UNIVERSITAT DE BARCELONA

Undergraduate Thesis

MAJOR IN MATHEMATICS and BUSINESS ADMINISTRATION

Introduction to Time Series and Forecasting

Eloi Castaño Camps

Advisors: **Dr. Josep Vives Santa-Eulàlia**
Dept. de Matemàtiques i Informàtica
Dr. Javier Martínez de Albéniz
Dept. de Matemàtica Econòmica, Financera i Actuarial

Barcelona, June 2022

Abstract

Time series analysis allows complex processes to be expressed in simple terms to understand how these processes were generated and to predict future values. SARIMA models assume that the observations of a process depend on the previous observations and the variation between them to give an expression of the underlying data generating process.

To find the SARIMA model that better fits our data we introduce the Box and Jenkins method, based on three iterative steps: model identification, parameter estimation and fitness check. Once we have identified the most appropriate fitting model, we use it to forecast future values.

We have followed this methodology to find the model that best fits the Spanish unemployment series from 2002 to the first quarter of 2022 and to forecast the next 8 observations.

Resum

L'anàlisi de sèries temporals permet expressar processos complexos en termes simples per entendre com s'han generat aquests processos i predir valors futurs. Els models SARIMA suposen que les observacions d'un procés depenen de les observacions anteriors i de la variació entre aquestes per donar una expressió del procés generador de dades subjacent.

Per a trobar el model SARIMA que més s'ajusta a les nostres dades presentem el mètode de Box i Jenkins, basat en tres passos iteratius: la identificació de models, l'estimació dels paràmetres i la comprovació de l'ajust. Quan s'ha identificat el model que millor s'ajusta a les dades, s'utilitza per predir valors futurs.

Hem seguit aquesta metodologia per trobar el model que millor s'ajusta a la taxa d'atur espanyola des del 2002 fins al primer trimestre del 2022 i predir les 8 observacions següents.

2020 Mathematics Subject Classification. 62M10, 91B84

Acknowledgements

I would like to express my gratitude to my advisors, Josep Vives Santa-Eulàlia and Javier Martínez de Albéniz for their guide throughout this project and the time they have devoted to it.

I would also like to thank my family, friends and colleagues for their support and encouragement during the degree.

Contents

Introduction

Since the irruption of digital technologies, the amount of data collected has increased drastically. In recent years, smartphones and Internet of Things devices (amongst others) have allowed generating lots of data that we were not able to record before. The analysis of the data can help us spot trends or find correlations between different sets of data. This is useful for different scientific branches, business staff and governments to make decisions based on predictive analysis of the data.

In this project, we focus on the analysis of time series. Time series are series of observations recorded at a specific time or, in other words, sequences of data taken at equidistant points of time. The analysis of time series allows extracting meaningful statistics and other characteristics, such as trends and patterns, that can be attributed to dependency relationships among observations. The nature of these dependencies between observations has practical interest. Time series analysis considers different methods to study this dependence and models that can be used to understand the underlying forces that generated the observed data and to forecast future observations of the series.

Forecasting refers to the prediction of some future event or condition as a result of a study and analysis of available data.[1] Even though the prediction is based on statistical studies and not on guesses, it is almost impossible to forecast an exact value for a future observation. There are multiple forecasting methods and models and the selection of a particular one should be based on their expected accuracy and the previous analysis of the data.

The observation of regularities and/or trends in data is very old, but a scientific study taking into account the knowledge of statistics can be traced at the beginning of XX century, with the works of G. U. Yule [17] and G. Walker [13]. They made important contributions tot the theory and practice of correlation and regression and the definition of the autoregressive model.

It was not until 1970 when G. E. P. Box and G. Jenkins published the book "Time Series Analysis: Forecasting and Control" [5] that a method to estimate the parameters of the models in terms of likelihood was defined. They also described a method to find the best ARIMA model fitting a time series based on three iterative stages: identification of feasible models, estimation of their parameters and checking the fitness of the models.

The progress in the analysis of time series has also been tied to technological improvements. They have allowed to record, store and make accessible large data sets as well as to make the calculations easier thanks to computation.

[1] `https://www.merriam-webster.com/dictionary/forecasting`

About this work

The first chapter is devoted to introduce introduce some basic concepts of time series needed to understand the different models described later and some examples of series.

In subsequent chapters we describe the *autoregressive moving average models* (ARMA), used to express stationary time series in terms of polynomials. These models can be expanded to include non-stationary series. The *autoregressive integrated moving average models* (ARIMA) eliminate the non-stationarity in mean terms of time series through differentiation to fit an ARMA model on them and the *seasonal autoregressive integrated moving average models* (SARIMA) also take into account the seasonal behavior of the data.

Finally, in Chapter 4, we explain the Box and Jenkins method to find the best fitting model to a time series. We use this methodology in Chapter 5 to fit a SARIMA model to the Spanish unemployment rate, one of the main macroeconomic indicators of a country, and forecast some of their future values.

To perform the necessary calculations to apply the different methods, we use the R programming language, a language designed for statistical computing and graphics. It is an open-source implementation of the S programming language developed at Bell Laboratories. R was designed by R. Ihaka and R. Gentleman in 1993 (see `https://www.r-project.org/`).

Chapter 1

Time Series

In this chapter, we introduce the basic ideas of time series needed to understand the concepts described later. One of this concepts is the *stationarity* of time series. Stationary series are processes which some of their properties do not vary with time. We also introduce the *autocorrelation function* and *partial autocorrelation function*, which will be used later to identify the underlying process generating the observations of a time series.

Finally, we present some examples of time series, such as the white noise processes, the random walk processes and an actual economic time series.

In this chapter we follow the books by G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Ljung [7], P. J. Brockwell and R. A. Davis [9], [10] and P. S. P. Cowpertwait and A. V. Metcalfe [11], as in the subsequent chapters.

1.1 Time series and stochastic processes

A *time series* is a set of observations x_t generated sequentially over time t. We can distinguish between continuous time series if the observations are recorded continuously over time and discrete time series if the set of observations is discrete. In this thesis we consider only discrete time series where the observations are recorded at fixed intervals of time.

We can also differentiate the time series whose future values can be exactly determined by a mathematical functions, which we call *deterministic time series*, from the time series that have some random component which does not allow us to explicitly describe their behavior with an analytical expression, which we call *non-deterministic time series*.

In order to forecast future values of non-deterministic time series we have to assume that there is a probability model that generates the observations of the time series.

Definition 1.1. A *discrete time stochastic process* is a sequence of random variables $\{X_t\}$ defined over time $t \in \mathbb{Z}$.

This means that at every time t there is a random variable $\{X_t\}$ that will take different values depending on its probability distribution.

A stochastic process generates an infinite set of time series that could be observed. We can think of the time series that we analyze as a particular realization of a stochastic

process. The analysis of time series that we do in this thesis consists of deducing the stochastic process that has generated our time series from the observations that we have of the time series.

1.2 Stationarity of time series

When forecasting, we assume that some properties of the time series are maintained over time and that we can extrapolate them to the future. For example, if we detect that the observations of the time series tend to increase around a fixed interval with each observation since the beginning of the series or if the observations are always around the same value or that a change on the trend of the observations always implies a similar variation on the future data, it is not daring to think that this characteristics will also be present on future observations. Let's define these properties and the time series whose properties are constant over time.

Definition 1.2. Let $\{X_t\}$ be a time series. The *mean function* of $\{X_t\}$ is defined as

$$\mu(t) = E(X_t) \quad t = 1, 2, \ldots, n,$$

where $E(X_t)$ is the expected value of the random variable X_t.

Now we define the covariance function between two random variables of our time series.

Definition 1.3. Let $\{X_t\}$ be a time series. The *covariance function* of $\{X_t\}$ is

$$\begin{aligned}\gamma(t, t+h) &= \mathrm{Cov}(X_t, X_{t+h}) \\ &= E[(X_t - \mu(t)) \cdot (X_{t+h} - \mu(t+h))],\end{aligned}$$

where $t = 1, 2, \ldots, n$ and $h = 1, 2, \ldots, n - t$.

This function indicates the degree of association between two variables. If the value of X_{t+h} tends to be high when X_t is high or the value of X_{t+h} tends to be small when X_t is small, the value of the covariance is positive and different from zero. On the other hand, if the values of X_{t+h} tends to be high when X_t is small or vice-versa, the value of the covariance is negative and different from zero. Finally, if there is no relation between the two variables, then the value of the covariance is zero or near to zero.

Definition 1.4. Let $\{X_t\}$ be a time series. $\{X_t\}$ is *strictly stationary* if $(X_1, \ldots, X_n)$ and $(X_{1+h}, \ldots, X_{n+h})$ have the same joint distributions for all h.

That is to say that a time series is strictly stationary if its distribution is unchanged after any arbitrary time shift.

This definition of stationary is quite restrictive. Therefore, we define the less stringently *weakly stationary.*

Definition 1.5. Let $\{X_t\}$ be a time series. $\{X_t\}$ is *weakly stationary* if

1. $E(X_t^2) < \infty$	for all $t \in \mathbb{Z}$,
2. $\mu(r) = \mu(s)$	for all $r, s \in \mathbb{Z}$,
3. $\gamma(r, r+h) = \gamma(s, s+h)$	for all r, s and $h \in \mathbb{Z}$.

In other words, a time series is weakly stationary if its second-moment is finite, its mean is constant and its covariance depends only on the distance between observations, known as *lag*. From now on, we refer to weakly stationary series as stationary series and, when talking about stationary series, we refer to the mean as μ instead of $\mu(t)$, since it does not depend on time.

1.3 Correlation

We can see that, on stationary time series, since the covariance only depends on the lag h, we can define the covariance function of these time series with only one variable. This function is known as the *autocovariance function*.

Definition 1.6. Let $\{X_t\}$ be a stationary time series. The *autocovariance function* (ACVF) of $\{X_t\}$ at lag h is

$$\gamma(h) = \text{Cov}(X_t, X_{t+h}) = E[(X_t - \mu) \cdot (X_{t+h} - \mu)].$$

Notice that $\gamma(0) = E[(X_t - \mu)^2] = \sigma^2$ is the variance of the time series.

It's easy to check that $\gamma(0) \geq 0$, since $\gamma(0) = \text{Var}(X_t) \geq 0$ and that $|\gamma(h)| \leq \gamma(0)$ for all h, since from the Cauchy-Schwarz inequality

$$|\gamma(h)|^2 = (E\left[(X_t - \mu) \cdot (X_{t+h} - \mu)\right])^2 \leq E\left[(X_t - \mu)^2\right] \cdot E\left[(X_{t+h} - \mu)^2\right] = \gamma(0)^2.$$

From this definition of the autocovariance function, we can describe the autocorrelation function.

Definition 1.7. Let $\{X_t\}$ be a stationary time series. The *autocorrelation function* (ACF) of $\{X_t\}$ at lag h is

$$\rho(h) = \frac{\gamma(k)}{\gamma(0)}.$$

We can see that $\rho(0) = 1$ and, from the properties of the autocovariance function, $|\rho(h)| \leq 1$.

On some time series, the observations X_t and X_{t+h} are correlated because they are correlated with other observations between them even though there is no direct correlation between X_t and X_{t+h}. For example, there may be a correlation between the observations X_1 and X_3 because they are correlated with the observation X_2. The *partial autocorrelation function* gives the correlation between variables without taking into account the middle observations between them. We formally define this function on Section 2.2 because we need some concepts that we present on that chapter. On stationary time series we can also define this function in terms of the lag between variables.

The *correlograms* are the chart representation of the autocorrelations as a function of the lags. When we talk of the correlograms we are referring to the ACF correlogram and the PACF correlogram. These charts are useful for the model identification process on the Box-Jenkins method described on the Chapter 5.

Let's see some examples of time series, and analyze their plots, their main characteristics described on this chapter and their correlograms.

1.4 Some examples

This section is devoted to show several useful examples of time series and their characteristics.

Example 1.1. *White Noise.* This time series consist of a sequence of uncorrelated random variables $\{X_t\}$ with mean $\mu = 0$ and variance σ^2. This series is the simplest example of a stationary time series. If the random variables are independent and identically distributed, the series is called *IID noise* and if they follow a normal distribution, the series is called *Gaussian white noise*, which is an example of IID noise. Let's generate a Gaussian white noise series in R of 100 observations and $\sigma^2 = 1$. The series is plotted on Figure 1.1 and its correlograms are plotted on Figure 1.2. The code to generate them can be found on the Annex I on page 1.

Since the random variables are independent, they are not correlated, so its autocovariance function is σ^2 at lag 0 and 0 at lags $h > 0$, its autocorrelation function is 1 at lag 0 and 0 at lags $h > 0$ and its partial autocorrelation function is 0 at all lags.

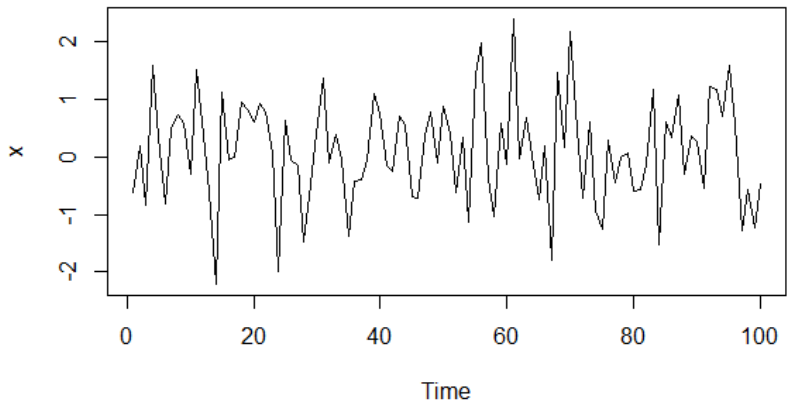

Figure 1.1: Plot of a Gaussian white noise series of 100 observations and $\sigma^2 = 1$.

On the correlograms we can see that the autocorrelations are not exactly 0. This is because of sampling variation. The dashed blue lines on the correlograms indicate the confidence intervals for the autocorrelations to be 0 with confidence level 95%. This means that the values lying within this interval are not *statistically significant* with confidence level 95%. By default R assumes that the series is a Gaussian white noise and shows the interval $[-1.96/\sqrt{n}, 1.96/\sqrt{n}]$, since 1.96 is the 0.975 quantile of the Gaussian distribution with mean 0 and $\sigma^2 = 1$.

Example 1.2. *Random Walk.* This time series is obtained by sequentially adding independent and identically distributed random variables to the observations of the time series.

Definition 1.8. Let $\{X_t\}$ be a time series and $\{W_t\}$ an IID noise time series. $\{X_t\}$ is a *random walk* if

$$\begin{aligned} X_1 &= W_1, \\ X_t &= X_{t-1} + W_t, \qquad \text{if } t > 1. \end{aligned}$$

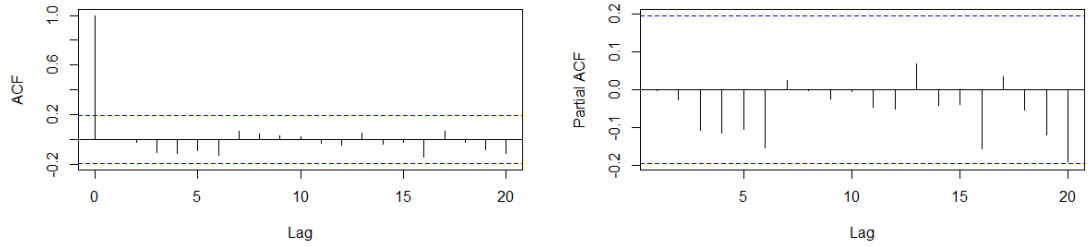

Figure 1.2: Correlogram (left) and partial correlogram (right) of a Gaussian white noise series of 100 observations and $\sigma^2 = 1$.

Notice that sequentially substituting X_{t-1} on the equation, a random walk can also be defines as

$$X_t = W_1 + W_2 + \cdots + W_t.$$

The mean of this time series is $\mu(t) = 0$ and its covariance is

$$\text{Cov}(X_t, X_{t+h}) = \text{Cov}\left(\sum_{i=1}^{t} W_i, \sum_{j=1}^{t+h} W_j\right) = \sum_{i=1}^{t} \text{Cov}(W_i, W_i) = t\sigma^2,$$

since $\text{Cov}(W_i, W_j) = 0$ if $i \neq j$ and $\text{Cov}(W_i, W_i) = \text{Var}(W_i) = \sigma^2$. Therefore, as the covariance depends on the time, this process is not stationary.

Random walks can be generated on R using the code on Annex I on page 1.

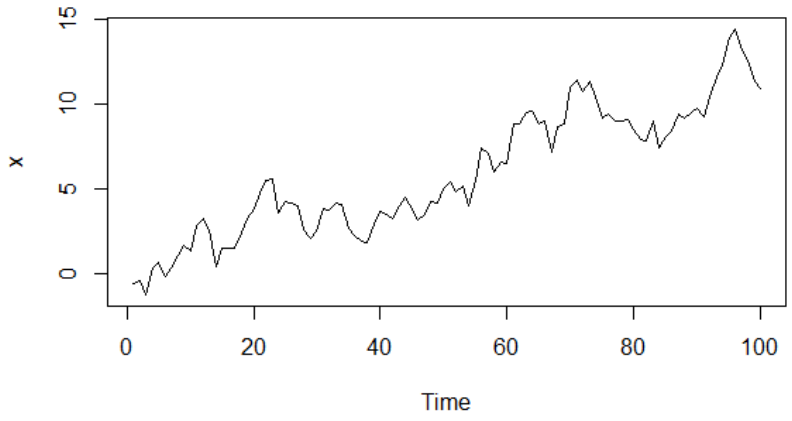

Figure 1.3: Plot of a random walk of 100 observations.

Looking at the plot of the generated series, we can see that there is an increasing trend. This pattern is explained due to the high serial correlation of the series and the randomness involved on generating this time series. Changing the seed on the R code shows that the increasing trend is not a characteristic of random walks but from this realization.

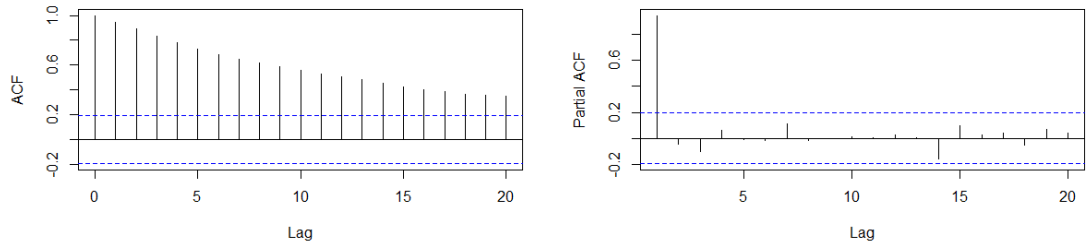

Figure 1.4: Correlogram (left) and partial correlogram (right) of a random walk of 100 observations.

The correlogram of this series starts at 1 and slowly tends to 0 and the partial correlogram has only one significant value at lag 1. Random walks are the special case of non-stationary AR(1) process (explained on Chapter 2) and these patterns on the correlograms are characteristic of this kind of series.

Example 1.3. *Quarterly earnings per Johnson & Johnson share*

Finally, let's see an example of real data. The time series of quarterly earnings in US dollars per Johnson & Johnson share from 1960 to 1980 is one of the data sets implemented on the default R packages and can be accessed using the call `JohnsonJohnson`.

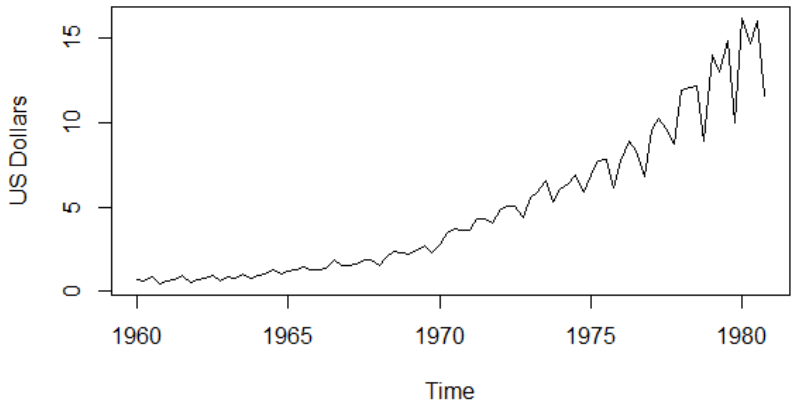

Figure 1.5: Plot of quarterly earnings per Johnson & Johnson share from 1960 to 1980. Source: R "datasets" package.

In Figure 1.5 we can see that the series follows an increasing trend, that there is a pattern on the data that repeats over time each year known as *seasonal component* and that the variance increases over time. Therefore, we can conclude that this time series is not stationary.

Chapter 2

Stationary processes

Stationary processes are series which some of their properties do not vary with time.

In this chapter, we introduce a representation for stationary processes as a linear combination of random variables and the concepts of causality and invertibility based on this representation.

We also define the autoregressive models (AR),that allows us to express the time series in terms of the previous observations and a random component; the moving average models (MA), that allows us to express the time series in terms of the current random component and the previous random components of the series; and the autoregressive moving average models (ARMA), that consider both dependencies at the same time. These methods were first defined by P. Whittle in 1951 [14] and were popularized by G. E. P. Box and G. Jenkins in 1970 [5].

2.1 Linear processes

According to Yule's [17] and Wold's [15] studies, all stationary time series $\{X_t\}$ can be characterized as linear processes. This is that they can be represented as

$$X_t = \mu + \sum_{i=-\infty}^{\infty} \psi_i W_{t-i}, \qquad \text{for all } t,$$

where $\mu \in \mathbb{R}$, $\{\psi_i\}$ is an absolutely summable sequence of constants and $\{W_t\}$ is a white noise series with mean 0 and variance σ^2. For a linear process to be stationary it's necessary that the series is $\{\psi_i\}$ is absolutely summable.[1]

Defining the *backward shift operator*, B, as $BX_t = X_{t-1}$ and $B^i X_t = X_{t-i}$, linear processes can also be represented as

$$X_t = \mu + \psi(B) W_t,$$

where $\psi(B) = \sum_{i=-\infty}^{\infty} \psi_i B^i$.

[1] Recall that a series $\sum_{n=-\infty}^{\infty} a_n$ is absolutely summable if $\sum_{n=-\infty}^{\infty} |a_n| < \infty$

We can see that the mean of a linear process $\{X_t\}$ is μ and its covariance function is

$$\gamma(h) = \text{Cov}(X_t, X_{t+h}) = \text{Cov}\left(\mu + \sum_{i=-\infty}^{\infty} \psi_i W_{t-i},\ \mu + \sum_{i=-\infty}^{\infty} \psi_i W_{t+h-i}\right)$$
$$= \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \,\text{Cov}(W_{t-i}, W_{t-i}) = \sigma^2 \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h},$$

since $\text{Cov}(W_{t-i}, W_{t+h-i}) = 0$ if $t - i \neq t + h - i$.

A linear process is said to be *causal* if X_t can be expressed as a linear combination of present and past values of the white noise W_s (such that $s \leq t$) and therefore is uncorrelated with the future observations of W_s (such that $s > t$). This property is formally described below.

Definition 2.1. A linear process is *causal* or *a causal function of* $\{Z_t\}$ if there exist constants $\{\psi_i\}$ such that $\sum_{i=0}^{\infty} |\psi_i| < \infty$ and

$$X_t = \sum_{i=0}^{\infty} \psi_i W_{t-i} \text{ for all } t.$$

On the other hand, a linear process is *invertible* if the white noise series $\{W_t\}$ can be represented as a causal function of $\{X_t\}$. This property is formally described below.

Definition 2.2. A linear process is *invertible* if there exist constants $\{\pi_i\}$ such that $\sum_{i=0}^{\infty} |\pi_i| < \infty$ and

$$W_t = \sum_{i=0}^{\infty} \pi_i X_{t-i} \text{ for all } t.$$

2.2 AR processes

Autoregressive models are based on the idea that the current value of the process can be expressed as a combination of the p previous observations of the series plus a random component.

Definition 2.3. Let $\{X_t\}$ be a time series and $\{W_t\}$ a white noise series. An *autorregressive model of order p* (or AR(p)) is defined as

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t,$$

where ϕ_i are constants for $i = 1, \ldots, p$ and $\phi_p \neq 0$.

Using the backward shift operator, the process can also be expressed as

$$\phi(B) X_t = W_t,$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$.

For the AR(1) process

$$X_t = \phi X_{t-1} + W_t,$$

the condition of invertibility is always fulfilled and the condition of stationarity is equivalent to the condition that $|\phi| < 1$. Notice that if $|\phi| > 1$ we can rewrite the process as

$$X_{t-1} = \phi^{-1} X_t - \phi^{-1} W_t,$$

or

$$X_t = \phi^{-1} X_{t+1} - \phi^{-1} W_{t+1}.$$

A simple manipulation leads us to obtain

$$X_t = -\sum_{i=1}^{\infty} \phi^{-i} W_{t+i},$$

and since $|\phi^{-1}| < 1$, the process is stationary but not causal.

We can see that the only non-stationary AR(1) process is the random walk defined on Example 1.2. From now on, we consider only causal autoregressive processes.

Taking expectations on the representation of the AR(1) process, we can see that its mean is:

$$\mu = E(X_t) = E(\phi X_{t-1} + W_t) = \phi E(X_{t-1}) + E(W_t),$$

and, since $\{W_t\}$ is a white noise sequence with mean 0, $\{X_t\}$ is stationary and $\phi \neq 0$, $E(X_t) = \phi E(X_t)$ implies that $\mu = 0$.

The autocovariance function of the AR(1) process is

$$\gamma(h) = \text{Cov}(X_t, X_{t-h}) = \text{Cov}(\phi X_{t-1}, X_{t-h}) + \text{Cov}(W_t, X_{t-h}) = \phi\, \gamma(h-1) + 0.$$

Iterating the process we have that $\gamma(h) = \phi^h \gamma(0)$ and therefore the autocorrelation function of the AR(p) process is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h.$$

The correlogram of this process starts at 1 at lag 0 and geometrically decreases to 0 and the partial autocorrelogram has a significant value at lag 1 and the rest are 0. If $\phi < 0$ the values of the ACF alternate between positive and negative. Let's see the correlograms of a generated AR(1) process with $\phi = 0.9$ and $\phi = -0.9$. The code used to generate the processess and the graphs can be found on Annex I, page 2.

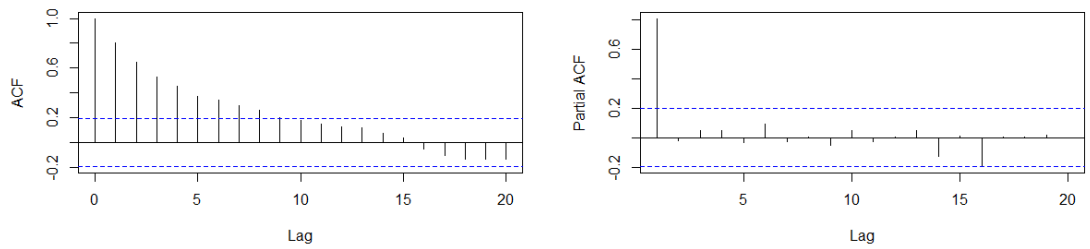

Figure 2.1: Correlogram (left) and partial correlogram (right) of an AR(1) process with $\phi = 0.9$.

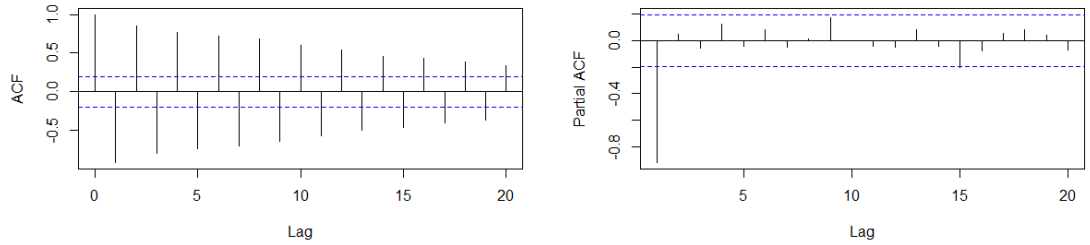

Figure 2.2: Correlogram (left) and partial correlogram (right) of an AR(1) process with $\phi = -0.9$.

For the general AR(p) process, since $\phi(B)$ is finite, all autoregressive processes are invertible. To check the stationary condition, we can write the process as:

$$X_t = \frac{1}{\phi(B)} W_t = \psi(B) W_t,$$

and we can see that for the process to exist and to be stationary, $\phi(B)$ must not have roots on the unit circle and to also be causal, the roots of $\phi(B)$ have to lie outside of the unit circle.

The autocorrelation function of a stationary autoregressive process can be obtained multiplying its formula by X_{t-h} for $h > 0$:

$$X_t X_{t-h} = \phi_1 X_{t-1} X_{t-h} + \cdots + \phi_p X_{t-p} X_{t-h} + W_t X_{t-h},$$

and taking the expected values on each side of the equation:

$$\gamma(h) = \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p) + E(W_t X_{t-h}),$$

where $E(W_t X_{t-h}) = 0$ since X_{t-h} can only be related to white noises up to time $t-h$. Now, dividing by $\gamma(0)$ we find the autocorrelation function

$$\rho(h) = \phi_1 \rho(h-1) + \cdots + \phi_p \rho(h-p).$$

The autocovariance function of an AR(p) process decreases to 0 geometrically as lags increase if the polynomial $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^P$ has no complex roots or converges to 0 following a sinusodial function if it has complex roots.

Substituting h on the autocorrelation function by $1, \ldots, p$ we obtain a set of linear equations for $\phi_1, \ldots, \phi_p$ in terms of $\rho(1), \ldots, \rho(p)$. These equations are known as *Yule-Walker equations* (see [17] and [13]).

$$\begin{array}{lcllll} \rho(1) & = & \phi_1 & + & \phi_2 \rho(1) & + \cdots + \phi_p \rho(p-1), \\ \rho(2) & = & \phi_1 \rho(1) & + & \phi_2 & + \cdots + \phi_p \rho(p-2), \\ \vdots & & \vdots & & \vdots & \vdots \\ \rho(p) & = & \phi_1 \rho(p-1) & + & \phi_2 \rho(p-2) & + \cdots + \phi_p. \end{array} \tag{2.1}$$

Notice that we can rewrite the equations in a matrix form

$$\rho(p) = P(p)\phi,$$

where $\rho = (\rho(1), \ldots, \rho(p))^T$, $\phi = (\phi_1, \ldots, \phi_p)^T$ and P is the matrix

$$P = \begin{pmatrix} 1 & \rho(1) & \rho(2) & \ldots & \rho(p-1) \\ \rho(1) & 1 & \rho(1) & \ldots & \rho(p-2) \\ \rho(2) & \rho(1) & 1 & \ldots & \rho(p-3) \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \rho(p-1) & \rho(p-2) & \rho(p-3) & \ldots & 1 \end{pmatrix}.$$

These equations are useful because they allow us to estimate the ϕ_i parameters replacing the theoretical autocorrelations $\rho(h)$ by their sample values.

Now we can formally define the *partial autocorrelation function* introduced on Section 1.3 as a function of the autocorrelations.

Definition 2.4. Let $\{X_t\}$ be a time series and ϕ_{ki} be the i-th coefficient in an autoregressive representation of order h of $\{X_t\}$.[2] We define the *partial autocorrelation function of* $\{X_t\}$ *at lag* h (PACF) as $\alpha(h) = \phi_{hh}$.

From the definition, we can see that the PACF of an AR(p) process is different from 0 if $h \leq p$ and 0 if $h > p$.

2.3 MA processes

Moving average models are based on the idea that the current value of the process can be expressed as a linear combination of the current white noise term and the q most recent past white noise terms.

Definition 2.5. Let $\{X_t\}$ be a time series and $\{W_t\}$ a white noise series. A *moving average model of order* q (or MA(q)) is defined as

$$X_t = W_t - \theta_1 W_{t-1} - \cdots - \theta_q W_{t-q},$$

where θ_i are constants for $i = 1, \ldots, q$ and $\theta_q \neq 0$.

Using the backward shift operator, the process can also be expressed as

$$X_t = \theta(B) W_t,$$

where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$.

Observations. *On the R language, the moving average models are defined using positive signs instead of negative signs between the coefficients*[3]*. This flips the signs of estimated coefficient values and* θ *terms in formulas like the ACF and PACF. We have to consider this notation when we use the software.*

[2] Recall that the set of ϕ_{hi} fulfill the equations

$$\rho(h) = \phi_{h1}\rho(h-1) + \cdots + \phi_{h(h-1)}\rho(j-h+1) + \phi_{hh}\rho(j-h) \quad \text{for } j = 1, \ldots, h.$$

(see the Yule-Walker equations (2.1))

[3] `https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/arima`

For the MA(1) process

$$X_t = W_t - \theta W_{t-1},$$

the condition of stationarity is always fulfilled. Rewriting the equation using the backwards shift operator B it is easy to check that the condition of invertibility is equivalent to the condition $|\theta| < 1$:

$$X_t = (1-\theta)W_t,$$
$$W_t = \frac{1}{1-\theta B}X_t,$$

where the root of $(1-\theta B)$ must lie outside the unit circle, meaning that $|\theta|$ must be less than 1.

The mean of the MA(1) process is 0 since it is a sum of zero mean white noises. Its autocovariance function is

$$\begin{aligned}\gamma(0) &= \mathrm{Var}(X_t) = E(W_t^2 + \theta^2 W_{t-1}^2 - 2\theta W_t W_{t-1}) \\ &= E(W_t^2) + \theta^2 E(W_{t-1}^2) - 2\theta E(W_t W_{t-1}) = \sigma^2(1+\theta^2),\end{aligned}$$

$$\begin{aligned}\gamma(1) &= \mathrm{Cov}(X_t, X_{t-1}) = E[(W_t - \theta W_{t-1})(W_{t-1} - \theta W_{t-2})] \\ &= E(W_t W_{t-1}) - \theta E(W_t W_{t-2}) - \theta E(W_{t-1}^2) + \theta^2 E(W_{t-1} W_{t-2}) \\ &= -\theta E(W_{t-1}^2) = -\theta\sigma^2,\end{aligned}$$

$$\begin{aligned}\gamma(h) &= \mathrm{Cov}(X_t, X_{t-h}) \\ &= E[(W_t - \theta W_{t-1})(W_{t-h} - \theta W_{t-h-1})] = 0, \qquad \text{if } h > 1.\end{aligned}$$

Hence its autocorrelation function is

$$\rho(h) = \begin{cases} \dfrac{\sigma^2\theta}{\sigma^2(1+\theta^2)} = \dfrac{-\theta}{1+\theta^2}, & \text{if } h = 1, \\ 0, & \text{if } h > 1. \end{cases}$$

Finally, substituting the autocorrelation function on the Yule-Walker equations (2.1), we get the partial autocorrelation function

$$\alpha(h) = \frac{-\theta^h(1-\theta^2)}{1-\theta^{2(h+1)}}.$$

This function gradually decreases to 0 and, if θ is positive, the function is negative whereas if θ is negative, the sign of the function alternates. In Figure 2.3 and Figure 2.4 we can see the correlograms of a generated MA(1) process with $\theta = 0.9$ and $\theta = -0.9$ respectively as an example. The code used to generate the processess and the graphs can be found on Annex I page 2.

For the general MA(q) process, since $\theta(B)$ is finite, all moving average processes are stationary. To check the invertibility condition, we can write the process as:

$$W_t = \frac{1}{\theta(B)}X_t = \pi(B)X_t,$$

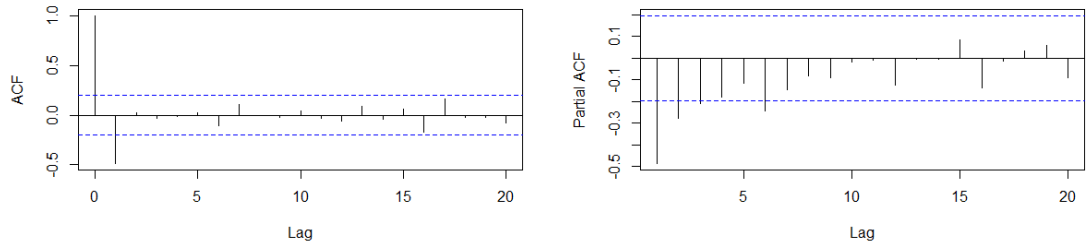

Figure 2.3: Correlogram (left) and partial correlogram (right) of an MA(1) process with $\theta = 0.9$.

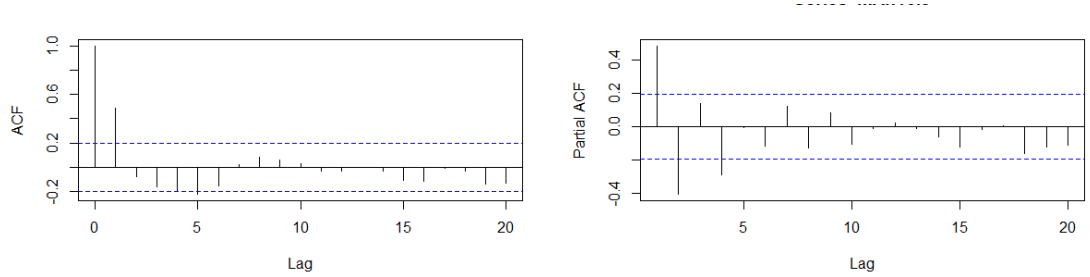

Figure 2.4: Correlogram (left) and partial correlogram (right) of an MA(1) process with $\theta = -0.9$.

and we can see that the roots of the polynomial $\theta(B)$ must lie outside the unit circle for the process to be invertible.

The autocovariance function of a moving average process can be obtained as we did for the AR(p) process, multiplying by X_{t-h} for $h > 0$ and taking the expected values on each side of the equation:

$$\begin{aligned}\gamma(h) &= E[X_t X_{t-h}] \\ &= E[(W_t - \theta_1 W_{t-1} - \cdots - \theta_q W_{t-q})(W_{t-h} - \theta_1 W_{t-h-1} - \cdots - \theta_q W_{t-h-q})] \\ &= -\theta_h E(W_{t-k}^2) + \theta_1\theta_{h+1} E(W_{t-h-1}^2) + \cdots + \theta_{q-h}\theta_q E(W_{t-q}^2),\end{aligned}$$

since W_t are uncorrelated.

In conclusion

$$\gamma(h) = \begin{cases} (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2, & \text{for } h = 0, \\ (-\theta_h + \theta_1\theta_{h+1} + \theta_2\theta_{h+2} + \cdots + \theta_{q-h}\theta_q)\sigma^2, & \text{for } h = 1, \ldots, q, \\ 0, & \text{for } k > q. \end{cases}$$

Dividing by $\gamma(0)$ we get the autocorrelation function

$$\rho(h) = \begin{cases} \dfrac{-\theta_h + \theta_1\theta_{h+1} + \cdots + \theta_{q-h}\theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2}, & \text{for } k = 1, \ldots, q, \\ 0, & \text{for } k > q. \end{cases}$$

The partial autocorrelation formula of an MA(q) process decays to 0 as the lag increases.

2.4 ARMA processes

The autoregressive model and the moving average model take into account different kinds of dependencies between observations over time. We can consider both dependencies at the same time on a unique autoregressive moving average process.

Definition 2.6. Let $\{X_t\}$ be a time series and $\{W_t\}$ a white noise series. An *autoregressive moving average process of order* (p, q) (or ARMA(p, q)) is defined as

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t - \theta_1 W_{t-1} - \cdots - \theta_q W_{t-q},$$

where ϕ_i are constants for $i = 1, \ldots, p$, $\phi_p \neq 0$, θ_j are constants for $j = 1, \ldots, q$ and $\theta_q \neq 0$.

Using the backward shift operator, the process can be expressed as

$$\phi(B)X_t = \theta(B)W_t,$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$.

Writing the process as

$$X_t = \frac{\theta(B)}{\phi(B)} W_t, \quad \text{or} \quad \frac{\phi(B)}{\theta(B)} X_t = W_t,$$

we can see that $\phi(B)$ and $\theta(B)$ do not have common factors and that the conditions of stationarity, causality and invertibility are the same for $\phi(B)$ and $\theta(B)$ that the ones on pure autoregressive models and pure moving average models respectively.

In particular, we can think an ARMA(p, q) process as an AR(p) process $\phi(B)X_t = Y_t$ where Y_t follows a MA(q) process $Y_t = \theta(B)W_t$ or as a MA(q) process $X_t = \theta(B)Y_t$ where Y_t follows an AR(p) process $\phi(B)Y_t = W_t$

The autocorrelation function and the partial autocorrelation function of an ARMA (p, q) process both gradually decrase to 0 geometrically or following a sinus wave, depending on p and q and the sign of the parameters.

For the ARMA$(1, 1)$ process

$$X_t = \phi X_{t-1} + W_t - \theta W_{t-1},$$

the stationary condition is $\phi \neq 1$, the causal condition is $|\phi| < 1$ and the invertibility condition is $|\theta| < 1$.

The autocovariance function is

$$\begin{aligned}\gamma(0) &= E[(\phi X_{t-1} + W_t - \theta W_{t-1})^2] = \phi^2 E(X_{t-1}^2) + E(W_t^2) + \theta^2 E(W_{t-1}^2) \\ &+ 2\phi E(X_{t-1}W_t) - 2\phi\theta E(X_{t-1}W_{t-1}) - 2\theta E(W_t W_{t-1}) \\ &= \phi^2\gamma(0) + \sigma^2 + \theta^2\sigma^2 - 2\phi\theta\sigma^2 = \phi^2\gamma(0) + \sigma^2(1 + \theta^2 - 2\phi\theta) \\ &= \frac{\sigma^2(1 + \theta^2 - 2\phi\theta)}{1 - \phi^2},\end{aligned}$$

$$\begin{aligned}\gamma(1) &= E[(\phi X_{t-1} + W_t - \theta W_{t-1})X_{t-1}] \\ &= \phi E(X_{t-1}^2) + E(W_t X_{t-1}) - \theta E(W_{t-1}X_{t-1}) = \phi\gamma(0) - \theta\sigma^2 \\ &= \frac{\sigma^2(1 - \phi\theta)(\phi - \theta)}{1 - \phi^2},\end{aligned}$$

$$\begin{aligned}\gamma(h) &= E(X_t X_{t-h}) = E[(\phi X_{t-1} + W_t - \theta W_{t-1})X_{t-h}] \\ &= \phi E(X_{t-1}X_{t-h}) + \phi E(W_t X_{t-h}) - \theta E(W_{t-1}X_t t - h) \\ &= \phi\gamma(h-1) = \phi^{h-1}\gamma(1), \qquad \text{if } h > 1.\end{aligned}$$

Hence, the autocorrelation function is

$$\begin{aligned}\rho(1) &= \frac{(1 - \phi\theta)(\phi - \theta)}{1 + \theta^2 - 2\phi\theta}, \\ \rho(h) &= \phi\rho(h-1) = \phi^{h-1}\rho(1), \qquad \text{for } h > 1.\end{aligned}$$

So the correlogram decreases geometrically from lag 2.

The partial autocorrelation function of an ARMA$(1,1)$ is $\rho(1)$ at lag 1 and then behaves like the PACF of a MA(1) process (see Fig.2.3 and Fig.2.4). Let's see the correlograms of a generated ARMA$(1,1)$ with parameters $\phi = 0.9$ and $\theta = 0.7$. The code used to generate the process and the graphs can be found on Annex I, page 2.

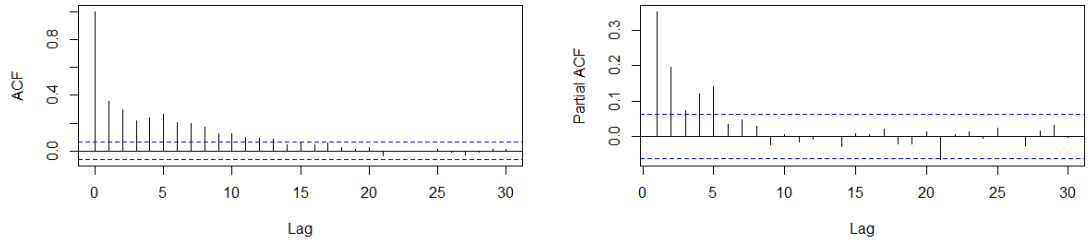

Figure 2.5: Correlogram (left) and partial correlogram (right) of an ARMA$(1,1)$ process with $\phi = 0.9$ and $\theta = 0.7$.

Chapter 3

Non-stationary Models

Some time series are not stationary because of trends or seasonal effects. The non-stationary series due to trends can be transformed into stationary series by differentiating them. Once differentiated we can fit an ARMA process on them. These processes are known as *autoregressive integrated moving average processes* (or ARIMA) since the differentiated series needs to be summed or integrated to recover the original series. ARIMA models were introduced by A. M. Yaglom (1955) [16] and expanded among other authors by G. E. P. Box and G. M. Jenkins (1962) [4].

The seasonal component of a time series is a change in the observations that is repeated cyclically over time at the same frequency. The ARIMA models can be extended to include the analysis of the seasonal component. This is done by considering additional parameters and differentiation for the seasonal period. The ARIMA models that include the seasonal parameters are known as *seasonal autoregressive integrated moving average models* (or SARIMA).

3.1 ARIMA processes

A time series that is not stationary in terms of mean can be differentiated as many times as needed until it is stationary. It is done by subtracting the previous observation to the current observation. We can define the *differential operator* ∇ as

$$\nabla := (1 - B), \qquad \text{and} \qquad \nabla X_t = (1 - B)X_t = X_t - X_{t-1},$$

where B is the backwards shift operator.

As an example, let $\{X_t\}$ be a random walk. This process was defined in Example 1.2 and can be expressed as

$$X_t = X_{t-1} + W_t,$$

where $\{W_t\}$ is a white noise series.

We saw that this process is not stationary, but if we differentiate it:

$$\nabla X_t = X_{t-1} + W_t - X_{t-1} = W_t,$$

so the differentiated process is stationary.

A series is *integrated* of order d if it is not stationary but its d difference is stationary. Fitting an ARMA to an integrated process is known as fitting an ARIMA model.

Definition 3.1. Let $\{X_t\}$ be a time series and d a non negative integer. $\{X_t\}$ is an *autoregressive integrated moving average processes of order* (p, d, q) (or ARIMA(p, d, q) if $Y_t := (1 - B)^d X_t$ is a causal ARMA(p, q) process.

Substituting Y_t, we get the general form:

$$\phi(B)(1 - B)^d X_t = \theta(B)W_t,$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ and $\{W_t\}$ is a white noise process.

Notice that if $d = 0$, this model represents a stationary ARMA(p, q) process.

To detect if a process will be better fitted by an ARIMA model than an ARMA model, we can look at its plot for trends or parts with stationary behaviour but with different means. We can also look at its correlogram, since the ACF of integrated processes are characterized by a slow decay towards zero instead of a geometrical decay of ARMA processes. Let's generate an ARIMA$(1, 1, 0)$ and an ARIMA$(0, 1, 1)$ and look at their plots, differentiated plots and correlograms.

Example 3.1. *ARIMA* $(1, 1, 0)$

On this example, we generate an ARIMA$(1, 1, 0)$ process with $\phi = 0.8$ of 100 observations. All the code used to generate the process and the figures on this example can be found on Annex I, page 2. We use the seed (2) instead of the seed (1) used on previous examples since the process generated is more interesting.

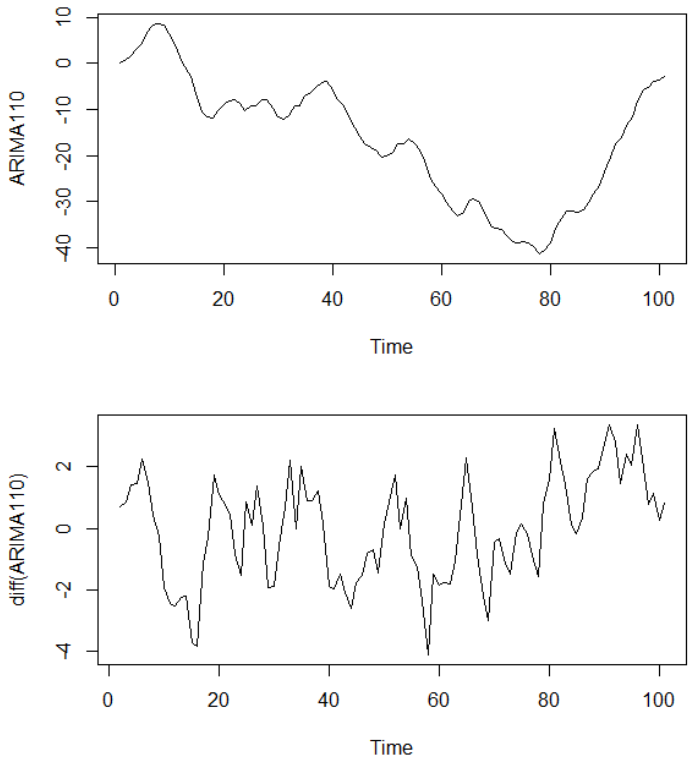

Figure 3.1: Plot of an ARIMA$(1, 1, 0)$ process with $\phi = 0.8$ of 100 observations (above) and plot of the differentiated process (below).

In Figure 3.1 we can see that the plot of the process has a clear decreasing trend followed by an increasing trend from observation 78. This is a non stationary behavior and leads us to consider differentiating the process. The plot of the differentiated process in Figure 3.1 has a stationary behavior, so from the observation of the plots we can conclude that the process is integrated of order 1.

Now, looking at the correlograms of the differentiated series in Figure 3.2 we can identify that it follows an AR(1) model, as described in Section 2.2.

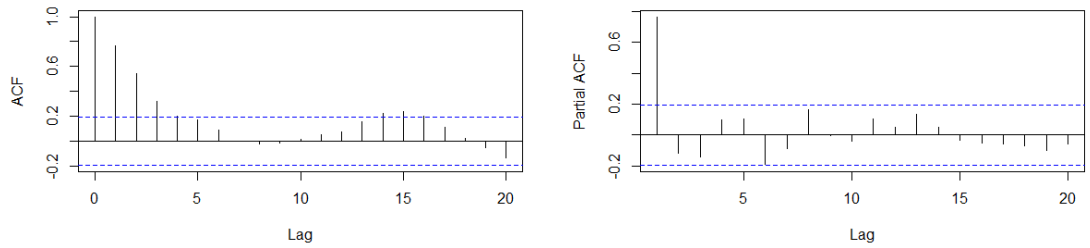

Figure 3.2: Correlogram (left) and partial correlogram (right) of a differentiated ARIMA$(1, 1, 0)$ process with $\phi = 0.8$ of 100 observations.

If we started looking at the correlogram instead of the plot of the series or if we wrongly assumed that the process was stationary and proceeded to analyze the correlograms, we would have seen that the correlogram of the series (Figure 3.3) slowly decays to 0, thing that suggests that the series should be differentiated.

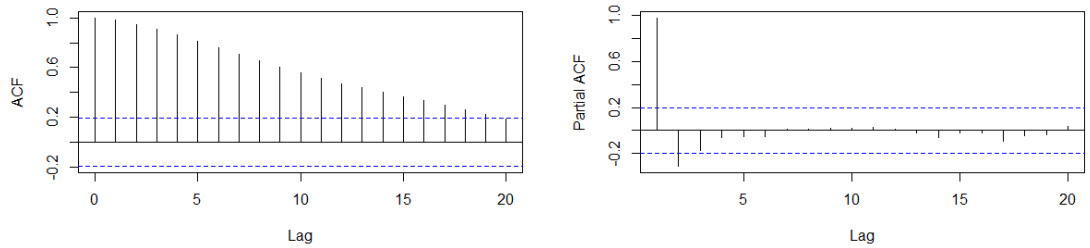

Figure 3.3: Correlogram (left) and partial correlogram (right) of an ARIMA$(1, 1, 0)$ process with $\phi = 0.8$ of 100 observations.

Looking at both correlograms, we might also think that an AR(2) model could fit the process, as the correlogram decreases and the partial correlogram only has significant values at lags 1 and 2. On Section 4.3 we define the *Akaike information criterion* (or AIC), that helps us choose which model fits better the underlying process.

Now, let's look at an ARIMA$(0, 1, 1)$ and check that its correlograms resemble the ones of the ARIMA$(1, 1, 0)$, with a slowly decreasing ACF.

Example 3.2. *ARIMA*$(0, 1, 1)$

The code used to simulate the ARIMA$(0, 1, 1)$ process with $\theta = 0.8$ and its graphs can be found on Annex I, page 3.

Looking first at the correlograms of the series this time (Figure 3.4), we can see that they have a pattern more resembling the correlograms of the ARIMA$(1, 1, 0)$ than the ones of the stationary processes studied on Chapter 2, with a slow leaning towards 0 of

the ACF. This is a sign that the series is not stationary and that we should differentiate it.

The correlograms of the differentiated series on Figure 3.5 have a close behavior to the correlograms of the MA(1) process defined on Section 2.3.

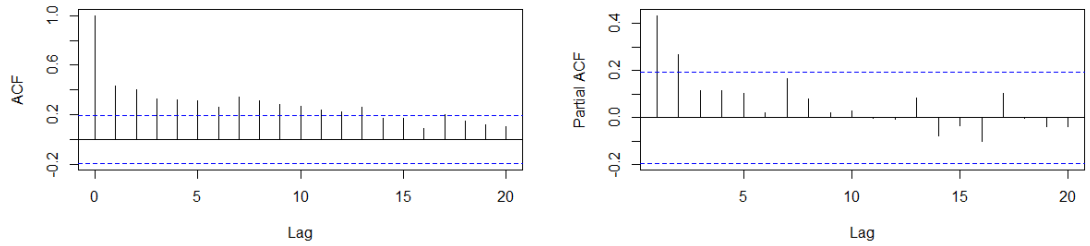

Figure 3.4: Correlogram (left) and partial correlogram (right) of an ARIMA$(0, 1, 1)$ process with $\theta = 0.8$ of 100 observations.

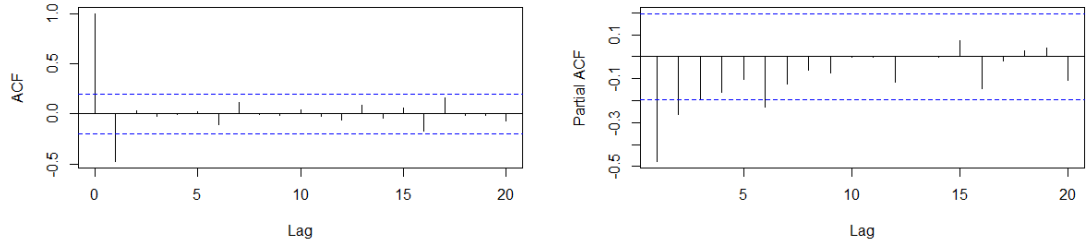

Figure 3.5: Correlogram (left) and partial correlogram (right) of a differentiated ARIMA$(0, 1, 1)$ process with $\theta = 0.8$ of 100 observations.

If we look at the plot of the process (Figure 3.6) we can see that it has an increasing trend, hence it is non-stationary and we should differentiate it in order to fit an ARMA model. In Figure 3.7 we can see that the differentiated series is stationary.

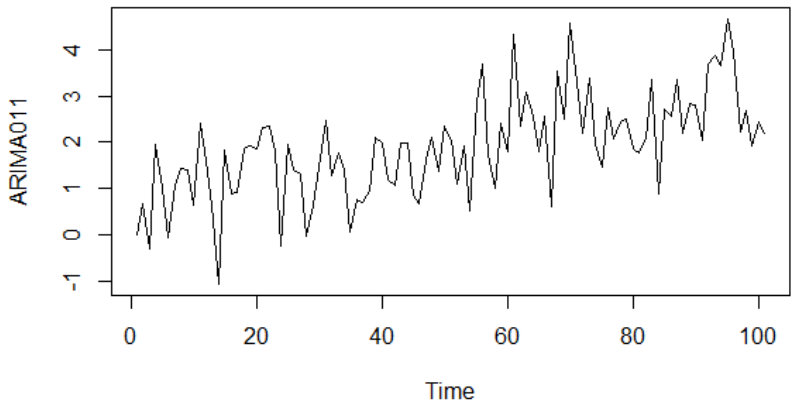

Figure 3.6: Plot of an ARIMA$(0, 1, 1)$ process with $\theta = 0.8$ of 100 observations.

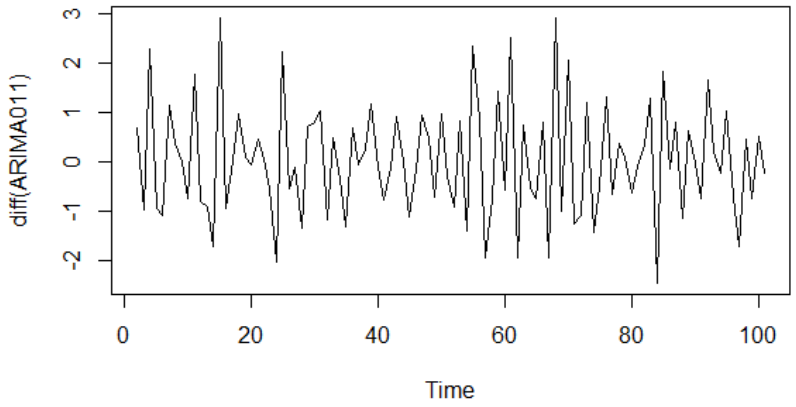

Figure 3.7: Plot of a differentiated ARIMA$(0, 1, 1)$ process with $\theta = 0.8$ of 100 observations.

To conclude this section, we want to remark that we have to be careful to not overdifferentiate the series once stationarity has been achieved. It introduces extra serial correlation and increases the complexity of the model. As an example, consider again a random walk process $\{X_t\}$. We have seen at the beginning of the section that differentiating the process transforms it into a white noise and thus a stationary process. If we differentiate it again, we get

$$\nabla^2 X_t = \nabla W_t = (1 - B)W_t = W_t - W_{t-1}.$$

This means that the model for $\{X_t\}$ is an ARIMA$(0, 2, 1)$ with $\theta = 1$ instead of an ARIMA$(0, 1, 0)$. Apart from being a more complicated process, the value of $\theta = 1$ means that it is non invertible and would cause problems when estimating the parameters.

In practice, most processes can be well fitted with $d \leq 2$.

3.2 SARIMA processes

Seasonal time series are defined by a strong serial correlation at the seasonal lag and (possibly) at their multiples. They can be *pure seasonal models* if there only exist dependencies among variables from one season to the next one or *multiplicative seasonal models* if there are dependencies between values from one season to the next as well as between the near observations of the series. The *seasonal ARIMA* (or SARIMA) models allow to study both kinds of processes. These models are an extension of the ARIMA models described previously including including autoregressive and moving average terms at lag s.

Definition 3.2. Let $\{X_t\}$ be a time series, and d, D non negative integers. $\{X_t\}$ is an *seasonal autoregressive integrated moving average process of order* $(p, d, q) \times (P, D, Q)_s$ *with period* s (or SARIMA$(p, d, q) \times (P, D, Q)_s$) if the process $Y_t = (1 - B)^d(1 - B^2)^D X_t$ is a causal ARMA process defined by

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)W_t,$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, $\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$, $\Theta(B^s) = 1 - \Theta_1 B - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$ and $\{W_t\}$ is a white noise process.

Substituting Y_t, we get the general form:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B)^D X_t = \theta(B)\Theta(B^s)W_t$$

Notice that if p, d and q are zero, the formula describes a pure seasonal process. Otherwise, if P, D or Q is non zero the formula describes a multiplicative seasonal process.

The conditions of stationary, causality and invertibility for the seasonal processes are the same as the ones of the non seasonal processes but considering Φ_i instead of ϕ_i and Θ_i instead of θ_i and their correlograms behave the same but s-lagged. In Figure 3.8 we can see the correlograms of a simulated pure seasonal autoregressive process of order 1 with period 12 of 500 observations. Notice that the integer lags of the correlogram tend to 0 geometrically and there is only one significant value in the partial autocorrelogram at lag 1. Notice that on R, the lags are counted based on the seasonality and therefore, what we would count as lag 12 appears as lag 1 on the graph.

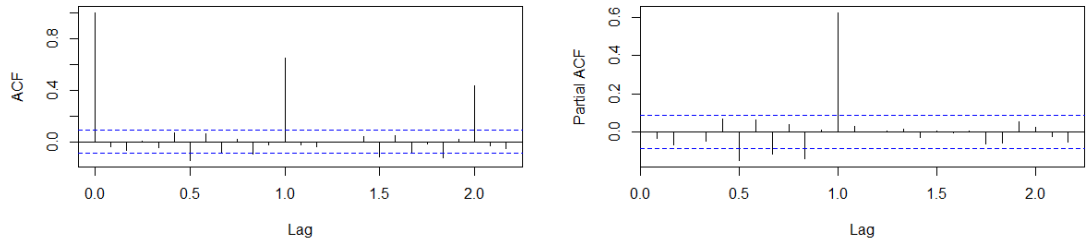

Figure 3.8: Correlogram (left) and partial correlogram (right) of a SARIMA$(0,0,0) \times (1,0,0)_{12}$ process with $\Phi = 0.7$ of 500 observations.

For the multiplicative seasonal processes the correlograms show both the regular component as described in Section 3.1 and the seasonal component as described previously in this section. In Figure 3.9 we can see the correlograms on a SARIMA$(1,0,0) \times (1,0,0)$. The code use to simulate the processes and generate the graphs on Figures 3.8 and 3.9 can be found on Appendix A.3, Listing A.9.

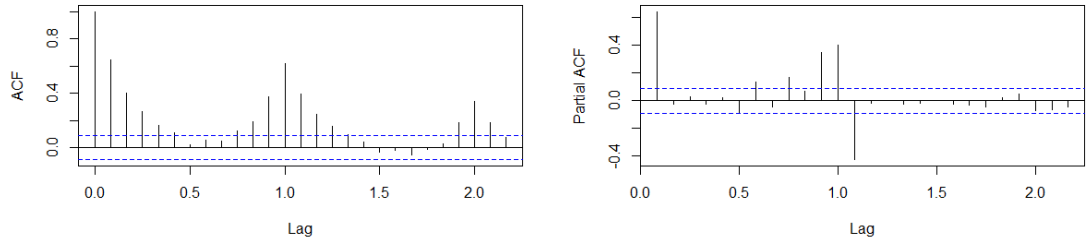

Figure 3.9: Correlogram (left) and partial correlogram (right) of a SARIMA$(1,0,0) \times (1,0,0)_{12}$ process with $\phi = 0.7$ and $\Phi = 0.7$ of 500 observations.

Chapter 4

Model identification and forecasting

The main reason for applying the studied models and methods of time series is for forecasting purposes. Therefore, it is necessary to know which model is worth using to get accurate future predictions.

First, we identify which model could explain the behavior of our data and estimate the parameters of the model. Then we apply different methods to check if there is any lack of fit on the model that we have selected and diagnose the cause of it. This will be done by using the Akaike information criterion, studying the significance of the parameters and analyzing the residuals of the fitted model. The *Akaike information criterion* (or AIC) was first introduced by H. Akaike at a 1971 symposium, published in 1973 [1] and formally defined in 1974 [2]. It is a measure of the relative quality of statistical models for a given set of data.

If we identify any inadequacy, we start again the process identifying another tentative model that could explain our data behavior using the information from the diagnosis. This process is repeated until a suitable model is found.

Finally, when the model that approximates better our data has been identified, we use it to predict future values.

We explain this process following the classical example of the monthly totals of international airline passengers from January 1949 to December 1960. This data was first used as an example in 1976 by Box and Jenkins [6] and has been used by many authors since then (see for example Brockwell and Davis, 2006 [9], 2016 [10] and Cowpertwait and Metcalfe, 2009 [11]). The data set is build-in on R's default database and can be accessed using the call `AirPassengers`.

4.1 Model identification

The goal of this stage is to identify models that could potentially fit our data or, in other words, of which model could our analyzed data be a particular realization. There can be many potential models identified, since the exact behavior of the data depends on the behavior of the physical world and it cannot be described by purely mathematical arguments.

The first thing that we have to do for trying to fit an ARMA model is to make sure that the time series that we are analysing is stationary. We have seen in Chapter 3 how to identify and transform non stationary series in mean terms into stationary series by differentiating. To identify non stationarity in variance terms we can look at its plot for an increase of variability over time, this is that the values gradually tend to be further from the mean. Non stationarity in variance can be addressed applying the *power transformation* defined by Box and Cox in 1964 [3].

Definition 4.1. Let $\{X_t\}$ be a time series. We define the *Box-Cox transformation* f_λ as

$$f_\lambda(X_t) = \begin{cases} \dfrac{X_t^\lambda - 1}{\lambda}, & \lambda > 0, \\ \\ \ln X_t, & \lambda = 0, \end{cases}$$

where λ is a real parameter.

In practice, if the transformation is needed, $\lambda = 0$ is often an adequate parameter. The transformation has to be applied before the differentiation of the series both in the regular part or the seasonal part. Notice that it can only be applied on positive processes.

Example 4.1. *Previous transformations and stationarity of the international airline passengers series.*

Let's look at the plot of the airline passengers data to see if this transformation is necessary.

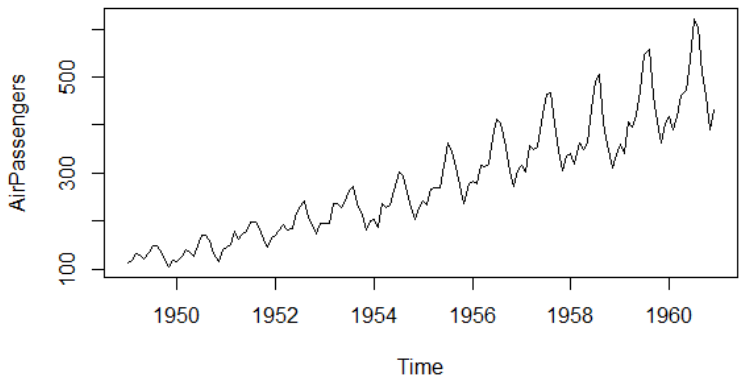

Figure 4.1: Plots of the monthly total international airline passengers from January 1949 to December 1960. Source: R "datasets" package.

In Figure 4.1 we can see that the airline passengers has an increasing trend, a strong seasonal effect of cycle 12 and an increase on variance over time. Therefore, we need to apply a log-transformation and consider differentiating the series on its regular part, on its seasonal part or on both.

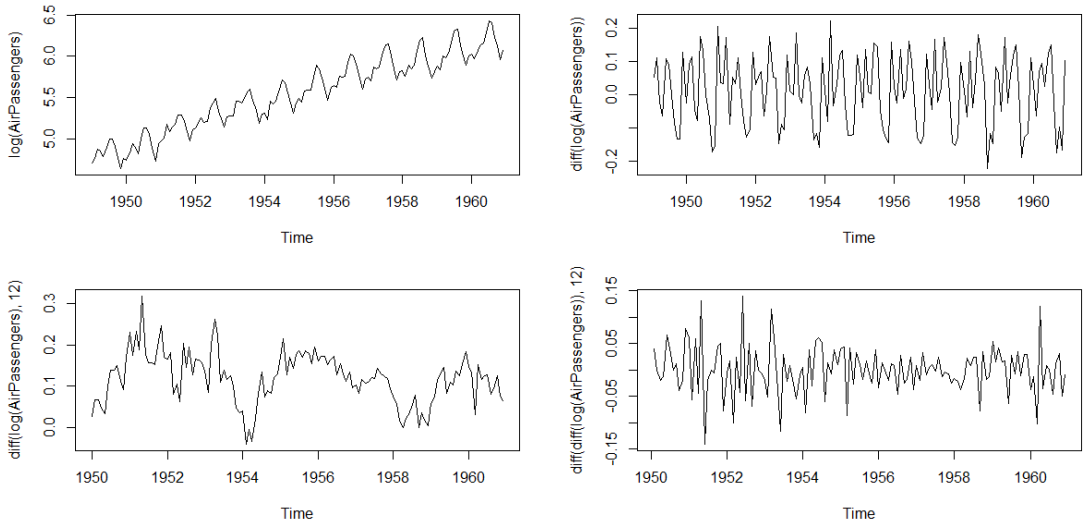

Figure 4.2: Plots of the log-transformed airline passengers series: undifferentiated series (top left), regular differentiated series (top right), seasonal differentiated series (bottom left) and regular and seasonal differentiated series (bottom right).

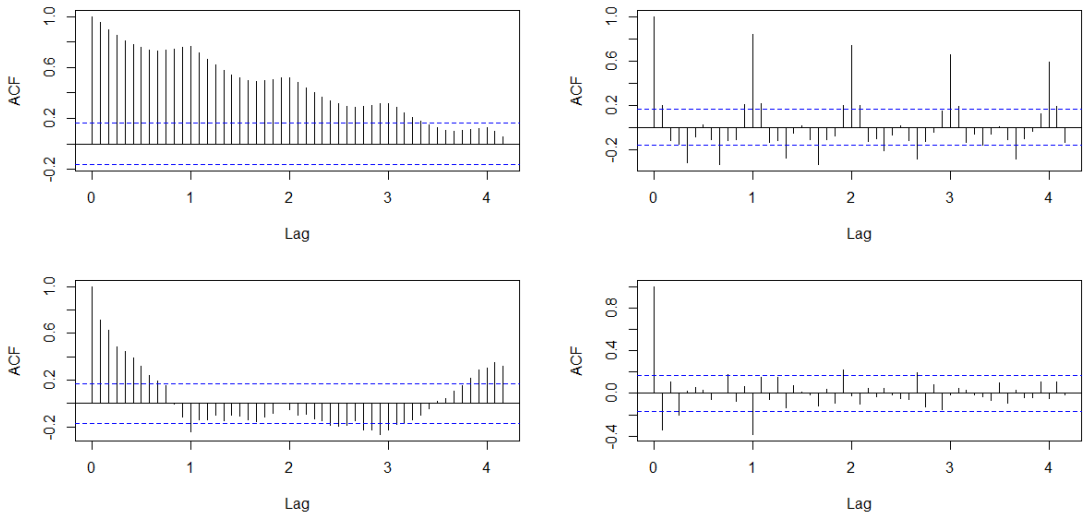

Figure 4.3: Sample ACF of the log-transformed airline passengers series: undifferentiated series (top left), regular differentiated series (top right), seasonal differentiated series (bottom left) and regular and seasonal differentiated series (bottom right).

Looking at the plots of the differentiated series in Figure 4.2 and its sample auto-covariarnce functions in Figure 4.3, it is clear that the series needs to be differentiated once on regular terms in order to be stationary as the difference removes the trend and the sample ACF tends to 0 slowly without the difference. Regards the need of seasonal differentiating, notice that the sample ACF of the regular differentiated series tends to 0

slowly and considering a first order difference both on regular terms and seasonal terms the sample ACF follows a stationary pattern. Finally, we can also see that the plot of the regular and seasonal differentiated series has a stationary behavior. Hence we should consider a seasonal multiplicative model.

Once we have transformed (or not) the series into a stationary series we can begin to speculate on which model could fit better our data. In Chapter 2 and Chapter 3 we have seen that each process following one of the models studied has a characteristic behavior of its plot or correlograms. We consider the different tentative models based on the particularities that we identify looking at their sample correlograms. A summary of these patterns for the regular part can be found in Table 4.1. The patterns for the seasonal part with frequency s are the same as the ones for the regular part but spaced s lags.

Table 4.1: Summary of the behavior of the ACF and the PACF for AR(p), MA(q) and ARMA(p, q) processes.

	AR(p)	MA(q)	ARMA(p, q)
AFC	Decays geometrically to 0.	Significant values for the first q lags. Non significant values afterwards.	Decays geometrically to 0 from lag q.
PACF	Significant values for the first p lags. Non significant values afterwards.	Decays geometrically to 0.	Decays geometrically to 0 from lag p.

Example 4.2. *Model identification for the log-transformed airline passengers series.*

The sample correlograms of the series in Figure 4.4 both have significant spikes at lags 1 and 12 (remember that R shows lag 12 as lag 1 as $s = 12$) and the rest are non significant. Since there is no geometrical trend towards 0 on any correlogram, we consider 5 potential models that could fit the series: SARIMA$(0,1,1) \times (0,1,1)_{12}$, SARIMA$(0,1,1) \times (1,1,0)_{12}$, SARIMA$(1,1,0) \times (0,1,1)_{12}$, SARIMA$(1,1,0) \times (1,1,0)_{12}$ and SARIMA$(1,1,1) \times (1,1,1)_{12}$. We will see which model fits better the series when diagnosing the series at Section 4.3.

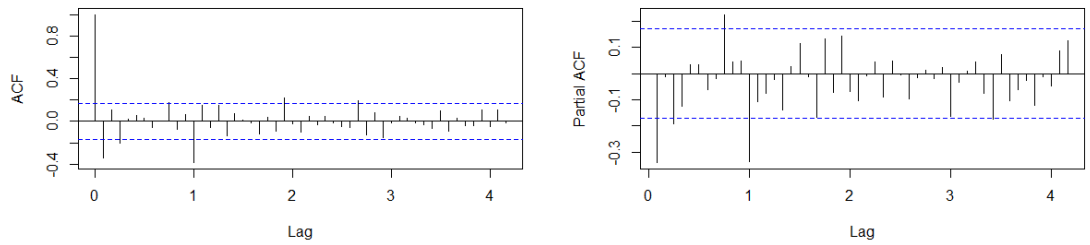

Figure 4.4: Correlogram (left) and partial correlogram (right) of the regular and seasonal differentiated log-transformed airline passengers series.

4.2 Parameter estimation

Once some feasible models have been identified we have to estimate the parameters $\phi_1, \ldots, \phi_p, \Phi_1, \ldots, \Phi_P, \theta_1, \ldots, \theta_q, \Theta_1, \ldots, \Theta_Q$ of that model. For pure autoregressive processes the ϕ_i parameters can be estimated using the Yule-Walker equations (2.1) replacing the autocorrelations $\rho(h)$ by the sample autocorrelations, but this method is not used for moving average processes since it would imply solving a complicated nonlinear equations system.

One of the most used methods for parameter estimation and the one used by default by the `arima` function in R is the *maximum likelihood estimation*. This method estimates the parameters that maximize the probability of the observed series to be a particular realization of the estimated model. In this way the estimated parameters of the model using this method are the ones with the highest probability of obtaining the observed series.

Definition 4.2. Let $\{X_t\}$ be a time series, $\mathbf{X}_n = (X_1, \ldots, X_n)$ and Γ_n the covariance matrix $\Gamma_n = E(\mathbf{X}_n' \mathbf{X}_n)$. Assuming that Γ_n is nonsingular, the *function of likelihood of* $\mathbf{X_n}$ is

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{X}_n \Gamma_n^{-1} \mathbf{X}_n'\right).$$

This definition can be found in Brockwell and Davis, 2016 [10].

Notice from the autocovariance formulas calculated in Chapter 2 that the covariance matrix Γ_n depends on the parameters $\phi_1, \ldots \phi_p$, $\theta_1, \ldots, \theta_q$ and on σ^2. Therefore, L depends on the chosen model. In Brockwell and Davis, 2016 [10] it is also shown how to obtain an expression of the likelihood function for an ARMA process using the innovators algorithm and the conditions for the parameter estimators to be the maximum likelihood ones.

The estimators that maximize the likelihood function are found differentiating $\ln L(\Gamma_n)$ partially with respect to σ^2 and finding the parameters that make $\frac{\partial}{\partial \sigma^2} \ln L(\Gamma_n) = 0$.

Example 4.3. *Parameter estimation of the potential models for the airline passengers series.*

Using the code on Annex I, page 3, we fit each potential model considered in Example 4.2 for the monthly airline passengers series and get the estimations for the parameters of Table 4.2.

Table 4.2: Estimated parameters for the potential models to fit the log-transformed airline passengers series

Model	ϕ	θ	Φ	Θ
SARIMA$(0,1,1) \times (0,1,1)_{12}$		0.4018280		0.5569448
SARIMA$(0,1,1) \times (1,1,0)_{12}$		0.4423334	-0.4742972	
SARIMA$(1,1,0) \times (0,1,1)_{12}$	-0.3395210			0.5618858
SARIMA$(1,1,0) \times (1,1,0)_{12}$	-0.3744776		-0.4637481	
SARIMA$(1,1,1) \times (1,1,1)_{12}$	0.1666474	0.5614956	-0.0990091	0.4973187

4.3 Model Diagnostic Checking

Once we have a set of feasible models and their estimated parameters, we have to choose which one could fit better our analyzed series and check if the model is adequate. If we find evidences that the model could be inadequate, we will need to know the reasons why the model is inadequate to modify it and find an appropriate model.

To choose the model with the highest potential of "good fitting" we use the *Akaike information criterion*. This criterion estimates the quality of a model relative to other models considered based on the information that is lost by using the model instead of the others. The AIC also takes into account the simplicity of the model, penalizing models with more parameters if they do not improve substantially the loss of information.

Definition 4.3. Let $\{X_t\}$ be a time series and L be the likelihood function of the model given on Definition 4.2. The *Akaike information criterion* for a SARIMA$(p,d,q)\times(P,D,Q)_s$ is

$$AIC = 2(p+q+P+Q+1) - 2\ln L.$$

The model that minimizes the loss of information is the one with the minimum AIC value.

Example 4.4. AIC of the potential models to fit the airline passengers series.

In Table 4.3 we can see that the considered model with smaller AIC from the ones that we considered on Example 4.2 is the SARIMA$(0,1,1)\times(0,1,1)_{12}$. Hence it is the model that minimizes the loss of information among them and is the one that we will continue diagnosing on further examples.

Table 4.3: AIC of the potential models fitted to the log-transformed airline passengers series

Model	AIC
SARIMA$(0,1,1)\times(0,1,1)_{12}$	-483.3991
SARIMA$(0,1,1)\times(1,1,0)_{12}$	-477.4053
SARIMA$(1,1,0)\times(0,1,1)_{12}$	-481.4896
SARIMA$(1,1,0)\times(1,1,0)_{12}$	-474.8188
SARIMA$(1,1,1)\times(1,1,1)_{12}$	-480.3109

The code used to calculate these values can be found on Annex 1, page 3.

Now that we have found the model that minimizes the loss of information, let's check if it is an adequate model. The first validation that we will do is on the significance of the estimated parameters. If any of the parameters is not significantly different from zero we should consider a simpler model. The significance of each parameter is tested by analyzing the ratio between the parameter estimation $\hat{\beta}_i$ and its standard error

$$\sigma_i = \sqrt{\frac{\sum_{j=0}^{i-1}\hat{\beta}_j^2}{n}}\ .$$

If the ratio $|\hat{\beta}_i/\sigma| < 1.96$ we can conclude that the parameter is not significant. Notice that the condition on the ratio is equivalent to the condition that 0 is not in the 95% confidence interval of the parameters $\hat{\beta}_i = [\hat{\beta}_i - 1.96\sigma, \hat{\beta}_i + 1.96\sigma]$.

Example 4.5. *Significance of the estimated parameters of the SARIMA*$(1,1,1)\times(1,1,1)_{12}$ *and SARIMA*$(0,1,1)\times(0,1,1)_{12}$ *models fitted to the log-transformed airline passengers series.*

On R, the `arima` function returns a list of information including the estimated parameters and its standard error. In Listing 4.1 we can see the output of fitting a SARIMA$(0,1,1)\times(0,1,1)_{12}$ model to the log-transformed airline passengers series.

Listing 4.1: R output for ap011011

```
Call:
arima(x = log(AirPassengers), order = c(0, 1, 1),
    seasonal = list(order = c(0, 1, 1), 12))

Coefficients:
          ma1      sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,
aic = -483.4
```

We can see that both θ and Θ are significant since their ratios $0.4018/0.0896 = 4.48$ and $0.5569/0.0731 = 7.62$ respectively are greater than 1.96.

If we firstly assumed that the SARIMA$(1,1,1)\times(1,1,1)_{12}$ model fits better the series, checking the significance of its parameters (Listing 4.2) we would see that ϕ and Φ are not significant, since their ratios are $0.1666/0.2459 = 0.68$ and $0.099/0.154 = 0.64$ respectively and we should consider a model without those parameters. Notice that θ and Θ are still significant since their ratios are $0.5615/0.2116 = 2.65$ and $0.4973/0.1360 = 3.66$ respectively, suggesting to reduce the model to a SARIMA$(0,1,1)\times(0,1,1)_{12}$ model.

Listing 4.2: R output for ap111111

```
Call:
arima(x = log(AirPassengers), order = c(1, 1, 1),
    seasonal = list(order = c(1, 1, 1), 12))

Coefficients:
         ar1      ma1     sar1     sma1
      0.1666  -0.5615  -0.099  -0.4973
s.e.  0.2459   0.2116   0.154   0.1360

sigma^2 estimated as 0.001336:  log likelihood = 245.16,
aic = -480.31
```

Finally, we compare the predicted values of the fitted model to our actual observations and check if, in fact, it provides a good fit or not. The *residuals* of the model are the difference between the predicted values and the actual observations. The predicted values

of the series can be computed substituting the real values on the formula of the model. If the considered model was the process that generated our observations, the residuals would be uncorrelated or, in other words, the series of the residuals would behave as a white noise process (recall Example 1.1). If the residuals show any kind of correlation we might have missed something on the identification process and a better fitting model should be considered.

One way to check if the residuals are uncorrelated is looking at their correlograms. The residuals behave as a white noise if their correlograms have no significant spikes. If the residuals' correlograms have any significant spike, we might have forgotten to include an ARMA process to the model. The forgotten process can be identified analyzing the behavior of the residuals' correlograms as in Section 4.1.

Example 4.6. *Analysis of the residuals' correlograms of the* $SARIMA(0,1,1)\times(0,1,1)_{12}$ *and* $SARIMA(0,1,0)\times(0,1,1)_{12}$ *models fitted to the log-transformed airline passengers series.*

In Figure 4.5 we see that only the value of the ACF at lag 23 lies outside of the significance bounds. Since less that 5% of the values of the AFC and PACF of the residuals of fitting a $\text{SARIMA}(0,1,1)\times(0,1,1)_{12}$ model on the data lie outside of the significant bounds, we do not reject the hypothesis that the residuals behave as a white noise and the model is well fitted.

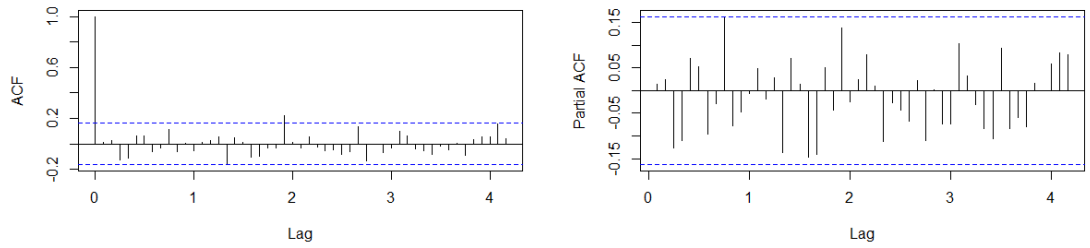

Figure 4.5: Correlogram (left) and partial correlogram (right) of the residuals of the $\text{SARIMA}(0,1,1)\times(0,1,1)_{12}$ model fitted to the log-transformed airline passengers series.

From the residuals' correlograms of the $\text{SARIMA}(0,1,0)\times(0,1,1)_{12}$ model fitted to the log-transformed airline passengers series in Figure 4.6 we can see that the residuals do not behave as a white noise and that an AR(1), MA(1) or ARMA(1, 1) could be added to the model, since both ACF and PACF have a significant spike at lag 1.

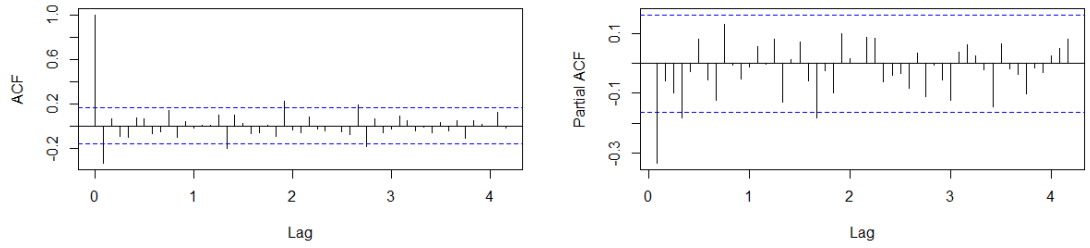

Figure 4.6: Correlogram (left) and partial correlogram (right) of the residuals of the $\text{SARIMA}(0,1,1)\times(0,1,1)_{12}$ model fitted to the log-transformed airline passengers series.

Another way to check that there is no correlation between the residuals of the fitted model is to perform portmanteau tests in the residuals' series with the null hypothesis

that the residuals are not correlated. These tests are a great addition to the analysis of the correlograms on the diagnosis of the fitted models, but should not substitute them. *Box-Pierce test* and *Ljung-Box test* are two portmanteau tests with the null hypothesis that the first H residuals are not correlated.

The first test was developed by George E. P. Box and David A. Pierce (1970) [8] and states that if the residuals are not correlated, then

$$Q = n \sum_{h=1}^{H} \rho_r^2(h) \sim \chi^2(H - p - q),$$

where n is the sample size minus d and ρ_r are the autocorrelations of the residuals. A large value of Q means that the autocorrelation of the residuals is to high to be considered white noise. We reject the null hypothesis at significance level $\alpha = 5\%$ if $Q > \chi^2_{1-\alpha}(H - p - q)$. On most statistical programs, the test returns the p-value of $Q < \chi^2_{1-\alpha}(H - p - q)$.

In 1978 Greta M. Ljung and George E. P. Box [12] showed that the χ^2 distribution did not provide an adequate approximation to the distribution of the Q statistic and proposed a modification of the statistic

$$\tilde{Q} = n(n+2) \sum_{h=1}^{H} \rho_r^2(h)/(n-h).$$

Example 4.7. *Portmanteau tests on the residuals of the SARIMA*$(0,1,1) \times (0,1,1)_{12}$ *and SARIMA*$(0,1,0) \times (0,1,1)_{12}$ *models fitted to the log-transformed airline passengers series.*

The code used to perform the tests can be found in Annex I, page 3.

Performing both tests at lag 24 on the residuals of the SARIMA$(0,1,1) \times (0,1,1)_{12}$ model we get that the p-value for the Box-Pierce test is 0.5008 and 0.3309 for the Ljung-Box test. Since they are greater than 0.05 we do not reject the null hypothesis of the independence of the residuals.

On the other hand, performing the tests on the residuals of the SARIMA$(0,1,0) \times (0,1,1)_{12}$ model, that we already know does not fit well the data, we get that the p-values for the Box-Pierce and Ljung-Box tests are 0.006347 and 0.001985. Hence we reject the null hypothesis and conclude that the residuals are correlated. A different SARIMA model should be considered.

After all the checking, we can conclude that the best fitting SARIMA model for the log-transformed airline passengers series is the SARIMA$(0,1,1) \times (0,1,1)_{12}$ and therefore, is the model that we will use to predict future values of the series.

4.4 Forecasting

Now that we have identified and checked the best model to explain the underlying process of our observed data, we use it to forecast the future values of the observed time series. At time t we have the observations $\{x_1, x_2, \ldots, x_t\}$ of the time series $\{X_t\}$ and we want to forecast the value of the observation x_{t+i}. For the underlying SARIMA process, this observation can be directly computed from the equation of the model

$$\phi(B)\Phi(B^s)(1-B)^d(1-B)^D x_{t+i} = \theta(B)\Theta(B^s)w_{t+i},$$

recursively substituting the observed values on the expanded form of the equation

$$x_{t+i} = \sum_{j=1}^{p+P+d+D} \pi_j x_{t+i-j} - \sum_{j=1}^{q+Q} \psi_j w_{t+i-j},$$

where π_j are obtained expanding $\phi(B)\Phi(B^s)(1-B)^d(1-B)^D$ and ψ_j are obtained expanding $\theta(B)\Theta(B^s)$.

It can be proved (G. E. P. Box and G. M. Jenkins [7]) that for the time series, the estimator $\hat{x}_t(i)$ of the observation x_{t+i} is the one that minimizes the square error of the forecast, where is the $\hat{x}_t(i)$ is the expectation of x_{t+i} conditional on the past observed values and the parameters

$$\hat{x}_t(i) = E[x_{t+i}|x_1, \ldots, x_t; \phi, \Phi, \theta, \Theta].$$

To simplify the notation, we will use the square brackets to refer to the conditional expectation at time t on the observed values and the parameters of the model.

$$[x_{t+i}] = E[x_{t+i}|x_1, \ldots, x_t; \phi, \Phi, \theta, \Theta].$$

Taking conditional expectations at time t in the expanded equation we get

$$\hat{x}_t(i) = [x_{t+i}] = \sum_{j=1}^{p+P+d+D} \pi_j [x_{t+i-j}] - \sum_{j=1}^{q+Q} \psi_j [w_{t+i-j}].$$

Notice that

$$\begin{array}{ll}
[x_{t+i}] = x_{t+i}, & \text{if } i = 0, -1, \ldots, -t, \\
[x_{t+i}] = \hat{x}_t(i), & \text{if } i \geq 1, \\
[w_{t+i}] = w_{t+i} = x_{t+i} - \hat{x}_{t+i-1}(1), & \text{if } i = 0, -1, \ldots, -t, \\
[w_{t+i}] = 0, & \text{if } i \geq 1.
\end{array}$$

Therefore, the forecasts $\hat{x}_t(i)$ $(j \geq 1)$ can be calculated recursively substituting the observed values, the forecast for the previous $i-1$ values and the errors of the one-step-ahead forecast of the observed values (notice that the expected errors for the future observations is 0).

Writing the equation of x_{t+i} as a function of $\{w_t\}$

$$x_{t+i} = \sum_{j=0}^{\infty} \psi_j w_{t+i-j},$$

where ψ_j are obtained expanding $\phi(B)^{-1}\Phi(B^s)^{-1}(1-B)^{-d}(1-B)^{-D}\theta(B)\Theta(B^s)$, we can calculate the forecast errors $\varepsilon_t(i)$

$$\begin{aligned}
x_{t+i} =& (w_{t+i} + \psi_1 w_{t+i-1} + \cdots + \psi_{i-1} w_{t+1}) \\
&+ (\psi_i w_t + \psi_{i+1} w_{t-1} + \ldots) \\
=& \varepsilon_t(i) + \hat{x}_t(i),
\end{aligned}$$

From the standard deviation of the forecast errors

$$\sigma^2(i) = \sqrt{\mathrm{Var}[\varepsilon_t(i)]} = (1 + \psi_1^2 + \cdots + \psi_{i-1}^2)^{1/2} \sigma_w$$

and assuming that $\{w_t\}$ follow a normal distribution we can obtain the bounds of the confidence intervals for the forecasts $\hat{x}_t(i)$. The 95% confidence interval of the forecasts $\hat{x}_t(i)$ are then $[\hat{x}_t(i) - 1.96\sigma(i),\ \hat{x}_t(i) + 1.96\sigma(i)]$.

Example 4.8. *Forecasting future values of the airline passengers series.*

On R, we can predict future values of a fitted SARIMA model using the function `predict`. This function returns the forecasts $\hat{x}_t(i)$ and standard errors of the forecast. We will use it to calculate the 24 next values (two times the season length). The code used to generate Listing 4.3, Figure 4.7 and Figure 4.8 can be found on Appendix A.4, Listing A.10.

Listing 4.3: R output for p.ap

```
$pred
          Jan      Feb      Mar      Apr      May      Jun
1961 6.110186 6.053775 6.171715 6.199300 6.232556 6.368779
1962 6.206435 6.150025 6.267964 6.295550 6.328805 6.465028
          Jul      Aug      Sep      Oct      Nov      Dec
1961 6.507294 6.502906 6.324698 6.209008 6.063487 6.168025
1962 6.603543 6.599156 6.420947 6.305257 6.159737 6.264274

$se
             Jan        Feb        Mar        Apr        May
1961 0.03671562 0.04278291 0.04809072 0.05286830 0.05724856
1962 0.09008475 0.09549708 0.10061869 0.10549195 0.11014981
             Jun        Jul        Aug        Sep        Oct
1961 0.06131670 0.06513124 0.06873441 0.07215787 0.07542612
1962 0.11461854 0.11891946 0.12307018 0.12708540 0.13097758
             Nov        Dec
1961 0.07855851 0.08157070
1962 0.13475740 0.13843405
```

Notice from Listing 4.3 and Figure 4.7 that the standard errors increase over time, decreasing the accuracy of the predicted values.

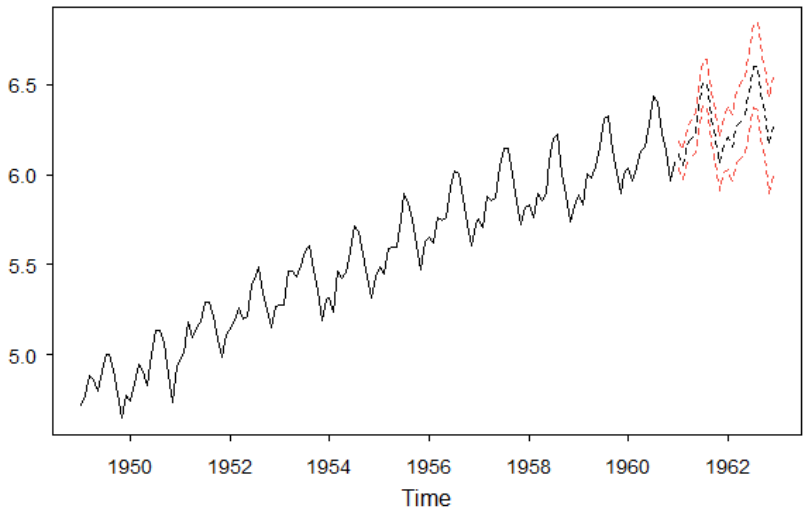

Figure 4.7: Plot of the log-transformed airline passengers series with 24 predicted months (dashed) and their 95% confidence intervals (red dashed).

Recall that we applied a log-transformation on the data. To get the forecasts for the original series we have to undo the transformation by applying the exponential function on each forecast. The plot of the original data and its forecasted values can be found in Figure 4.8.

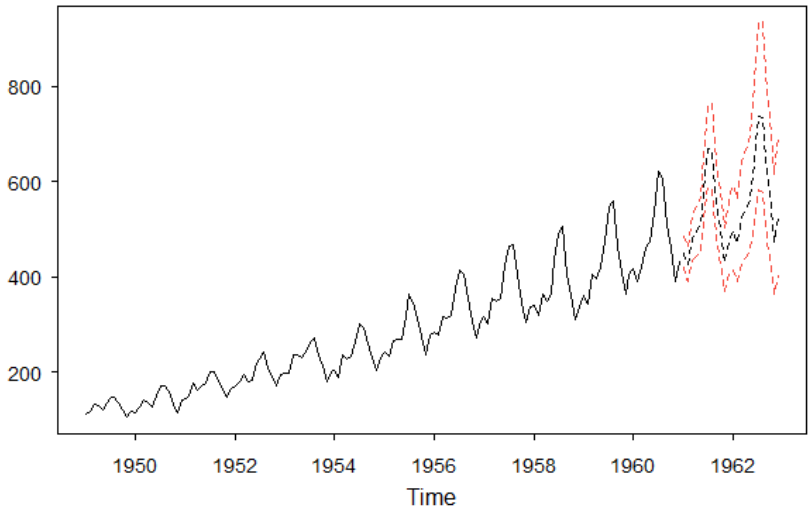

Figure 4.8: Plot of the airline passengers series with 24 predicted months (dashed) and their 95% confidence intervals (red dashed).

Chapter 5

Analysis of the Spanish unemployment rate

The goal of this chapter is to fit a SARIMA model to the Spanish unemployment rate following the process defined in Chapter 4 and forecast some future values for it. All the code used for the graphs and tables of this chapter can be found in Appendix A.5, Listing A.11.

The unemployment rate is one of the main macroeconomic indicators of the situation of the economy of a country and has been a theme of discussion for economists over the years.

According to the OECD (Organisation for Economic Co-operation and Development) "The unemployed are people of working age who are without work, are available for work, and have taken specific steps to find work".[1] The unemployment rate is defined as the relation between the number of unemployed people and the sum of the unemployed people plus those in employment.

A high unemployment rate has negative effects for a country both in the economic and social sense. It means that there is a waste of resources (workforce) and that a significant amount of people is unable to earn money to meet their financial obligations, increasing inequality and driving people to poverty. This can lead into an increase of conflict in the country since the population tends to attribute the state of the economy to the government and the unrest can end up in riots. It also usually involves an increase of crime, since it becomes the only way for people without income for a long period of time to survive.

In Spain, this indicator is estimated through the "Encuesta de población activa", a quarterly survey on households to obtain data on the workforce of the country. The outcome of this survey can be accessed through the INE ("Instituto Nacional de Estadística") webpage.[2] To work with the data on R we use the "API JSON INE", a service that allows to access all the data available through an URL petition in JSON format. To find the URL petition that we need, we can use their URL generator (`https://www.ine.es/dyngs/DataLab/es/manual.html?cid=66`). We explicitly request for the data from the first quarter of 2002 to the first quarter of 2022 in order to maintain the usability of this work over time.

[1] `https://data.oecd.org/unemp/unemployment-rate.htm`
[2] `https://www.ine.es/index.htm`

5.1 Model identification

Before checking for stationary conditions on the series, we have to perform some transformation on the data obtained to format it as a time series object, needed to apply some of the functions that we use. Once formatted, we can see in Figure 5.1 that the series is not stationary both in mean and variance terms. Therefore, we apply a log-transformation to the series and proceed to analyze the differences needed to transform the series into a stationary one.

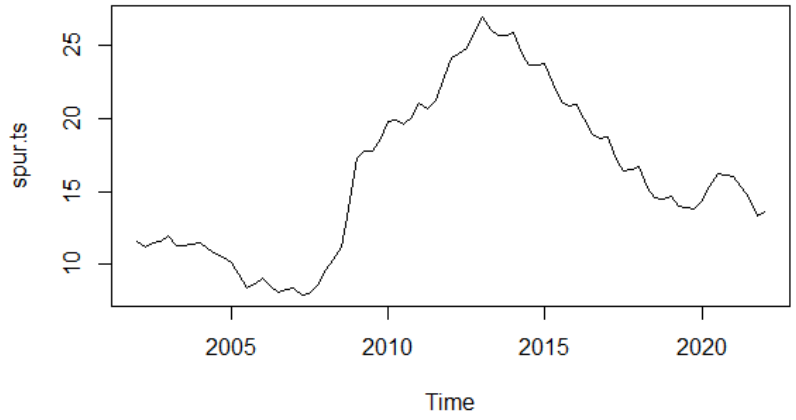

Figure 5.1: Plot of the quarterly unemployment rate of Spain from Q1 2002 to Q1 2022.

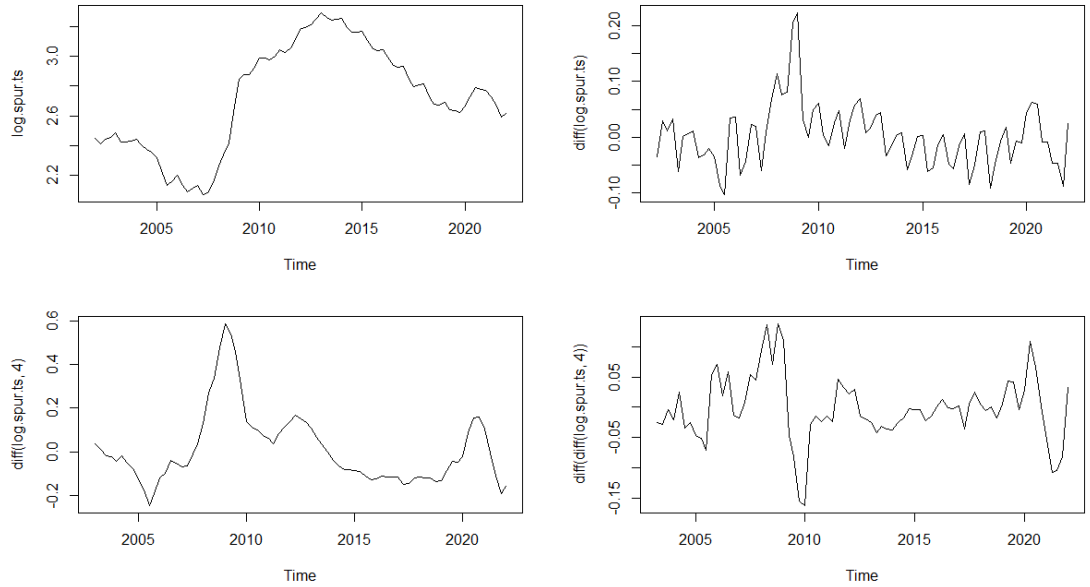

Figure 5.2: Plots of the log-transformed Spanish unemployment rate series: undifferentiated series (top left), regular differentiated series (top right), seasonal differentiated series (bottom left) and regular and seasonal differentiated series (bottom right).

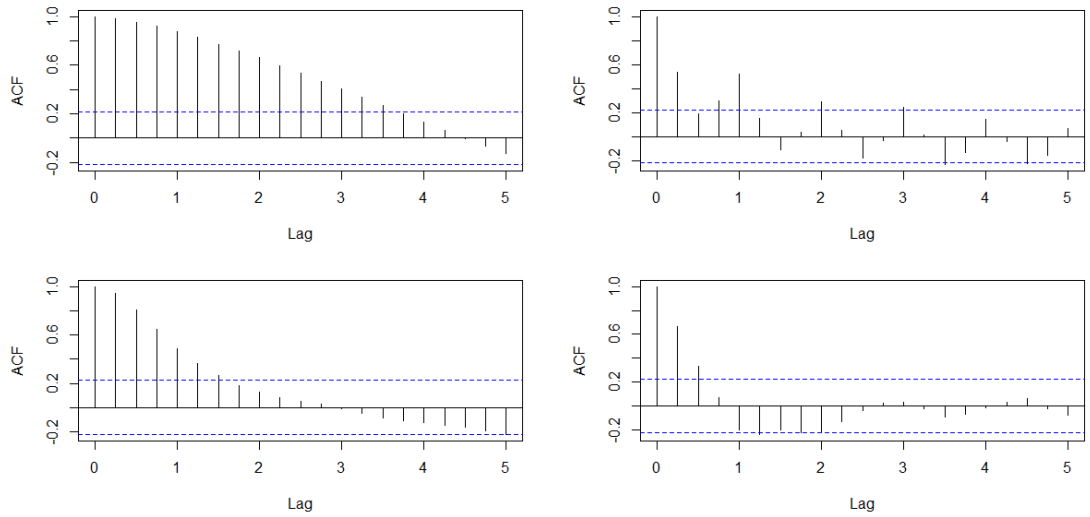

Figure 5.3: Sample ACF of the log-transformed Spanish unemployment rate series: undifferentiated series (top left), regular differentiated series (top right), seasonal differentiated series (bottom left) and regular and seasonal differentiated series (bottom right).

From the plots of the differentiated series in Figure 5.2 and its sample autocovariance functions in Figure 5.3 we can see that the seasonal differentiated series and the regular and seasonal differentiated series show a stationary behavior. Let's check the correlograms of both series and identify feasible models.

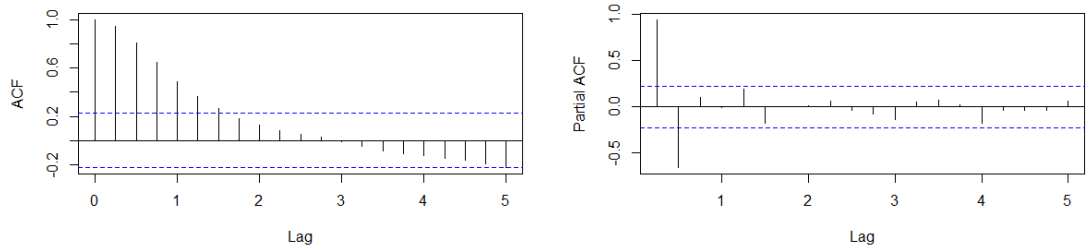

Figure 5.4: Correlogram (left) and partial correlogram (right) of the seasonal differentiated log-transformed Spanish unemployment rate series.

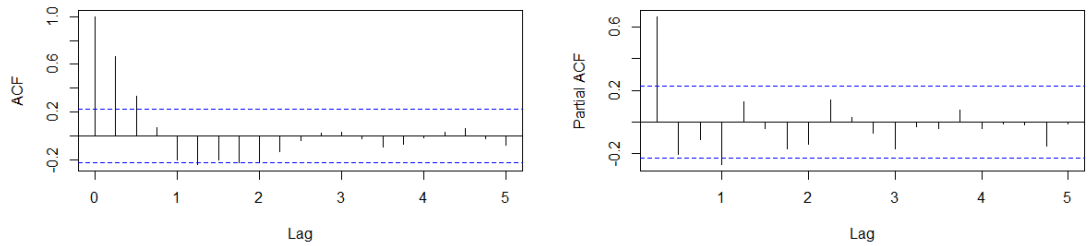

Figure 5.5: Correlogram (left) and partial correlogram (right) of the regular and seasonal differentiated log-transformed Spanish unemployment rate series.

Looking at the sample correlograms of the seasonal differentiated series in Figure 5.4 we can clearly identify the pattern of an AR(2) process, since the ACF decays to 0 and the significant values of the PACF are only the 2 first. This suggests that the SARIMA$(2,0,0) \times (0,1,0)_4$ is one of the models to consider. On the other hand, the sample correlograms of the seasonal differentiated series in Figure 5.5 suggest that the regular part may follow an AR(1) process and the seasonal part may follow an $AR(1)_4$, a $MA(1)_4$ or an ARMA(1,1)$_4$ process. This means that we should also consider the SARIMA$(1,1,0) \times (1,1,0)_4$, SARIMA$(1,1,0) \times (0,1,1)_4$ and SARIMA$(1,1,0) \times (1,1,1)_4$ models.

5.2 Parameter estimation and model diagnostic checking

On R, the statistics used to evaluate the considered models are calculated at the same time that the parameters of the model and are stored in the same object of the class "Arima". So, we will fit the models to the sample series and show their parameters at the same time that we diagnose them.

We will start the diagnostic check the significance of the parameters of the models, reconsider them if necessary and then check which is the one with smaller AIC.

Listing 5.1: R output for sp200010

```
Call:
arima(x = log.spur.ts, order = c(2, 0, 0),
    seasonal = list(order = c(0, 1, 0), 4))

Coefficients:
         ar1      ar2
      1.6119  -0.7055
s.e.  0.0774   0.0778

sigma^2 estimated as 0.001403:  log likelihood = 141.84,
aic = -277.69
```

Listing 5.2: R output for sp110110

```
Call:
arima(x = log.spur.ts, order = c(1, 1, 0),
        seasonal = list(order = c(1, 1, 0), 4))

Coefficients:
         ar1     sar1
      0.6988  -0.3563
s.e.  0.0811   0.1119

sigma^2 estimated as 0.001446:  log likelihood = 140.11,
aic = -274.23
```

Listing 5.3: R output for sp110011

```
Call:
arima(x = log.spur.ts, order = c(1, 1, 0),
      seasonal = list(order = c(0, 1, 1), 4))

Coefficients:
          ar1      sma1
       0.7218   −0.6811
s.e.   0.0802    0.1577

sigma^2 estimated as 0.001271:  log likelihood = 144.13,
aic = −282.27
```

Listing 5.4: R output for sp110111

```
Call:
arima(x = log.spur.ts, order = c(1, 1, 0),
    seasonal = list(order = c(1, 1, 1), 4))

Coefficients:
          ar1     sar1      sma1
       0.6968   0.2901   −0.9038
s.e.   0.0845   0.1663    0.1550

sigma^2 estimated as 0.001193:  log likelihood = 145.47,
aic = −282.94
```

Dividing each parameter by their standard error we can see that the only parameter that is not significant from the considered models is the Φ of the SARIMA$(1,1,0) \times (1,1,1)_4$ model. The ratio of the parameter is $0.2901/0.1663 = 1.7444$ and it is the only one smaller than 1.96. Therefore, we can eliminate this model from the considered ones.

Table 5.1: AIC of the potential models fitted to the log-transformed Spanish unemployment rate series

Model	*AIC*
SARIMA$(2,0,0) \times (0,1,0)_4$	-277.69
SARIMA$(1,1,0) \times (1,1,0)_4$	-274.23
SARIMA$(1,1,0) \times (0,1,1)_4$	-282.27
SARIMA$(2,0,0) \times (0,1,1)_4$	-286.31

In Table 5.1 we can see that the model with smaller *AIC* of the considered ones is the SARIMA$(1,1,0) \times (0,1,1)_4$ model. The function `auto.arima` of the "forecast" library allow us to easily calculate the *AIC* of a large set of models, even models that we have initially not considered. It identifies the fitted model for a time series with smaller *AIC* within a given maximum order for the processes. Applying the function to our series, returns the SARIMA$(2,0,0) \times (0,1,1)_4$ model, with an *AIC* of -286.31.

Notice that if we had started the diagnostic by analyzing the residuals of the series, we would had identified that the residuals of the fitted SARIMA$(2,0,0) \times (0,1,0)_4$ model follow a SARIMA$(0,0,0) \times (0,0,1)$ process and we would had considered the SARIMA$(2,0,0) \times (0,1,1)_4$ model.

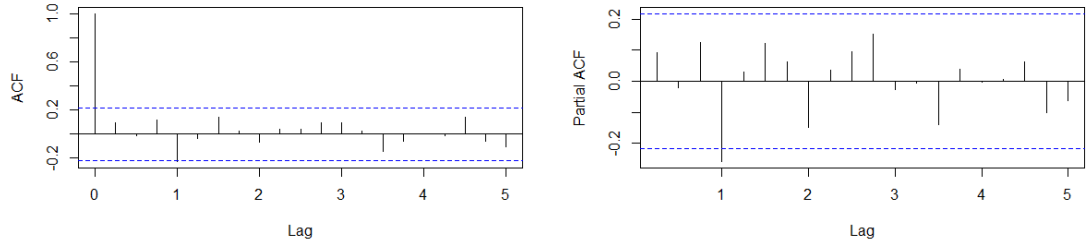

Figure 5.6: Correlogram (left) and partial correlogram (right) of the residuals of the SARIMA$(2,0,0) \times (0,1,0)_4$ model fitted to the log-transformed Spanish unemployment rate series.

Listing 5.5: R output for sp200011

```
Call:
arima(x = log.spur.ts, order = c(2, 0, 0),
      seasonal = list(order = c(0, 1, 1), 4))

Coefficients:
          ar1      ar2     sma1
       1.6827  -0.7180  -0.5771
s.e.   0.0776   0.0755   0.1777

sigma^2 estimated as 0.001199:  log likelihood = 147.15,
aic = -286.31
```

In Listing 5.5 we can see that all the parameters of the SARIMA$(2,0,0) \times (0,1,1)_4$ model are significant and that the process is invertible, since $|\Theta| = 0.5771 < 1$ and causal, since the roots of $1 - \phi_1 x - \phi_2 x^2 = 1 - 1.6827x + 0.7180x^2$ are $1.18015e^{\pm 0.119061i}$, that lie outside the unit circle.

Finally, performing the portmanteau tests on the residuals of the fitted model at lag 8 we get that the p-value for the Box-Pierce test is 0.9281 and the p-value for the LjungBox test is 0.9113. Since both of them are greater than 0.05 and since there are no significant values on the ACF and PACF of the residuals, we do not reject that the residuals are not correlated.

The diagnostic that we have performed concludes that the SARIMA$(2,0,0) \times (0,1,1)_4$ model is the one that fits better the log-transformed Spanish unemployment series. Hence, we will use this model for forecasting.

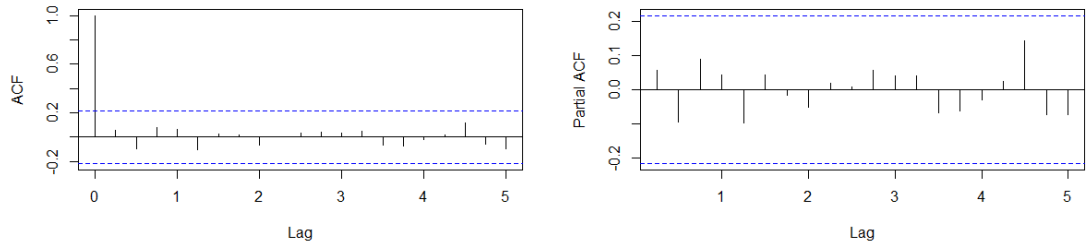

Figure 5.7: Correlogram (left) and partial correlogram (right) of the residuals of the SARIMA$(2,0,0) \times (0,1,1)_4$ model fitted to the log-transformed Spanish unemployment rate series.

5.3 Forecasting

We get the forecasted values of the log-transformed series for the 8 next values in Listing 5.6 using the function `predict` and use them and their standard errors to plot the graph in Figure 5.8

Listing 5.6: R output for p.sp

```
$pred
          Qtr1      Qtr2      Qtr3      Qtr4
2022              2.602433  2.602337  2.576267
2023  2.607806  2.602322  2.606409  2.583199
2024  2.616548

$se
             Qtr1         Qtr2         Qtr3         Qtr4
2022                0.03462427  0.06777357  0.09974046
2023  0.12868042  0.16230575  0.19499427  0.22429201
2024  0.24933459
```

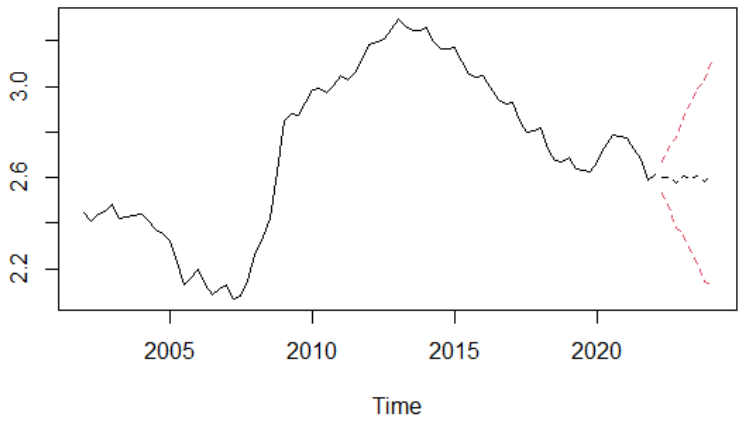

Figure 5.8: Plot of the log-transformed Spanish unemployment rate series with 8 predicted quarters (dashed) and their 95% confidence intervals (red dashed).

Finally, we undo the log-transformation applying the exponential function on each forecast. The forecasted values can be found in Listing 5.7 and The plot of the original data and its forecasted values can be found in Figure 5.9

Listing 5.7: R output for exp(p.sp$pred)

```
      Qtr1     Qtr2     Qtr3     Qtr4
2022          13.49653 13.49524 13.14796
2023 13.56925 13.49503 13.55031 13.23942
2024 13.68838
```

Table 5.2: 95% confidence intervals of the 8 forecasted values of the Spanish unemployment rate series

	Q1	Q2	Q3	Q4
2022		[12.610999, 14.44425]	[11.816545, 15.41241]	[10.813284, 15.98671]
2023	[10.544377, 17.46188]	[9.817852, 18.54946]	[9.246271, 19.85782]	[8.529977, 20.54898]
2024	[8.396809, 22.31465]			

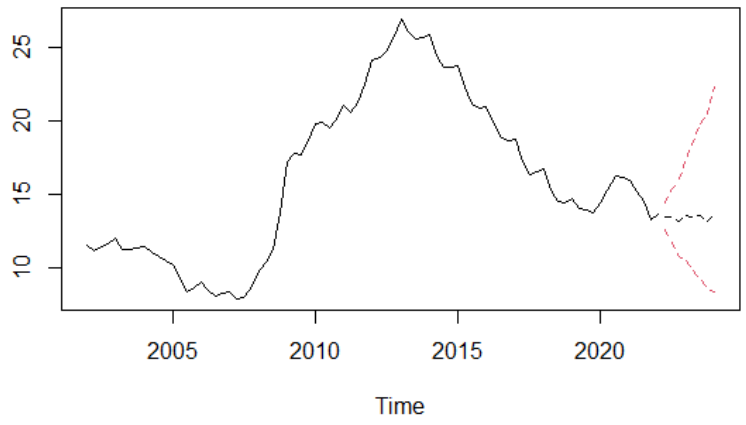

Figure 5.9: Plot of the Spanish unemployment rate series with 8 predicted quarters (dashed) and their 95% confidence intervals (red dashed).

Chapter 6

Conclusions

Even though the SARIMA model selected for forecasting was the best fitting model for the Spanish unemployment rate, we can see in Table 5.2 that the length of the 95% confidence intervals of the forecasted value for the first unknown data is quite narrow, but they increase in length rapidly over time, suggesting that we should take with a grain of salt any forecast for a quarter later than the last of 2022.

This limited accuracy could be explained by different factors, being the first one the lack of observations. Many authors conclude that to try fitting a SARIMA model we should have at least 50 observations of the time series, but preferably more than 100. We have 81 observations of the series, which is over the recommended minimum, but perhaps more observations could have helped identifying undetected underlying processes and making the estimations of the parameters more accurate.

Another possible reason is that the best fitting model has changed over time and including old observations in the analysis does not allow us to identify the current underlying process. It is not daring to think that the behavior of the Spanish unemployment rate of the early 00's is completely different from the actual behavior and therefore it should not be included in the analysis.

In addition, processes like the evolution of the unemployment rate are highly related with other macroeconomic processes. The increase of the unemployment in 2008 and the posterior decrease in 2013 can be attributed as a consequence of the economic recession that took place between those years and the spike in 2020 can be explained by the impact of the COVID-19 in the economy. With the SARIMA models we only analyze the relationships of the series with itself. To expand the knowledge of this process it could be interesting to perform a multivariate analysis with series like the GDP of Spain. This kind of analysis also study the relationships between different time series.

Besides, I want to remark the importance that the evolution of computation and the development of specific software has meant for this branch of mathematics. With just one function on R I have been able to estimate the parameters of tenths of models and compare one of their statistics in a couple of seconds. This would have take several hours if it had to be done by hand.

To conclude, I think it's interesting to have a thought about the phrase *"All models are wrong, but some of them are useful"*, generally attributed to G. E. P. Box, one of the main contributors to the analysis of time series. It is almost impossible to find a model that is an exact representation of the reality. If we were able to find it, it would mean

that there were no random component and it would be a fact rather than a model. With methods like the ones introduced in this thesis we are able to give an approximation for complex processes in a simple model that are close enough to reality and can help us understand how the process has been generated and forecast accurate approximations of future values.

Bibliography

[1] Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle, in Petrov, B. N.; Csáki, F. (eds.)*, 2nd International Symposium on Information Theory.

[2] Akaike, H. (1974). *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19 (6): 716–723.

[3] Box, G. E. P.; Cox, D. R. (1964). *An analysis of transformations (with discussion)*. Journal of the Royal Statistical Society B, 26, 211–252.

[4] Box, G. E. P.; Jenkins, G. M. (1962).*Some statistical aspects of adaptive optimization and control*, Journal of the Royal Statistical Society B, 24, 297–331.

[5] Box, G. E. P.; Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

[6] Box, G. E. P.; Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, rev. ed., Oakland, California: Holden-Day.

[7] Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C.; Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, Hoboken.

[8] Box, G. E. P.; Pierce, D. A. (1970). *Distribution of residual autocorrelations in autoregressive integrated moving average time series models*, J. Am. Stat. Assoc., 65, 1509–1526.

[9] Brockwell, P. J.; Davis R. A. (2006). *Time Series: Theory and Methods*, Springer International Publishing.

[10] Brockwell, P. J.; Davis R. A. (2016). *Introduction to Time Series and Forecasting*, Springer International Publishing.

[11] Cowpertwait, P. S. P.; Metcalfe A. V. (2009). *Introductory time series with R*, Springer.

[12] Ljung, G. M.; Box, G. E. P. (1978). *On a measure of lack of fit in time series models*, Biometrika, 65, 297–303.

[13] Walker, G. (1931). *On periodicity in series of related terms*, Proc. R. Soc., A131, 518–532

[14] Whittle, P. (1951). *Hypothesis testing in times series analysis*. Uppsala: Almqvist & Wiksells Boktryckeri AB.

[15] Wold, H. O. (1938). *A Study in the Analysis of Stationary Time Series*, Almqvist & Wiksell, Uppsala, Sweden; 2nd ed., 1954.

[16] Yaglom, A. M. (1955). *The correlation theory of processes whose n-th difference constitute a stationary process*, Mat. Sb., 37(79), 141.

[17] Yule, G. U. (1927), *On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers*, Philos. Trans. R. Soc., A226, 267–298.

Appendix A

R code

A.1 Chapter 1

Listing A.1: Example 1.1

```
set.seed(1)
x <- rnorm(100, mean = 0, sd = 1)
plot(x, type = "l", xlab = 'Time')
acf(x, main = "")
pacf(x, main = "")
```

Listing A.2: example 1.2

```
set.seed(1)
x <- w <- rnorm(100, mean = 0, sd = 1)
for(t in 2:100) x[t] <- x[t-1] + w[t]
plot(x, type = "l", xlab = 'Time')
acf(x, main = "")
pacf(x, main = "")
```

Listing A.3: example 1.3

```
library(datasets)
plot(JohnsonJohnson)
```

A.2 Chapter 2

Listing A.4: Section 2.2

```
set.seed(1)
AR10.9 <- arima.sim(list(order = c(1,0,0), ar = 0.9), n=100)
acf(AR10.9, main = "")
pacf(AR10.9, main = "")
AR1n0.9 <- arima.sim(list(order = c(1,0,0),
    ar = -0.9), n=100)
acf(AR1n0.9, main = "")
pacf(AR1n0.9, main = "")
```

Listing A.5: Section 2.3

```
set.seed(1)
MA10.9 <- arima.sim(list(order = c(0,0,1), ma = -0.9), n=100)
acf(MA10.9, main = "")
pacf(MA10.9, main = "")
MA1n0.9 <- arima.sim(list(order = c(0,0,1), ma = 0.9), n=100)
acf(MA1n0.9, main = "")
pacf(MA1n0.9, main = "")
```

Listing A.6: Section 2.4

```
set.seed(1)
ARMA11 <- arima.sim(list(order = c(1,0,1), ar = 0.9,
    ma = -0.7), n=1000)
acf(ARMA11, main = "")
pacf(ARMA11, main = "")
```

A.3 Chapter 3

Listing A.7: Example 3.1

```
set.seed(1)
ARIMA110 <- arima.sim(list(order = c(1,1,0), ar = 0.8),
    n=100)
plot(ARIMA110)
plot(diff(ARIMA110))
acf(ARIMA110, main = "")
pacf(ARIMA110, main = "")
acf(diff(ARIMA110), main = "")
pacf(diff(ARIMA110), main = "")
```

Listing A.8: Example 3.2

```
set.seed(1)
ARIMA110 <- arima.sim(list(order = c(0,1,1), ma = -0.8),
    n=100)
plot(ARIMA011)
plot(diff(ARIMA011))
acf(ARIMA011, main = "")
pacf(ARIMA011, main = "")
acf(diff(ARIMA011), main = "")
pacf(diff(ARIMA011), main = "")
```

Listing A.9: Section 3.2

```
install.packages ("astsa")
library (astsa)
set.seed(1)
SARIMA1 <- sarima.sim(sar = 0.7, S = 12, n=500)
acf(SARIMA1, main = "")
pacf(SARIMA1, main = "")

set.seed(1)
SARIMA2 <- sarima.sim(ar = 0.7, sar = 0.7, S = 12, n=500)
acf(SARIMA2, main = "")
pacf(SARIMA2, main = "")
```

A.4 Chapter 4

Listing A.10: Monthly total international airline passengers from January 1949 to December 1960

```
plot(AirPassengers)

plot(log(AirPassengers))
plot(diff(log(AirPassengers)))
plot(diff(log(AirPassengers), 12))
plot(diff(diff(log(AirPassengers)), 12))

acf(log(AirPassengers), 50, main = "")
acf(diff(log(AirPassengers)), 50, main = "")
acf(diff(log(AirPassengers), 12), 50, main = "")
acf(diff(diff(log(AirPassengers)), 12), 50, main = "")

pacf(diff(diff(log(AirPassengers)), 12), 50, main = "")

ap011011 <- arima(log(AirPassengers), order = c(0,1,1),
    seasonal = list(order = c(0,1,1), 12))
ap011110 <- arima(log(AirPassengers), order = c(0,1,1),
```

```
    seasonal = list(order = c(1,1,0), 12))
ap110011 <- arima(log(AirPassengers), order = c(1,1,0),
    seasonal = list(order = c(0,1,1), 12))
ap110110 <- arima(log(AirPassengers), order = c(1,1,0),
    seasonal = list(order = c(1,1,0), 12))
ap111111 <- arima(log(AirPassengers), order = c(1,1,1),
    seasonal = list(order = c(1,1,1), 12))

ap011011$coef
ap011110$coef
ap110011$coef
ap110110$coef
ap111111$coef

ap011011$aic
ap011110$aic
ap110011$aic
ap110110$aic
ap111111$aic

ap011011
ap111111

acf(ap011011$residuals, 50, main = "")
pacf(ap011011$residuals, 50, main = "")
ap010011 <- arima(log(AirPassengers), order = c(0,1,0),
    seasonal = list(order = c(0,1,1), 12))
acf(ap010011$residuals, 50, main = "")
pacf(ap010011$residuals, 50, main = "")

Box.test(ap011011$residuals, lag = 24, "Box-Pierce")
Box.test(ap011011$residuals, lag = 24, "Ljung-Box")
Box.test(ap010011$residuals, lag = 24, "Box-Pierce")
Box.test(ap010011$residuals, lag = 24, "Ljung-Box")

p.ap <- predict(ap011011, 24)
p.ap
ts.plot(log(AirPassengers), p.ap$pred, p.ap$pred - 1.96 *
    p.ap$se, p.ap$pred + 1.96 * p.ap$se, lty = c(1,2,2,2),
    col = c(1,1,2,2))
ts.plot(AirPassengers, exp(p.ap$pred), exp(p.ap$pred - 1.96 *
    p.ap$se), exp(p.ap$pred + 1.96 * p.ap$se),
    lty = c(1,2,2,2), col = c(1,1,2,2))
```

A.5 Chapter 5

Listing A.11: Analysis of the Spanish unemployment rate

```
install.packages ("rjson")
library (rjson)


#Access the data
spur <- fromJSON(file = "https://servicios.ine.es/wstempus/
js/ES/DATOS_SERIE/EPA86913?date=20020101:20220511")


#Format the data as a ts object
spur.values <- seq(1 : length(spur$Data))
for (i in 1:length(spur$Data)){
    spur.values[i] <- spur$Data[[i]]$Valor
}
spur.ts <- ts(spur.values, st = 2002, fr = 4)


#Model identification
plot(spur.ts)

log.spur.ts <- log(spur.ts)
plot(log.spur.ts)
plot(diff(log.spur.ts))
plot(diff(log.spur.ts, 4))
plot(diff(diff(log.spur.ts, 4)))

acf(log.spur.ts, 20, main = "")
acf(diff(log.spur.ts), 20, main = "")
acf(diff(log.spur.ts, 4), 20, main = "")
acf(diff(diff(log.spur.ts, 4)), 20, main = "")

pacf(diff(log.spur.ts, 4), 20, main = "")
pacf(diff(diff(log.spur.ts, 4)), 20, main = "")

#Parameter estimation and model diagnostic checking
sp200010 <- arima(log.spur.ts, order = c(2,0,0),
    seasonal = list(order = c(0,1,0), 4))
sp110110 <- arima(log.spur.ts, order = c(1,1,0),
    seasonal = list(order = c(1,1,0), 4))
sp110011 <- arima(log.spur.ts, order = c(1,1,0),
    seasonal = list(order = c(0,1,1), 4))
sp200010
sp110110
sp110011
sp110111
```

```
auto.arima(log.spur.ts)

acf(sp200010$residuals, 20, main = "")
pacf(sp200010$residuals, 20, main = "")

sp200011 <- arima(log.spur.ts, order = c(2,0,0),
    seasonal = list(order = c(0,1,1), 4))
sp200011

acf(sp200011$residuals, 20, main = "")
pacf(sp200011$residuals, 20, main = "")

Box.test(sp200011$residuals, lag = 8, "Box-Pierce")
Box.test(sp200011$residuals, lag = 8, "Ljung-Box")


#Forecasting
p.sp <- predict(sp200011, 8)
p.sp

ts.plot(log.spur.ts, p.sp$pred, p.sp$pred - 1.96 *
    p.sp$se, p.sp$pred + 1.96 * p.sp$se, lty = c(1,2,2,2),
    col = c(1,1,2,2))
ts.plot(spur.ts, exp(p.sp$pred), exp(p.sp$pred - 1.96 *
    p.sp$se), exp(p.sp$pred + 1.96 * p.sp$se),
    lty = c(1,2,2,2), col = c(1,1,2,2))
```