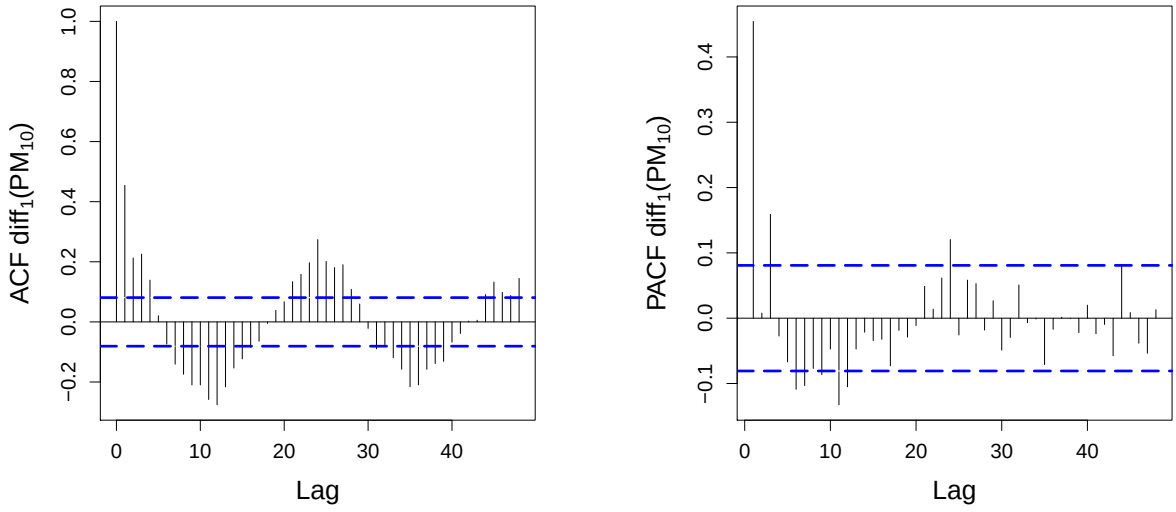

FIGURE A.49: ACF and PACF plots for the $diff_1(PM_{10})$ transformation.

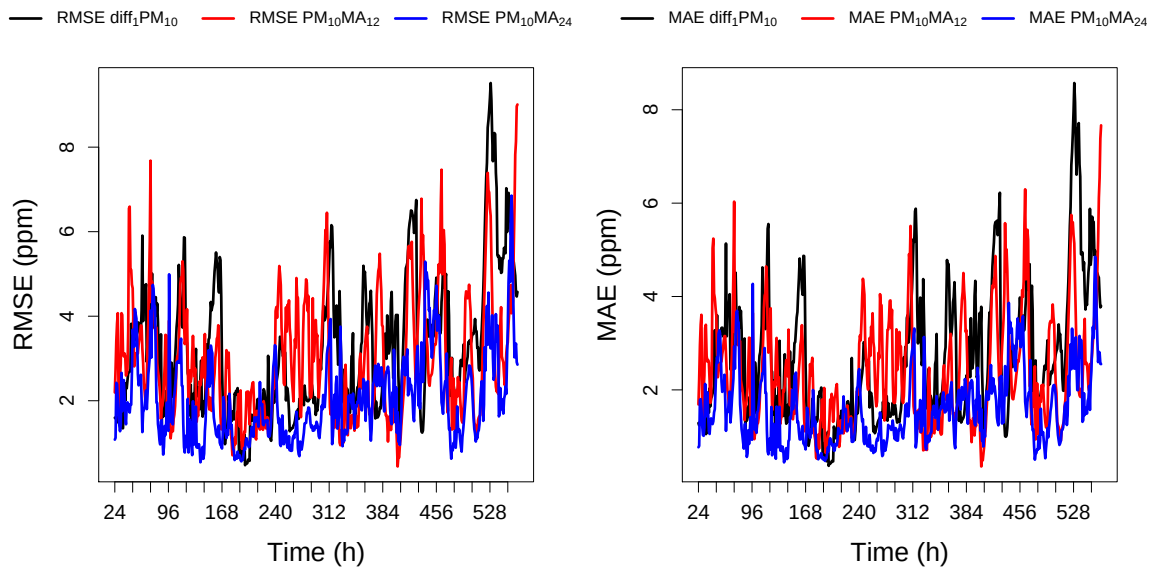

FIGURE A.50: *Left:* RMSE of different models plotted as a function of data points used by the model. *Rigth:* MAE of different models plotted as a function of data points used by the model.

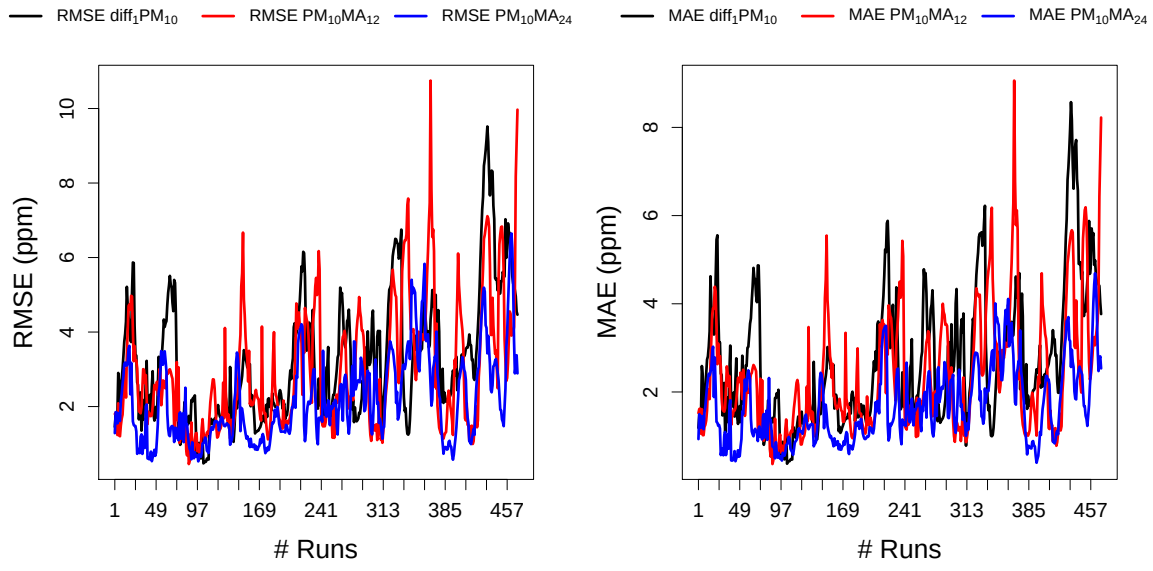

FIGURE A.51: *Left:* RMSE of different models plotted as a function of the number of fixed windows. *Rigth:* MAE of different models plotted as a function of fixed windows.

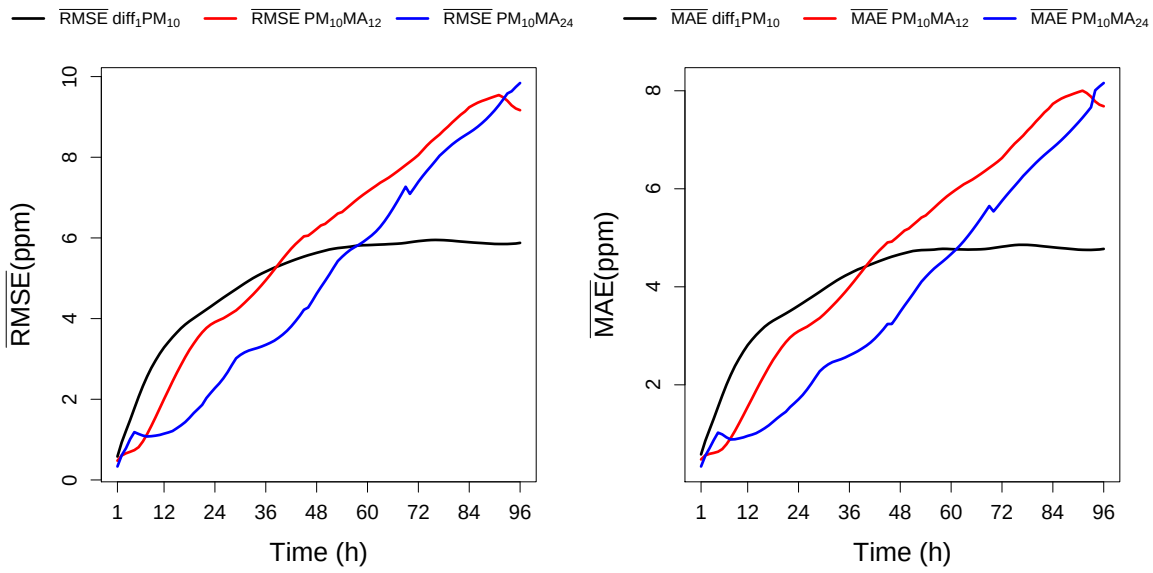

FIGURE A.52: *Left:* Average RMSE of different models plotted as a function of the forecast length. *Rigth:* Average MAE of different models plotted as a function of the forecast length.

A.5 $PM_{2.5}$ - Particulate matter

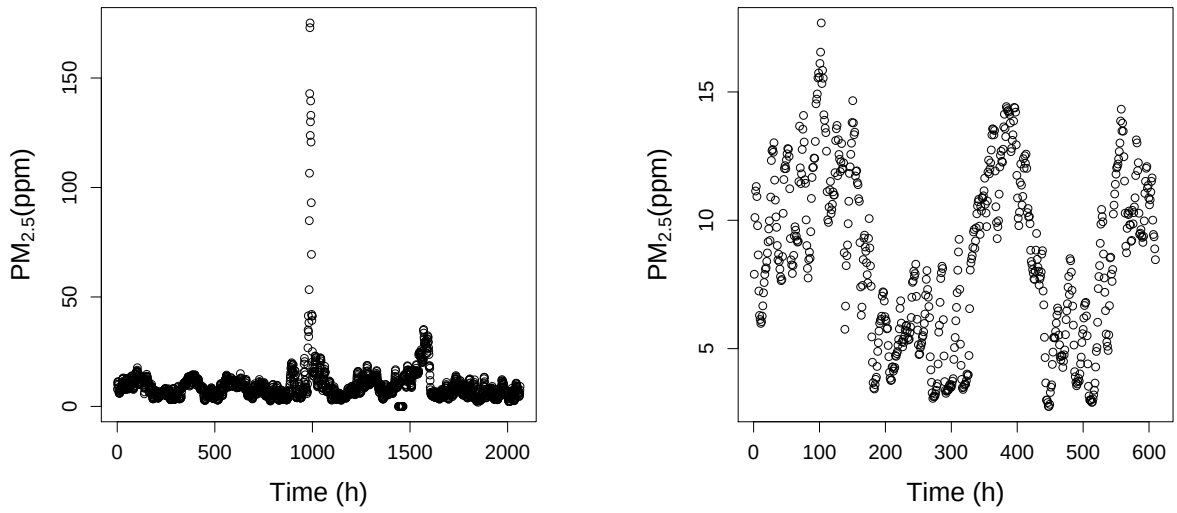

FIGURE A.53: *Left:* $PM_{2.5}$ concentration plotted over time for the complete data. *Right:* $PM_{2.5}$ concentration plotted over time for the longest outlier-free data points, common to every pollutant.

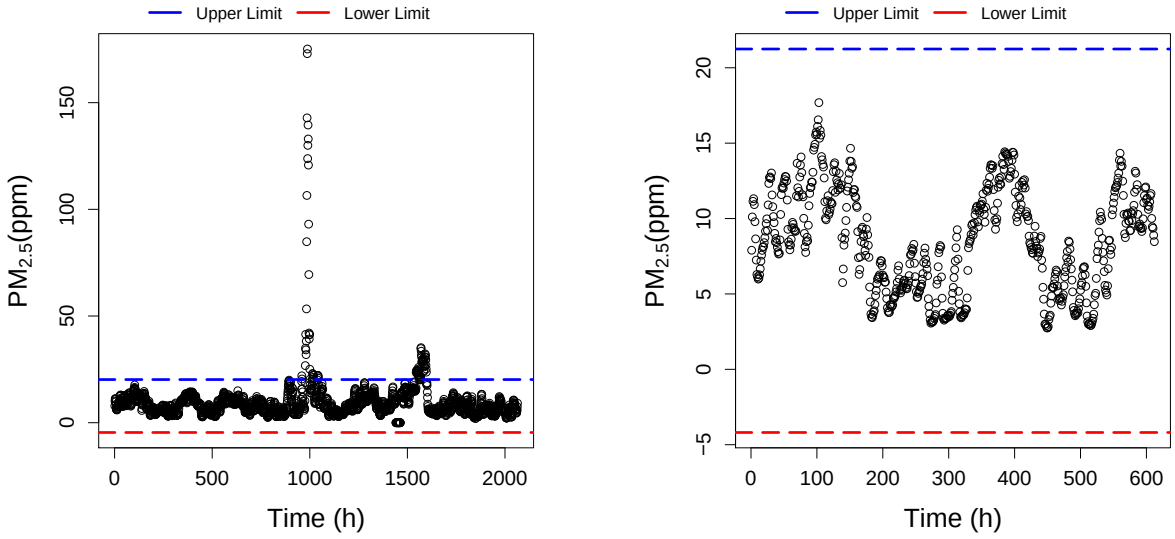

FIGURE A.54: *Left:* $PM_{2.5}$ concentration plotted over time for the complete data and MAD. The upper limit in blue and the lower limit in red. *Right:* The longest outlier-free data points for the first 612 data points and MAD. The upper limit in blue and the lower limit in red.

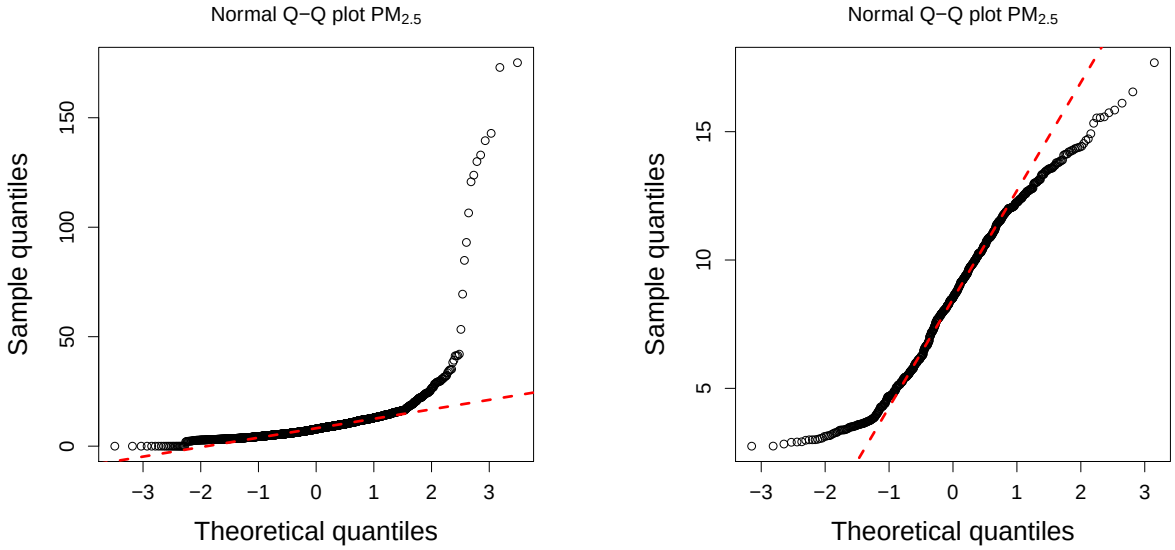

FIGURE A.55: *Left:* QQ plot for the complete data. *Right:* QQ plot for the longest outlier-free data points.

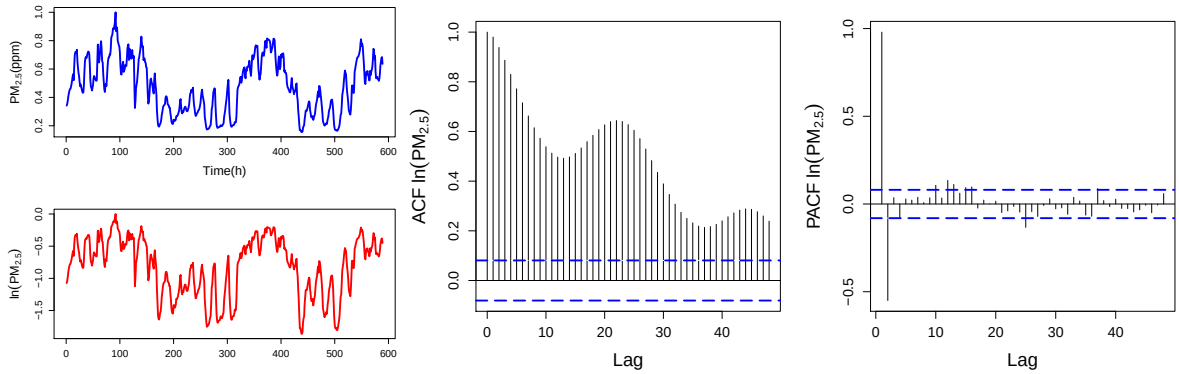

FIGURE A.56: *Left:* Normalized $PM_{2.5}$ compared to $\ln(PM_{2.5})$ plotted over time. *Center:* ACF plotted over lags for $\ln(PM_{2.5})$. *Right:* PACF plotted over lags for $\ln(PM_{2.5})$.

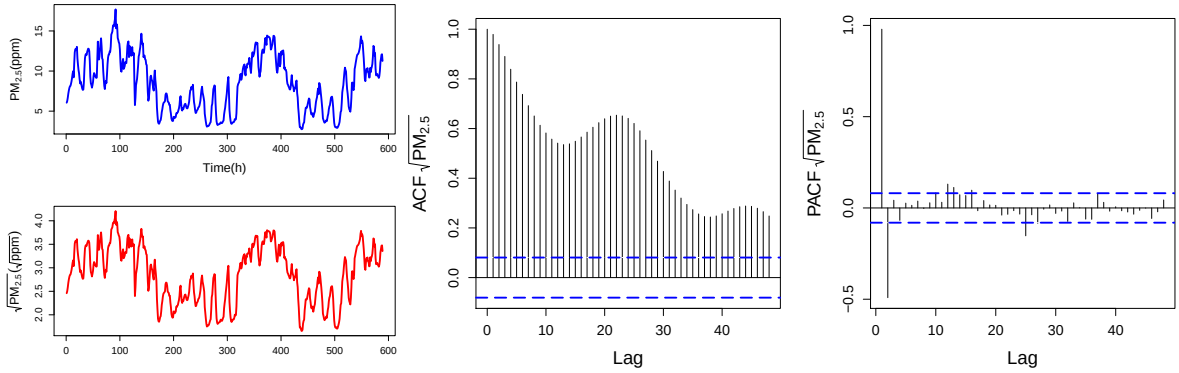

FIGURE A.57: *Left:* $PM_{2.5}$ concentrations compared to $\sqrt{PM_{2.5}}$ plotted over time. *Center:* ACF plotted over lags for $\sqrt{PM_{2.5}}$. *Right:* PACF plotted over lags for $\sqrt{PM_{2.5}}$.

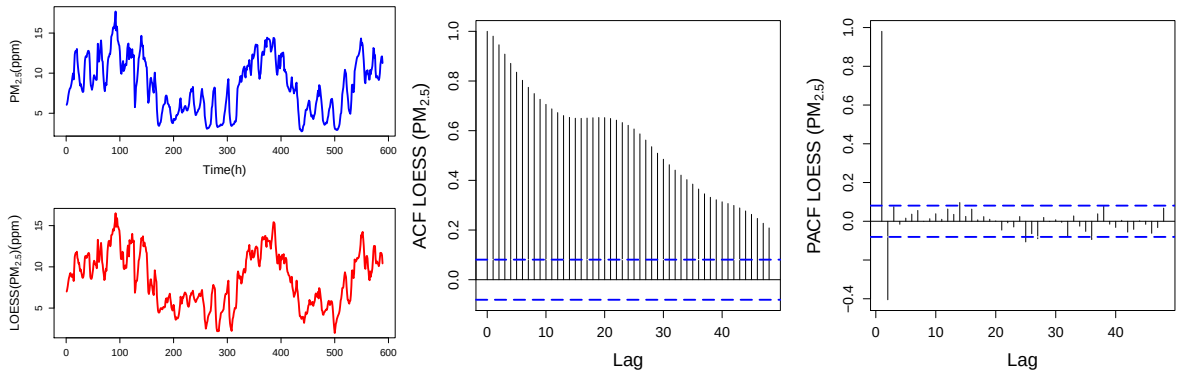

FIGURE A.58: *Left:* $PM_{2.5}$ concentrations compared to LOESS($PM_{2.5}$) plotted over time. *Center:* ACF plotted over lags for LOESS($PM_{2.5}$). *Right:* PACF plotted over lags for LOESS($PM_{2.5}$).

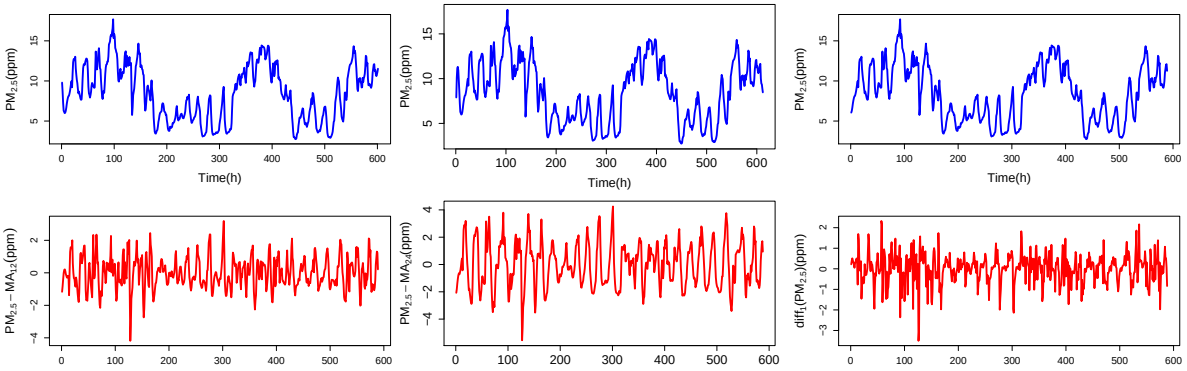

FIGURE A.59: *Left:* $PM_{2.5}$ concentrations compared to $PM_{2.5}$-MA_{12} plotted over time. *Center:* $PM_{2.5}$ concentrations compared to $PM_{2.5}$-MA_{24} plotted over time. *Right:* $PM_{2.5}$ concentrations compared to $\text{diff}_1(PM_{2.5})$ plotted over time.

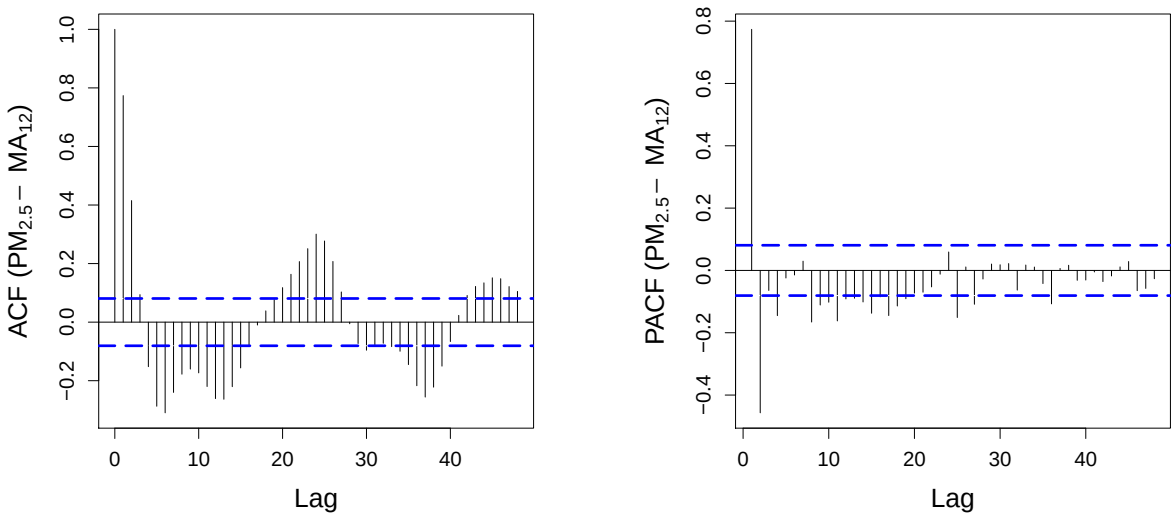

FIGURE A.60: ACF and PACF plots for the $PM_{2.5}$-MA_{12} transformation.

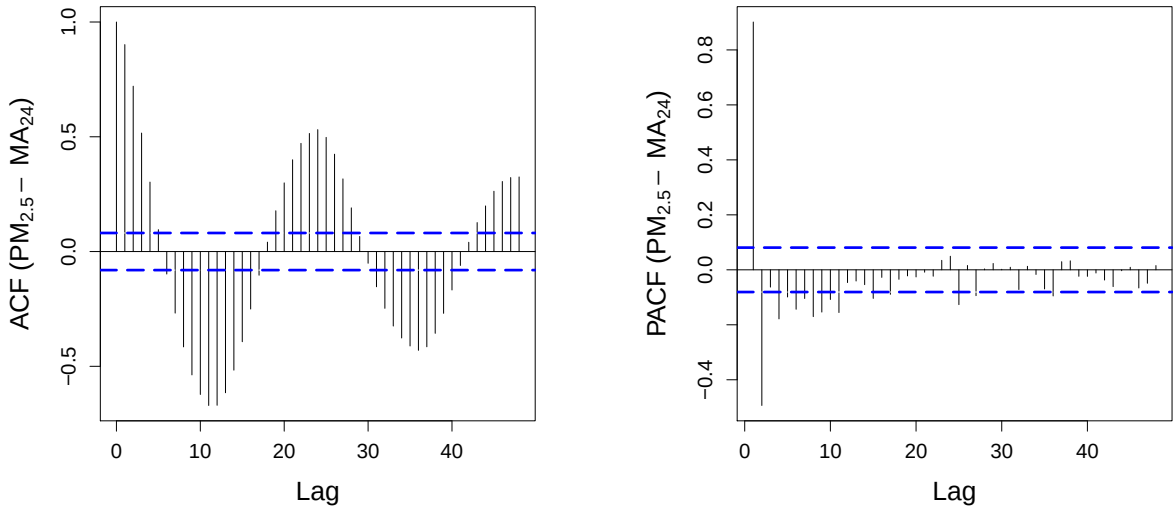

FIGURE A.61: ACF and PACF plots for the $PM_{2.5}$-MA_{24} transformation.

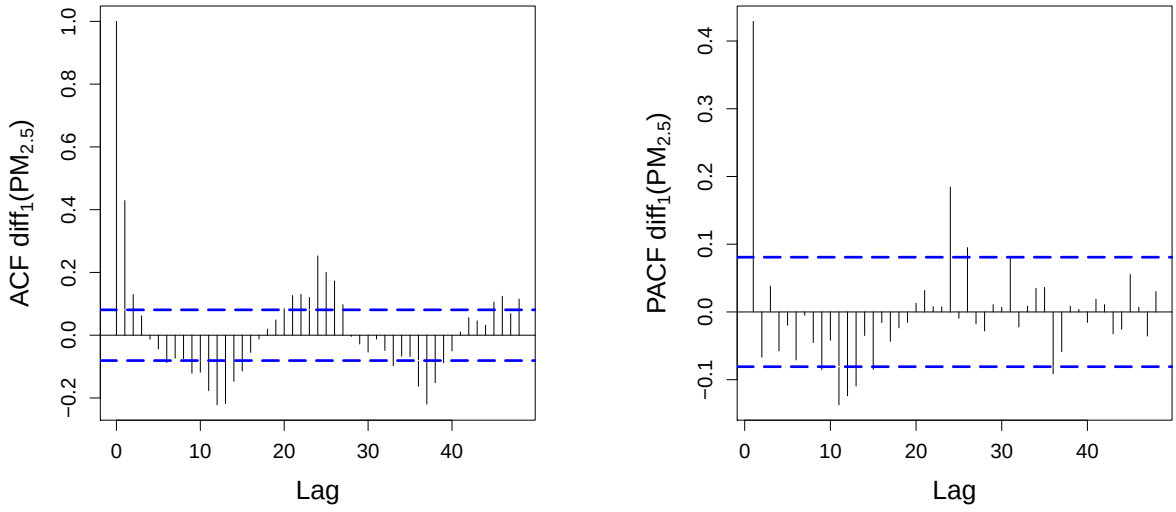

FIGURE A.62: ACF and PACF plots for the $diff_1(PM_{2.5})$ transformation.

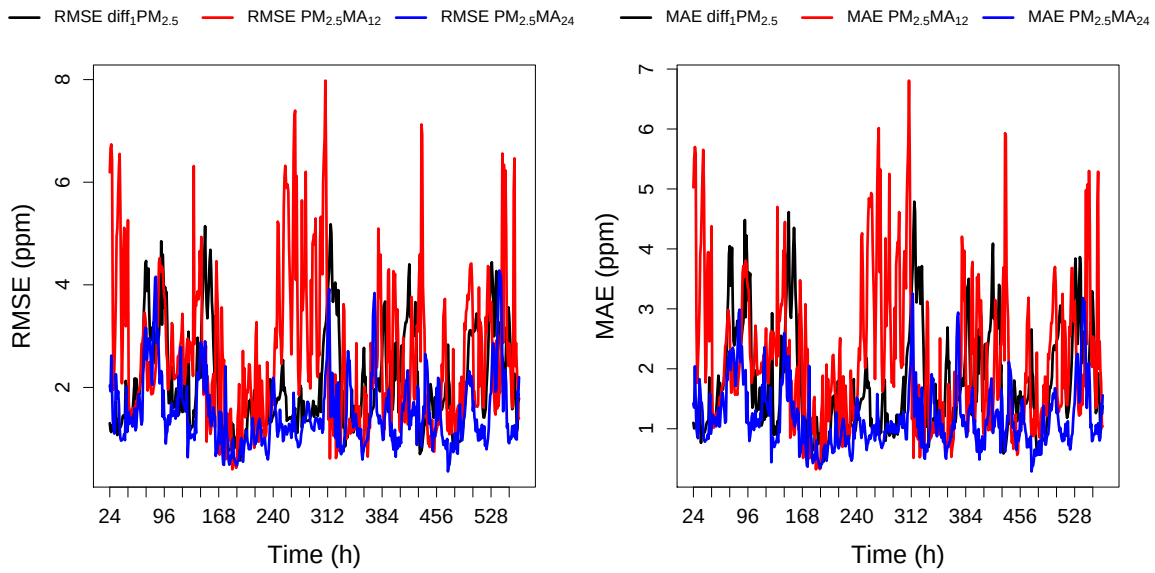

FIGURE A.63: *Left:* RMSE of different models plotted as a function of data points used by the model. *Rigth:* MAE of different models plotted as a function of data points used by the model.

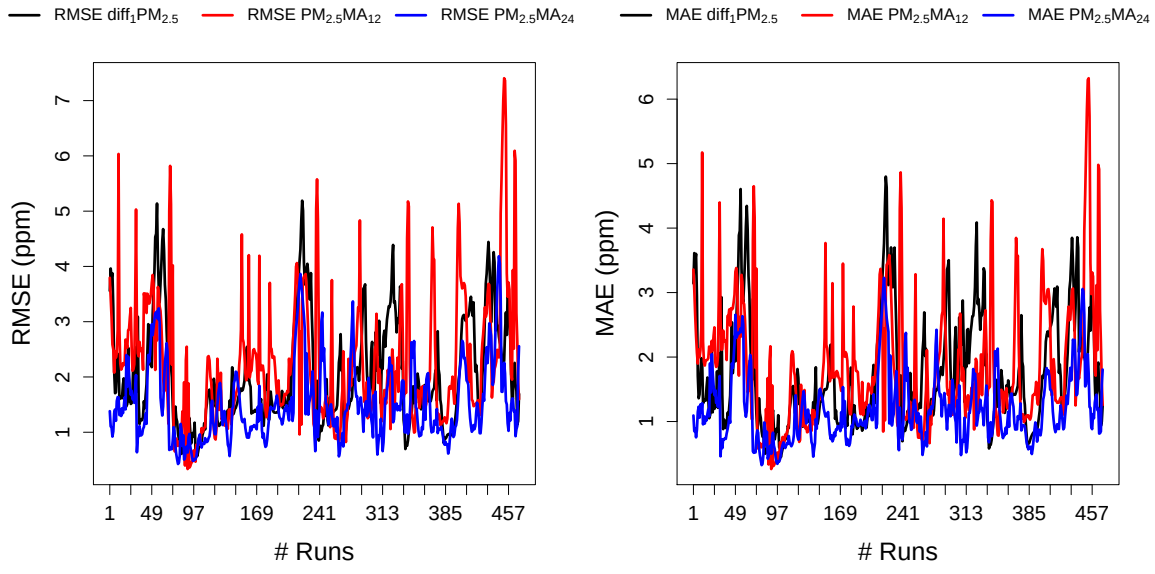

FIGURE A.64: *Left:* RMSE of different models plotted as a function of the number of fixed windows. *Rigth:* MAE of different models plotted as a function of fixed windows.

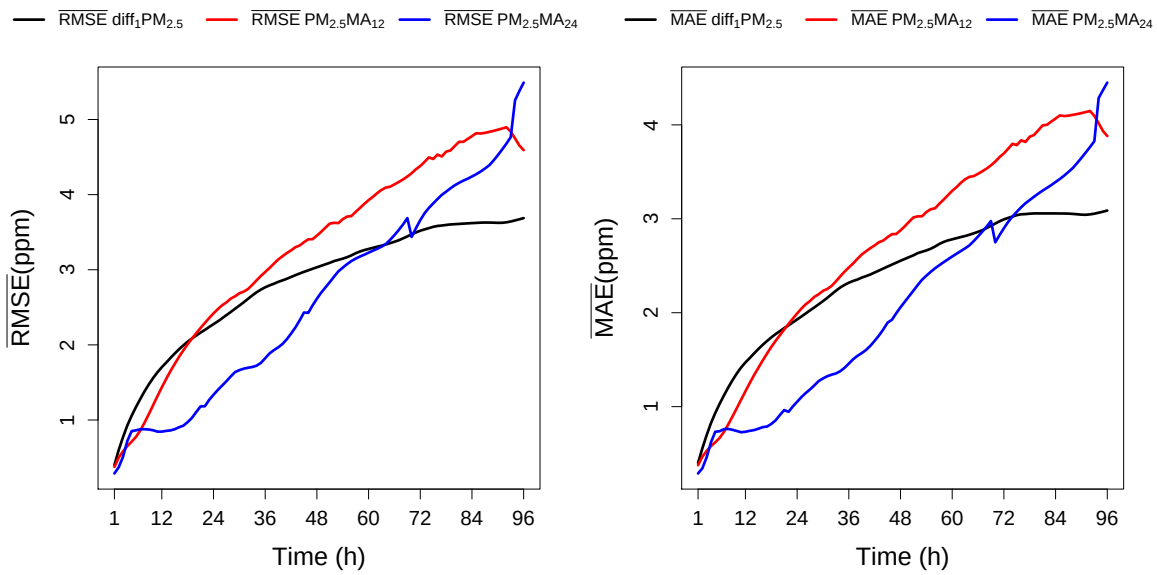

FIGURE A.65: *Left:* Average RMSE of different models plotted as a function of the forecast length. *Rigth:* Average MAE of different models plotted as a function of the forecast length.

Bibliography

[1] Klea Katsouyanni. "Ambient air pollution and health". In: *British Medical Bulletin* 68 (2003), pp. 143–156. ISSN: 00071420. DOI: 10.1093/bmb/ldg028.

[2] Michelle L. Bell, Devra L. Davis, and Tony Fletcher. "A Retrospective Assessment of Mortality from the London Smog Episode of 1952: The Role of Influenza and Pollution". In: *Environmental Health Perspectives* 112.1 (Oct. 2003), pp. 6–8. ISSN: 0091-6765. DOI: 10.1289/ehp.6539. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241116/.

[3] European Environment Agency. *Air quality in Europe- 2017 Report*. 2017. ISBN: 978-92-9213-921-6. DOI: 10.2800/850018. URL: https://www.eea.europa.eu/publications/air-quality-in-europe-2017.

[4] Manuel Ferreira de Magalhães et al. "Cost of asthma in children: A nationwide, population-based, cost-of-illness study". In: *Pediatric Allergy and Immunology* 28.7 (Nov. 2017), pp. 683–691. ISSN: 09056157. DOI: 10.1111/pai.12772. URL: http://doi.wiley.com/10.1111/pai.12772.

[5] Michael Guarnieri and John R. Balmes. "Outdoor air pollution and asthma". In: *The Lancet* 383.9928 (2014), pp. 1581–1592. ISSN: 1474547X. DOI: 10.1016/S0140-6736(14)60617-6. arXiv: arXiv:1011.1669v3. URL: http://dx.doi.org/10.1016/S0140-6736(14)60617-6.

[6] Yang Zhang et al. "Real-time air quality forecasting, part I: History, techniques, and current status". In: *Atmospheric Environment* 60 (2012), pp. 632–655. ISSN: 13522310. DOI: 10.1016/j.atmosenv.2012.06.031. URL: http://dx.doi.org/10.1016/j.atmosenv.2012.06.031.

[7] Luis A. Díaz-Robles et al. "A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile". In: *Atmospheric Environment* 42.35 (2008), pp. 8331–8340. ISSN: 13522310. DOI: 10.1016/j.atmosenv.2008.07.020.

[8] Alain Louis Dutot et al. "A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions". In: *Environmental Modelling and Software* 22.9 (2007), pp. 1261–1269. ISSN: 13648152. DOI: 10.1016/j.envsoft.2006.08.002. arXiv: 0802.3969.

[9] Yang Zhang et al. "Real-time air quality forecasting, Part II: State of the science, current research needs, and future prospects". In: *Atmospheric Environment* 60 (2012), pp. 656–676. ISSN: 13522310. DOI: 10.1016/j.atmosenv.2012.02.041. URL: http://dx.doi.org/10.1016/j.atmosenv.2012.02.041.

[10] Agência Portuguesa do Ambiente. *Previsão Diária do Índice de Qualidade do Ar*. 2018. URL: http://www.prevqualar.org/jsp/pt/previsao%7B%5C_%7Dcidades.jsp.

[11] L. Menut et al. "CHIMERE 2013: a model for regional atmospheric composition modelling". In: *Geoscientific Model Development* 6.4 (2013), pp. 981–1028. ISSN: 1991-9603. DOI: 10.5194/gmd-6-981-2013. URL: http://www.geosci-model-dev.net/6/981/2013/.

[12] H. Schmidt et al. "A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe". In: *Atmospheric Environment* 35.36 (2001), pp. 6277–6297. ISSN: 13522310. DOI: 10.1016/S1352-2310(01)00451-4.

[13] B. Bessagnet et al. "Aerosol modeling with CHIMERE - Preliminary evaluation at the continental scale". In: *Atmospheric Environment* 38.18 (2004), pp. 2803–2817. ISSN: 13522310. DOI: 10.1016/j.atmosenv.2004.02.034.

[14] Wei Yin Loh. "Classification and regression trees". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (2011), pp. 14–23. ISSN: 19424787. DOI: 10.1002/widm.8. arXiv: arXiv:1011.1669v3.

[15] Universidade NOVA de Lisboa PrevQualar. *Qualidade do ar - Sobre*. 2015. URL: http://www.prevqualar.org/content.action?cid=contentProject.

[16] Raquel Prado and Mike West. *Time Series: Modelling, Computation and Inference*. 2010, p. 368. ISBN: 9781420093360.

[17] G E P Box and G M Jenkins. *Time Series Analysis: Forecasting and Control*. Vol. Third. 1994, pp. 1–7. ISBN: 0130607746. DOI: 10.1111/j.1467-9892.2009.00643.x. URL: http://doi.wiley.com/10.1111/j.1467-9892.2009.00643.x.

[18] Ujjwal Kumar and V. K. Jain. "ARIMA forecasting of ambient air pollutants (O3, NO, NO2and CO)". In: *Stochastic Environmental Research and Risk Assessment* 24.5 (2010), pp. 751–760. ISSN: 14363240. DOI: 10.1007/s00477-009-0361-8.

[19] Muhammad Hisyam Lee et al. "Seasonal ARIMA for Forecasting Air Pollution Index: A Case Study". In: *American Journal of Applied Sciences* 9.4 (2012), pp. 570–578. ISSN: 1546-9239. URL: http://thescipub.com/PDF/ajassp.2012.570.578.pdf.

[20] Rati Wongsathan and Isaravuth Seedadan. "A Hybrid ARIMA and Neural Networks Model for PM-10 Pollution Estimation: The Case of Chiang Mai City Moat Area". In: *Procedia Computer Science* 86.March (2016), pp. 273–276. ISSN: 18770509. DOI: 10.1016/j.procs.2016.05.057. URL: http://dx.doi.org/10.1016/j.procs.2016.05.057.

[21] Ramalingam Shanmugam, Peter J. Brockwell, and Richard A. Davis. *Introduction to Time Series and Forecasting*. Vol. 39. 4. 1997, p. 426. ISBN: 0387953515. DOI: 10.2307/1271510. arXiv: arXiv:1011.1669v3. URL: http://www.jstor.org/stable/1271510?origin=crossref.

[22] C. De Wispelaere, ed. *Air Pollution Modeling and Its Application II*. Boston, MA: Springer US, 1983. ISBN: 978-1-4684-7943-0. DOI: 10.1007/978-1-4684-7941-6. URL: http://link.springer.com/10.1007/978-1-4684-7941-6.

[23] Carlos Borrego and Guy Schayes, eds. *Air Pollution Modeling and Its Application XV*. Boston, MA: Springer US, 2002. ISBN: 978-0-306-47294-7. DOI: 10.1007/b105277. URL: http://www.lob.de/cgi-bin/work/suche2?titnr=249420018%7B%5C&%7Dflag=citavi%20http://link.springer.com/10.1007/b105277.

[24] a. V. Metcalfe. "Introduction to Stochastic Processes." In: *The Statistician* 45.4 (1996), p. 533. ISSN: 00390526. DOI: 10.2307/2988557.

[25] John Lamperti. *Stochastic Processes - A Survey of the Mathematical Theory*. Springer-Verlag, 1977. ISBN: 9783540902751.

[26] Hans Von Storch and Francis W Zwiers. "Statistical Analysis in Climate Research". In: *Journal of the American Statistical Association* 95 (1999), p. 1375. ISSN: 01621459. DOI: 10.1017/CBO9780511612336. URL: https://goo.gl/2B3dxY.

[27] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. 2014. ISBN: 978-0-98-75071-0-5. DOI: 10.1017/CBO9781107415324.004.

[28] Wayne A Fuller. *Introduction to Statistical Time Series*. 1976. ISBN: 0471552399.

[29] NIST. *Engineering Statistics Handbook*. 2006. ISBN: 9780471388791. DOI: 10.1016/B978-081551447-3.50015-7.

[30] Robert B Cleveland et al. *STL: A seasonal-trend decomposition procedure based on loess*. 1990. DOI: citeulike-article-id:1435502.

[31] Encyclopedia of Mathematics. *Box-Cox transformation*. 2012. URL: https://goo.gl/iRL6Wx.

[32] Ken Aho, DeWayne Derryberry, and Teri Peterson. "Model selection for ecologists: the worldview of AIC and BIC". In: *Ecology* 95.3 (2014), pp. 631–636. ISSN: 00129658. DOI: 10.1890/13-1452.1. arXiv: arXiv:1011.1669v3. URL: http://internal-pdf//Aho%20et%20al.%202014-2324678410/Aho%20et%20al.%202014.pdf.

[33] Kenneth P. Burnham and David R. Anderson. "Multimodel inference: Understanding AIC and BIC in model selection". In: *Sociological Methods and Research* 33.2 (2004), pp. 261–304. ISSN: 00491241. DOI: 10.1177/0049124104268644. arXiv: arXiv:1011.1669v3.

[34] Rob J. Hyndman and Anne B. Koehler. "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4 (2006), pp. 679–688. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2006.03.001. arXiv: Rodgers,J.L.,{\&}Nicewander,W.A.(2008).ThirteenWaystoLookattheCorrelationCoefficient,42(1),59âĂŞ66..

[35] J. Scott Armstrong. *Principles of Forecasting*. Ed. by J. Scott Armstrong. Vol. 30. International Series in Operations Research & Management Science. Boston, MA: Springer US, July 2001. ISBN: 978-0-7923-7401-5. DOI: 10.1007/978-0-306-47630-3. URL: https://www.springer.com/la/book/9780792379300%20http://link.springer.com/10.1007/978-0-306-47630-3.

[36] University of Virginia Library Clay Ford. *Understanding Q-Q Plots*. 2015. URL: https://data.library.virginia.edu/understanding-q-q-plots/.

[37] Rob J. Hyndman and Yeasmin Khandakar. "Automatic time series forecasting: The forecast package for R". In: *Journal Of Statistical Software* 27.3 (2008), pp. C3–C3. ISSN: 10411135. DOI: 10.18637/jss.v027.i03. arXiv: 1701.05936. URL: http://www.robjhyndman.com/papers/forecastpackage.pdf.

[38] Xiaozhe Wang, Kate Smith, and Rob Hyndman. "Characteristic-based clustering for time series data". In: *Data Mining and Knowledge Discovery* 13.3 (2006), pp. 335–364. ISSN: 13845810. DOI: 10.1007/s10618-005-0039-x.

[39] Rob Hyndman et al. *Forecasting with Exponential Smoothing The state Space Approach*. 2008. ISBN: 978-3-540-71916-8. DOI: 10.1007/978-3-540-71918-2.

[40] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: 4 (2009), pp. 40–79. ISSN: 1935-7516. DOI: 10.1214/09-SS054. arXiv: 0907.4728. URL: http://arxiv.org/abs/0907.4728%7B%5C%%7D0Ahttp://dx.doi.org/10.1214/09-SS054.

[41] Leonard J. Tashman. "Out-of-sample tests of forecasting accuracy: an analysis and review". In: *International Journal of Forecasting* 16.4 (2000), pp. 437–450. ISSN: 01692070. DOI: 10.1016/S0169-2070(00)00065-0.

[42] et al. Leys, C. "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median". In: *Journal of Experimental Social Psychology* 49.4 (2013), Pages 764–766. URL: http://dx.doi.org/10.1016/j.jesp.2013.03.013.

[43] Keith Ord, Michèle Hibon, and Spyros Makridakis. "The M3-Competition". In: *International Journal of Forecasting* 16.4 (Oct. 2000), pp. 433–436. ISSN: 01692070. DOI: 10.1016/S0169-2070(00)00078-9. URL: https://doi.org/10.1016/S0169-2070(00)00078-9%20http://linkinghub.elsevier.com/retrieve/pii/S0169207000000789.