NATIONAL UNIVERSITY OF ENGINEERING

COLLEGE OF ECONOMICS AND STATISTICAL ENGINEERING AND SOCIAL SCIENCES

STATISTICAL ENGINEERING PROGRAM

EF012 PROJECT WORKSHOP

Design of a Statistical Processing System for Analyzing the Impact of Covid-19 on the Consumer Price Index cpi in Lima City.

by:

GARCÍA LOAYZA JESUS FRANCISCO

Professor

SIKOV ANNA

LIMA – PERU

2020

Abstract

The country is currently going through a health crisis, which is the product of the pan- demic caused by the recent virus called COVID-19, the rapid expansion of this virus has caused the health crisis to expand into an economic crisis, and that the virus forces people to stay home. These events generated that the government saw the need to declare the country in a state of emergency, the consequence of this decision brought with it an increase in the demand for basic necessities, causing a large part of the prices of the products to increase. This work will be based on measuring the impact of the Consumer Price Index (CPI) during the months of April, May, June, July and August; These months present the second quarter of the year and part of third quarter, a period in which the state of emergency was just beginning in the country and people were still not adapted.

To measure the Impact that was had in the second quarter and part of the third quarter of 2020, it is necessary to know what was or what were the forecasts for April, May, June, July and August. These forecasts will allow comparison with the real values of those months. It is known that the level of the CPI tends to grow due to the constant increase in prices suffered by products and services, but in a stable economy said inflation grows in a controlled manner. By measuring the impact that was had in the month of April, May, June, July and August, this month it will be assessed whether the Peruvian economy remained under control.

The information available is the monthly CPI levels of Lima Metropolitana from January 2001 to March 2020, with these data a time series will be built under the Box-Jenins methodology and thus establish a model that allows forecasting CPI levels for the months of April, May, June, July and August. In the preliminary analysis of the time series of the CPI levels of Lima Metropolitana, different components were identified, such as trend and seasonality.

The positive trend of the CPI is a natural component over time, in the same sense there is seasonality in the CPI levels because throughout the year there are months where the eco- nomic.

Therefore, it was proposed that the SARIMA model would be the most appropriate as it allows taking into account the existence of the seasonal component. In addition, for the evaluation of these models, it was decided to divide the amount of data into two groups, the first group would help to build the SARIMA candidate models and the second group would allow the validation of said models.

General Index

Chapter 1

INTRODUCTION

Peru throughout its history has experienced various economic crises that have directly affected the country's economy, currently several countries are in crisis as a result of the recent COVID-19 pandemic, which not only affects people's health but the economy directly. One of the indicators that shows the behavior of the economy is the Consumer Price Index (CPI), which represents the level of inflation in a country. It must be understood that inflation is the sustained and generalized increase in the prices of goods and services that occurs during a certain period of time. This as a consequence decreases the purchasing power of the currency, that is, with the same amount of money the acquisition of goods and services is limited.

In Peru, the Consumer Price Index is calculated through the National Survey of Family Budgets (ENAPREF) carried out by the National Institute of Statistics and Information (INEI). This survey makes it possible to measure the consumption pattern of households and also achieves knowledge of the income and expenditure of families. The objective of this survey is to verify in which living conditions the families live, determining the quantities, the cost of the items consumed and the services used, based on which the weights are considered for the elaboration of the family baskets.

There is a basic basket which is used as the basis for calculating the CPI. This basic basket is made up of 8 large consumption groups: a) food and beverages; b) Clothing and footwear, c) rental of housing, fuel and electricity; d) Furniture, fixtures and home maintenance; e) Health care and maintenance and measured services; f) Transportation and communications; g) recreation, entertainment, cultural and educational services; h) Other goods and services.

The target population to obtain the CPI are those over 14 years of age who also do not reside in collective housing, such as: hotels, hospitals, barracks, nursing homes, prisons, among others.

It is important to note that the rise in the CPI has both a positive and a negative side. The manifestation of controlled inflation is a reflection of the economic growth of a country, as it shows that people acquire higher cost goods and services because they have also increased their income level, that is, consumption is encouraged and more economic activity arises.

On the contrary, uncontrolled variations in inflation generate uncertainty in people, which is reflected in higher interest rates, which discourage investment and saving because individuals seek to protect themselves from inflation.

Some of the most representative factors for the growth of the CPI are due to the increase in the demand for products found in the basic family basket or also due to the shortage of products in said basket, since both factors bring as a consequence the increase in the cost of the products.

The current reality in the face of COVID-19 has caused many people to panic as a consequence, these people are in isolation and in the constant search for essential products to feel safer. In addition, the quarantine has caused many economic sectors to be paralyzed, so that production has also decreased.

So far it is reflected that the impact that this reality has had at present is very different from the forecasts made at the end of 2019, which indicated that the annual variation of the CPI would not exceed 3%, which in other words is 0.25% monthly.

1.1. Presentation of the Problem

Peru throughout its history has experienced various economic crises, which have directly affected people and the economy of the country. One of the most important events related to these crises was hyperinflation that took place in the 1980s, during the government of Alan Garcia. This period of crisis was so strong that it is still present in the memory of the majority of Peruvians, so there is a great interest of the population for the recurrent knowledge of this indicator. Another period of crisis was the one that happened in 2008, this time as a consequence of the international economic crisis that was experienced at that time, which caused the Peruvian GDP to decline. In this period, external demand suffered a significant drop, later bringing a decrease in Peruvian industrial production and an increase in inflation.

All this shows how inflation levels, measured by the CPI, are highly sensitive to changes or international crises. Therefore, it is evident how complicated it is to predict how the CPI levels will behave recurrently

At present, there is a terrible health crisis caused by the rapid expansion of the coronavirus, in this sense the Peruvian government made the decision to lead the population to social isolation, in addition to that there is an evident panic of the population during the first months of arrival of the virus.

All these events generate a constant search for basic necessities of the people to feel safer or feel more secure, therefore the consumption of these products increased. In addition, the quarantine has caused many economic sectors to be paralyzed, consequently there is a notable decrease in the level of production of various products.

It is evident that the CPI levels will vary in the first months affected by COVID-19, but it is unknown what impact or change the CPI levels have had as a result of the coronavirus, especially in the months of March, April and May; The are the first months affected by the coronavirus.

In order to have the desired knowledge of the levels of CPI product of the coronavirus in these affected months, it will be necessary to apply techniques or methodologies that allow modeling the behavior of its components over time. The preliminary analysis showed the remarkable behavior of the components of the CPI time series, which respond to a marked positive trend and seasonality in the months of April, July and December, where April is one of the months of interest. It should be noted that information from the last two decades is being used (from January 2001 to March 2020)

1.2. Objectives

1.2.1 General Objective

- Determine the impact of the CPI level caused by COVID-19 in Metropolitan Lima, during the months of April, May and June.

1.2.2. Specific Objectives

- Estimate the best time series models for predicting the Metropolitan Lima CPI without taking into account the behavior of Covid-19
- Determine the impact of the predicted Metropolitan Lima CPI without taking into account the Covid-19 versus the real value of the Metropolitan Lima CPI taking into account the Covid-19.

Chapter 2

Theoretical Foundations

2.1. Background

Victor Carrillo in his study called 'Peruvian Hyperinflation: A Systematic Review', shows how the crisis, in this case political, greatly affects the level of Peruvian society. Carrillo mentions that during the 80s a strong crisis was experienced as a result of the establishment of democracy after the paramilitary dictatorship that had been going on in the years prior to the 80s.

In 1988 the Peruvian economy shows one of the worst falls in the level of GDP in history, this fall was calculated at approximately 10%, resulting in an inflation level of 667%. To date, Peru has not registered a higher level of inflation than this. During this period hundreds of businesses went bankrupt due to the loss of capital value.

In 2008, Peru registered a marked slowdown in economic activity as a result of the effects of the international financial crisis. GDP growth fell from 9.8% in 2008 to 0.8% in 2009, mainly due to the sharp drop in external demand, with the consequent decline in industrial production, a strong adjustment process inventories and a significant reduction in private investment, as a result of lower demand and the uncertainty about the future of the international economy that prevailed at the end of 2008 and during 2009.

It was not until 2009 that the level of inflation, measured through the Lima consumer price index, fell significantly and reached 0.04% in the first 10 months of the year (0.7% in 12 months), due to the drop in the international prices of food and hydrocarbons, which had an impact on the reduction of the prices of transportation items and of domestic fuels and electricity. This point tells us that not only national crises directly affect the level of inflation in the economy, but also the increasing level of inflation is due to changes in the economy of other countries.

2.2. General Definitions

- Index: It is an indicator of central tendency of a set of elements that is usually expressed as a percentage. All index numbers have certain characteristics in common: Index numbers are ratios of one quantity in a current period to another quantity in a base period.

- Economic Indicator: A statistical indicator that is issued published by a government, this indicator shows current economic growth and stability. Common indicators include employment rates, Gross Domestic Product (GDP), inflation, retail sales, and so on.

- Price Index: Statistical indicator of the variation in the most representative prices of goods and services, which allows for intertemporal comparisons of their level and is also used to determine economic measures against inflationary or deflationary processes.

- Consumer Price Index: Index weighted according to the consumption that an average family unit carries out and that measures the general price level with respect to a previous period. It is the most used indicator to measure inflation and its acronym is CPI.

- Basic Basket: it is a set of basic necessity products and services that an average family needs to subsist for a certain period of time (usually per month), be it food, hygiene, clothing, health, transportation, entertainment, etc. others.

- Inflation: It is the sustained and generalized increase in the price level of goods and services, measured with respect to a stable purchasing power.

- National Survey of Family Budgets: It has to obtain information on the nature and destination of consumer spending, as well as on various characteristics related to the living conditions of households.

- Coronaviruses: They are an extensive family of viruses that can cause disease in both animals and humans. In humans, various coronaviruses are known to cause respiratory infections that can range from the common cold to more serious illnesses such as Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS). (WHO 2020)

- Covid-19: It is the infectious disease caused by the coronavirus discovered in Wuhan (China) in December 2019, which affects the respiratory tract

2.3. Statistical Definitions

- Time Series: It consists of a set of collected data recorded or observed in successive increments of time, a time series will generally have a random character, and can be interpreted as a sample of size one taken in successive periods of time, considered as a performance of a stochastic process (Uriel & Peiró, 2000, page 25).

- Components of a time series: The four components found in a historical series are (Hanke & Reitsch, 1996, p. 319)
  - Trend: it is the long-term component that constitutes the basis for the growth or decline of a historical series.
  - Cyclical component: it is the set of fluctuations in wave form or cycles around the trend, affected by external factors, these cyclical patterns tend to repeat themselves in the data from time to time.
  - Seasonal component: refers to a regularly recurring pattern of change over time.
  - Random Component - Measures the variability of the time series after the other components are removed.

- Stochastic process: a stochastic process is a family of random variables that we assume to be defined in the same probability space (Morettin, 2002, p. 15). For a fixed time t, Zt constitutes a random variable whose values form the state space and the sets of instants in time of the parameter space. The mean and variance are defined as:

$$E\{Z_t\} = m_t$$

$$V\{Z_t\} = s2t$$

- Stationary series: it is one whose average does not change over time (Hanke & Reitsch, 1996, p. 108).

- Ergocity: To make statistical inferences with a single random variable, it is necessary to resort to the repetition of experiments. It can be shown that, when a stochastic process meets certain conditions, it is possible to consistently estimate characteristics from a realization of the same, the processes that meet such conditions are called ergodic. A necessary but not sufficient condition is that the limit of ρk = 0, when k tends to infinity (auto correlation function) (Uriel & Peiró, 2000, pp. 27 - 30).

- White noise: it is a sequence of independent and identically distributed random variables with mean and finite variance. When it is normally distributed with zero mean and variance σ2, it is called Gaussian White Noise (Tsay, 2005, p. 31).

- Random walk: We call random walk to a stochastic process whose first differences form a white noise process (Hanke & Reitsch, 1996, p. 441).
- Stationarity: It is said that a stochastic process is stationary when, when making the same displacement in time of all the variables of any finite joint distribution, it turns out that this distribution does not vary (Uriel & Peiró, 2000, pp. 27 - 30).

- Simple Autocorrelation Function (FAC): It is the existing correlation between a variable lagged one or more periods and the same variable.

- Partial Autocorrelation Function (FACP): It is a function that identifies the degree of relationship between the real values of a variable and its previous values, while the effects of the other variables are kept constant.

- Parsimony: Rule that establishes that simple models are preferable to complicated ones. The model used should require the fewest possible number of parameters that adequately represent the pattern of the data. Parsimony indicates the number of parameters for an adequate representation of the model.

- Forecast error: or residual, it is the difference between the observed values and the forecast values.

$$e_t = Z_t - \widehat{Z}$$

Where

Z_t = Value of a time series in the period t
$\hat{Z}$= Forecast value for Z_t

2.4. Time Series Models

One of the most common assumptions made when working with a time series is that it is stationary. But, a large part of the series that are found in practice, present some form of non-stationarity, such as the economic and financial series that generally present increasing or decreasing trends, or even variance with changes over time. The series that do not present stationarity can present the so-called forms of explosive non-stationarity and those that do not present explosive behavior. This last type of non-stationarity is called homogeneous; The series can be stationary, floating around one level, for a time, then changing levels and floating around another, and so on, or changing inclination, or both.

A series can be stationary for a very long period or it can be stationary for only a very short periods, changing in level and / or inclination. This is just an intuitive idea of stationarity and now some definitions will be represented: weak (or broad or second order) and strong (or strict) stationarity.

Definition 3.4.1. A stochastic process {yy; t ∈ T} is stationary in a strong (or strictly stationary) sense if $F_{y_{t_1},\cdots,y_{t_n}}(y_1,\cdots,y_n) = P(y_{t_1}(\omega) \leq y_1,\cdots,y_{t_n}(\omega) \leq_n)$, for all $t_1 < t_2 < \cdot \cdot \cdot \leq t_n \in T$, $n \geq 1$ and for all ∈ T , that is, if all distributions with finite dimensions,

$F_{y_{t_1},\cdots,y_{t_n}}(y_1,\cdots,y_n) = F_{y_{t_1}+,\cdots,y_{t_n}+}(y_1,\cdots,y_n)$, (joint probability distribution function), remain the same over time.

In practice it is more difficult to use this definition and it is customary to define a less strict form of stationarity, as follows:

Definition 3.4.2. A stochastic process {Z_t, t = 1, 2, · · ·}, is weakly stationary (or second-order stationarity) if E(Z_t) = μ (constant media), var(Z_t) = σ^2 (constant variance) y cov(Z_t, yt+k) = γ(k) (covariance between Z_t and y_{t+k} a function that depends on k), for all t ∈ T and k ∈ N.

Definition 3.4.3. A stochastic process {y_y; t ∈ T} is Gaussian if, for any set $t_1 < t_2 < \cdot \cdot \cdot \leq t_n \in T$, $n \geq 1$, the random variables $y_{t_1}, \cdot \cdot \cdot, y_{t_n}$ have an n-varied normal distribution.

To define stationarity for linear models we need to perform Wald's theorem.

Theorem 3.4.1. (Wald's theorem) Let Z_t a stochastic process, then there is a stochastic process a_t, t = 1, 2, · · · such that

$$Z_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j} = \mu + \psi(b) a_t \tag{1}$$

where ψ(b) is a polynomial defined by $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \cdot \cdot \cdot$, con $\psi_0 = 1$, $(a_t) = 0$, t and $E(a_t a_s) = 0$ for all s ≠ t y $E(a^2) = \sigma_a^2$ where σ_a^2 is a constant variance.

Let a_t, with t = 1, 2, · · ·, independent and identically distributed random variables (iid), normal with zero mean and constant variation σ_a^2, namely $a_t \sim N(0, \sigma_a^2)$, in this case, $\{a_t, t \geq 0\}$ it is called white noise.

The process $Z_t = \psi(B)a_t$ is stationary in a weak sense if the roots of the polynomial ψ(B) = 0 they are all in or on the circle of unit radius. Let at is a process i.i.d. $N(0, \sigma_a^2)$, this condition of the roots ψ(B) = 0 guarantee stationarity in the strong sense of the process.

2.5. Functions of a Stationary Stochastic process

A stationary stochastic process is defined whether it meets conditions in the weak sense, that is, its mean and constant variance are known.

2.5.1. Autocovariance function:

It is defined by the different values that the covariance takes when one period is fulfilled with respect to the other. Analytically it is expressed as:

$$z_k = Cov(z_t, z_{t-k}) = E[(z_t - \mu)(y_{t-k} - \mu)]$$

Where, it can be shown that when "k.” is zero, we would have the variance of the function

$$z_0 = Cov(z, z_{t-0}) = E[(z_t - \mu)^2] = \sigma^2 \quad (2)$$

2.5.2. Autocorrelation function:

It is the cross correlation of a time variable with itself when compared to another period. It is usually used when you want to analyze how the observations of the past influence the current observation.

- Simple autocorrelation function (ACP): The simple autocorrelation function of a series provides the linear dependency structure of the series, that is, to see what degree of dependence the data shows now with data from k previous periods.

$$\rho_1 = \frac{Cov(z_t, z_{t-1})}{\sqrt{var(z_t)}\sqrt{var(z_{t-1})}}$$

Given the assumption of constant variance, $var(z_t) = var(x_{t-1})$

$$\rho_1 = \frac{Cov(z_t, z_{t-1})}{var(z_t)}$$

In general, for a lag of k periods we have to

$$\rho_k = \frac{Cov(z_t, z_{t-k})}{var(z_t)} \quad (3)$$

And when k = 0.

$$\rho_0 = \frac{Cov(z_t, z_t)}{var(z_t)} = \frac{var(z_t)}{var(z_t)} = 1$$

- Partial autocorrelation function (PACP): The partial autocorrelation of order k, measures the linear dependence of y_t and its lag y_{t-k} after removing the effect of the intermediate lags on both.

The partial autocorrelation of order k, is the coefficient of the partial regression in the population φ_{kk} in the autocorrelation of order k-th of the form.

$$z_t = \varphi_{0,k} + \varphi_{1,k} y_{t-1} + \cdot \cdot \cdot + \varphi_{k,k} y_{t-k} + \varepsilon$$

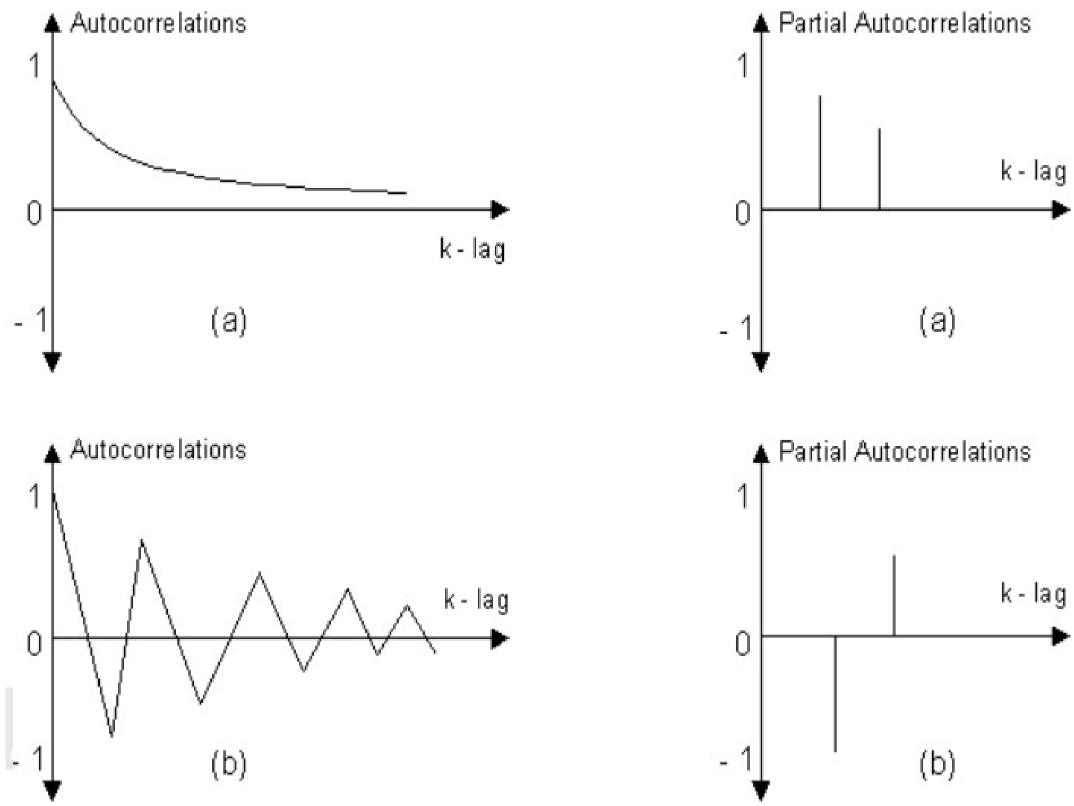

Figure 3.1: *Simple autocorrelation function and partial autocorrelation*

2.6. Smoothing a Time Series

Smoothing is a widely used time series technique to remove noise from the underlying data to help reveal important characteristics and components of the series (e.g. trend, seasonality, etc.). However, you can also use smoothing to fill in missing values and / or make a forecast.

2.6.1. Weighted Moving Averages

A moving average is usually used for time series data, where it is required to smooth out short-term fluctuations and highlight long-term trends or cycles. It is important to note that a weighted moving average contains multiplying factors to give different weights to the data at different positions in the sample window.

$$\hat{Z}_t = \frac{\sum_{i=1}^{n} W_i Z_{t-i}}{\sum_{i=1}^{n} W_i}$$

Where

- W_i is the i-th position weight factor
- Z_t is the observer value at time t
- $\hat{Z}_t$ is the smoothed value at time t

2.6.2. Holt-Winter exponential smoothing

The Holt-Winters multiplicative exponential smoothing method is a robust forecasting method for seasonal time series with additive trend.

$$\hat{F}_t(m) = (S_t + m \times b_t) \times C_{t-L+m}$$

$$S_{t>L} = \alpha(Z_t/C_{t-L}) + (1-\alpha)(S_{t-1} + b_{t-1})$$

$$b_{t>L} = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1}$$

$$C_{t>L} = \gamma(Z_t/S_t) + (1-\gamma)C_{t-L}$$

Where

- Z_t is the observation of the time series at time t.
- L is the length of the station.
- S_t is an estimate of the level component smoothing.
- b_t is an estimate of the softening of the trend component.
- C_t is an estimate of the smoothing of the seasonal index component.
- α is the level of smoothing coefficient,
- β is the trend of the smoothing coefficient.
- γ is the seasonal smoothing coefficient.
- $\hat{F}_t(m)$ is the forward-step smoothing forecast value m for X at time t

The smoothing coefficient α is used again to control the adaptation speed locally but a second smoothing constant β is used to control the degree of the trend and, finally, a third smoothing constant γ is presented to control the degree. of local seasonal indices carried out in forecast periods of multiple steps ahead.

2.7. Forecast models with one variable

This model predicts future values of a time series based only on previous values of the same time series. When using a univariate model, the above data is analyzed in order to identify a data pattern. Then, with the assumption that this will continue in the future, this pattern is extrapolated with the general object of predictions (Bowerman, Koehler, & O'Connell, 2007, p. 11)

2.7.1. Stationary Linear Models

a) Autoregressive Model: AR(p)

Autoregressive models (AR) express Z_t as a linear function of a certain number of real values to Z_t, that is, we write Z_t, as a function of the past values of the series itself, and we include in the expression a disturbance term or error, ε_t, which we assume behaves like white noise.

Their autocorrelation functions gradually drop to zero, while the partial autocorrelation coefficients drop to zero after the first lag period.

Model:

$$AR(p) : Z_t = \varphi_0 + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdot\cdot\cdot + \varphi_{p-1} Z_{t-p+1} + \varphi_p Z_{t-p} + \varepsilon_t \quad (5)$$

Where

- Z_t: Dependent variable
- Z_{t-1}, Z_{t-2}, $\cdot\cdot\cdot$,Z_{t-p}: Dependent variables that are lagged by a specific number of periods.
- $\varphi_0, \varphi_1, \varphi_2, \cdot\cdot\cdot , \varphi_p$ Regression Coefficients
- E_t: Residual term representing random events not explained by the model
- p: It is the number of periods of previous observations to include in the forecast for the next period.

b) Moving Average Model: MA(q)

Moving Average (MA) models provide Z_t forecasts based on a linear combination of previous errors. Your autocorrelation coefficient drops to zero after the ”q” lag period and its partial autocorrelation coefficient gradually drops to zero.

Model:

$$MA(q) : Z_t = \theta_0 + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_{q-1}\varepsilon_{t-q+1} - \theta_q\varepsilon_{t-q}$$

- Z_t: Dependent variable
- $\theta_0, \theta_1, \cdot\cdot\cdot, \theta_{tp}$:: Specific weight
- ε Residue or error

c) Autoregressive Moving Average Model: ARMA(p, q)

The Autoregressive and Moving Average Models *combine to form the general model called ARMA.*

In the ARMA Model (p, q) there will be "p" Autoregressive terms and "q" moving average terms. The ARMA (p, q) models offer the potential to fit models that could not be adequately fitted by the AR and MA models alone.

Model:

$$\begin{aligned} ARMA(p,q) : Z_t = \; & \phi_0 + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_{p-1} Z_{t-p+1} + \phi_p Z_{t-p} + \\ & \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_{q-1}\varepsilon_{t-q+1} - \theta_q\varepsilon_{t-q} \end{aligned} \tag{7}$$

d) Seasonal Autoregressive Moving Average Model: SARMA(p, q)(P, Q)

The SARMA models have the same moments and correlograms as the ARMA models but instead of considering the regular delays (1, 2, 3, etc.) the seasonal delays (s, 2s, 3s, etc.) are considered.

Model

$$\begin{aligned} AR(p): Z_t = \ & \phi_0 + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_{p-1} Z_{t-p+1} + \phi_p Z_{t-p} + \\ & \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_{q-1} \varepsilon_{t-q+1} - \theta_q \varepsilon_{t-q} + \\ & \Phi_1 Z_{t-s} + \Phi_2 Z_{t-2s} + \cdots + \Phi_P Z_{t-Ps} + \Phi_p Z_{t-s} - \\ & \Theta_s \varepsilon_{t-s} - \theta_2 \varepsilon_{t-2s} - \cdots - \theta_Q \varepsilon_{t-Qs} \end{aligned} \quad (8)$$

Where:

- P: is the number of seasonal autoregressive terms.
- Q: is the number of seasonal moving averages.

2.7.2. Non-stationary linear models

a) Autoregressive Integrated Moving Average Model: ARIMA (p, d, q)

Box and Jenkins (1970 and 1976) considered an extension of the ARMA models to deal with certain special types of non-stationary series, considering that such non-stationary series can become so if they are differentiated a sufficient number of times, leading to models ARIMA.

Autoregressive integrated moving average (ARIMA) models constitute a particular class of non-stationary processes, where "p" denotes the number of autoregressive terms, "d" the number of times the series must be differentiated to become stationary and "q" the number of terms of moving averages and is defined as:

$$\begin{aligned} ARIMA(p, q): W_t = \ & \phi_0 + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_{p-1} Z_{t-p+1} + \phi_p Z_{t-p} + \\ & \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_{q-1} \varepsilon_{t-q+1} - \theta_q \varepsilon_{t-q} \end{aligned} \quad (9)$$

Where

- W_t: $Z_t - Z_{t-d}$ Dependent variables
- d Number of Times that the series must be differentiated "p" to become stationary.

The ARIMA (p, q) model is usually written as follows:

$$\phi_p(B)\nabla^d Z_t = \theta_q(B)\varepsilon_t \tag{10}$$

b) Seasonal Autoregressive Integrated Moving Average Model: SARIMA (p, d, q)(P, D, Q)

It is similar to the SARIMA model, in this type of molds it constitutes a non-stationary process, where they have components such as; p, q, P, Q, d and D that represent the autoregressive terms, averages or moving averages and number of differentiations seasonally.

The SARIMA (p, d, q) (P, D, Q) model is usually written as:

$$\phi(B)\Phi_{12}(B)\nabla_{12}^{D}\nabla^d Z_t = \theta(B)\Theta_{12}(B)\varepsilon_t$$
$$(1-B)^d(1-B^{12})^D\phi(B)\Phi_{12}(B)Z_t = \theta(B)\Theta_{12}(B)\varepsilon_t \tag{11}$$

$$(1-\sum_{k=1}^{p}\phi_k B^k)(1-\sum_{k=1}^{P}\Phi_k^{12}B^k)(1-B^{12})^D(1-B)^d z_t = (1-\sum_{k=1}^{q}\theta_k B^k)(1-\sum_{k=1}^{Q}\Theta_k^{12}B^k)\varepsilon_t \tag{12}$$

- ∇: Symbolizes the existence of differentiation
- D: Shows the number of seasonal differentiations

2.8. Tests to contrast hypotheses in time series

For the correct adjustment of a typical time series model, it is necessary that the series to be proposed meet different conditions such as stationarity, normality and the non-correlation of the residuals of the adjusted model; and finally, contrast the parameters of the proposed model.

1. Increased Test de Dickey Fuller: The time series has a unit root (there is no stationarity)

$$H_0 : \gamma = 0 \ vs \ H_1 : \gamma < 0$$

The hypothesis test assumes that the series Zt ~ AR(p) can be expressed in its differentiated form in the following polynomial form:

$$\nabla Z_t = \alpha + \beta t + \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \nabla Z_{t-i} + e_t \tag{13}$$

Where:

- α: is a constant
- β: is the coefficient of the trend over time
- γ and δ_i: are the coefficients of each lag of the differentiated series.

The statistical value of this test is:

$$DF_\gamma = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

Where $SE(\hat{\gamma})$ refers to the error of the estimated coefficient, in addition this statistic does not follow a notable distribution, in fact it follows an asymmetric distribution called the Dickey Fuller distribution, which the higher the value (negative), the greater the chance of rejecting the null hypothesis. This test allows us to contrast the existence of a unit root.

2. Ljung Box test: The residuals of the adjusted SARIMA model $(p_i, d_i, q_i)(P_i, D_i, Q_i)_{12}$ are not correlated (white noise).

$$H_0 : \rho_1 = \rho_2 = \rho_3 = ... = \rho_m = 0 \qquad vs \qquad H_1 : \exists \rho_{1\leq k\leq m} \neq 0 \tag{14}$$

   The contrast of the hypothesis tests the non-correlation of the residuals of the adjusted model, that is, the residuals have the behavior of white noise. The statistic for this test is:

$$Q(m) = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}^2(k)}{T-k} \sim \chi^2(m)$$

   where m is the established lag limit, ρˆ(k) is the estimated autocorrelation in lag k, and t= 1, 2, ..., T. Therefore, if $Q > \chi^2(m)$ the null hypothesis is rejected, or in other words the p-value <0.05, it is concluded that the residuals do not behave like RB.

3. Wald test: The parameter θ_i of the SARIMA model $(p_i, d_i, q_i)(P_i, D_i, Q_i)_{12}$ adjusted is significant:

$$H_0: \theta_i = 0 \quad vs \quad H_1: \theta_i \neq 0$$

This contrast proves that the parameter associated with component i of the SARIMA model is significant, where the test statistic for this case tends to a standard normal distribution:

$$W_i = \frac{\hat{\theta}_i}{se(\hat{\theta}_i)} \sim N(0,1) \tag{15}$$

Therefore, if the p-value <5% H_0, is rejected, and it is concluded that the estimated parameter is significant for the model.

2.9. Estimation measures of the proposed model

- **AIC**

The Akaike information criterion is a measure of the goodness of fit of a statistical model. It could be said that it describes the relationship between bias and variance in the construction of the model, or speaking in a general way about the accuracy and complexity of the model.

It is necessary to specify that the AIC is not a test of the model in the sense of hypothesis testing. Rather, it provides a means for comparison between models with a model selection tool. Given a data set, several candidate models can be classified according to their AIC, with the model that has the minimum AIC being the best. From the AIC values it can also be inferred that, for example, the first two models are more or less tied and the rest are much worse. The AIC is defined as:

$$AIC = 2k - 2 \times Ln(L) \tag{16}$$

Where:

- k is the number of model parameters
- Ln(L) is the log-likelihood model for the statistical model.

- **BIC**

Bayesian Information Criterion (BIC) or Schwarz Criterion (SIC) is a measure of goodness of fit of a statistical model, and is usually used as a criterion for the selection of models among a finite set of models. This criterion is based on the logarithmic probability function (LLF) and is closely related to the Akaike information criterion (AIC).

Similar to the AIC, the BIC introduces a penalty term for the number of parameters in the model, but the penalty is greater than one in the AIC. The BIC is defined as:

$$BIC = k \times \ln(n) - 2 \times Ln(L) \tag{17}$$

where:

- k Similar to the AIC, the BIC introduces a penalty term for the number of parameters in the model, but the penalty is greater than one in the AIC. The BIC is defined as:
- Ln(L) is the log-likelihood model for the statistical model.

Given different estimated models, the model with the lowest BIC value is preferred; a low BIC implies fewer explanatory variables, better fit, or both.

2.10. Forecast error measures

When you have a certain model as a proposal, you usually make slight predictions and then compare these predictions with the real values. This will allow you to know if the model makes good or bad predictions. Statistics based on prediction errors allow us to know if the proposed model is good or not.

For the calculation of the error measurements, it is necessary to have the exact observations of the series (z_i) and the predicted values (x_i).

- **Average Absolute Error:**

  It measures the spread of the forecast error, or in other words, the measurement of the size of the error in units. It is the absolute value of the difference between the actual demand and the forecast, divided over the number of periods.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{18}$$

- **Mean Square Error:**

Like the MAD, the MSE is a measure of dispersion of the forecast error, however this measure maximizes the error by squaring, punishing those periods where the difference was higher compared to others. Consequently, the use of the MSE is recommended for periods with small deviations.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \tag{19}$$

- **Mean Absolute Percentage Error:**

The MAPE gives us the deviation in percentage terms and not in units like the previous measurements. It is the average of the absolute error or difference between the actual demand and the forecast, expressed as a percentage of the actual values.

$$MAPE = \frac{\frac{\sum_{i=1}^{n} |y_i - x_i|}{|y_i|}}{n} = \frac{\frac{\sum_{i=1}^{n} |e_i|}{|y_i|}}{n} \tag{20}$$

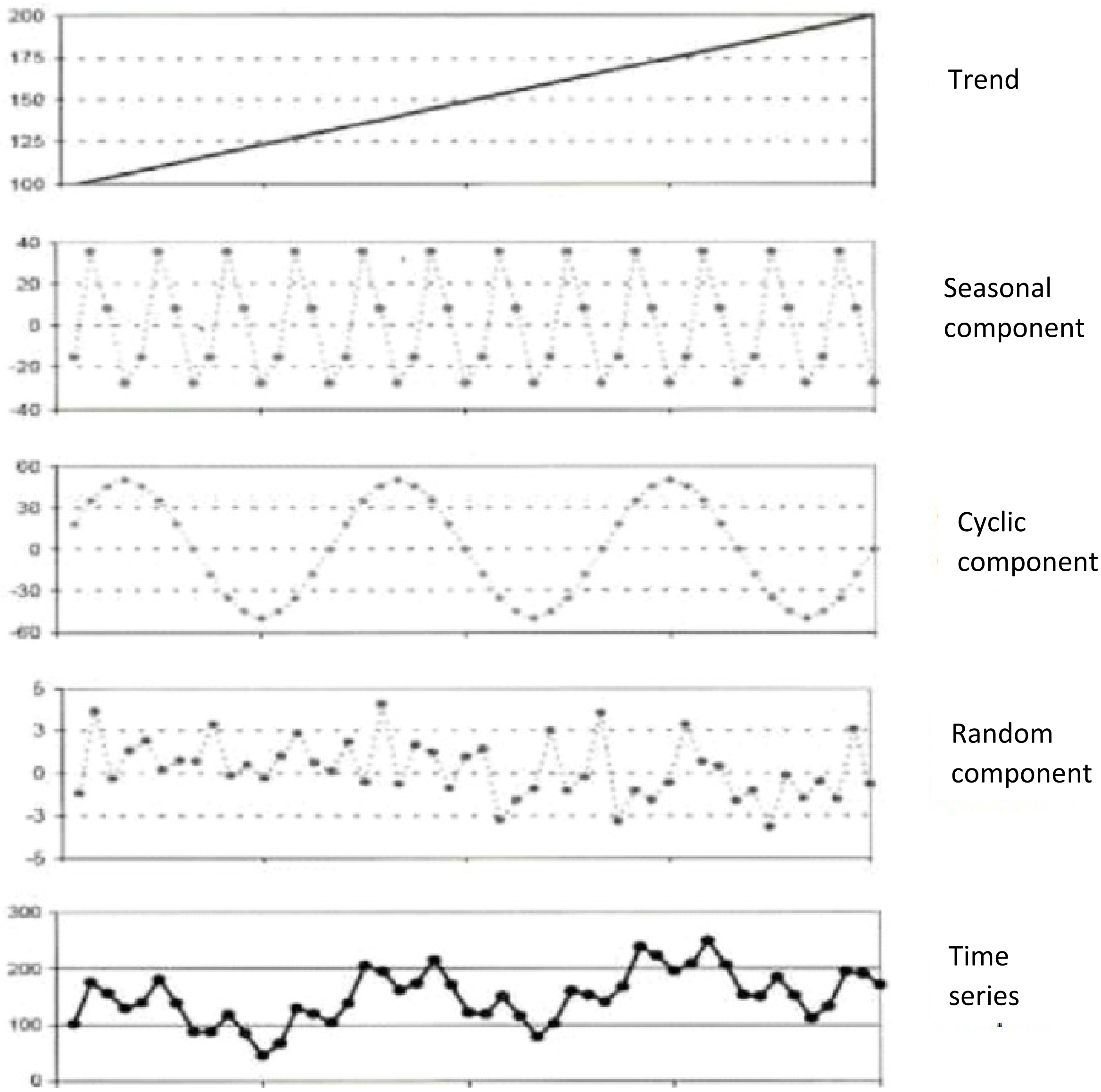

Figure 3.2: *Decomposition of the Components of a Time Series*

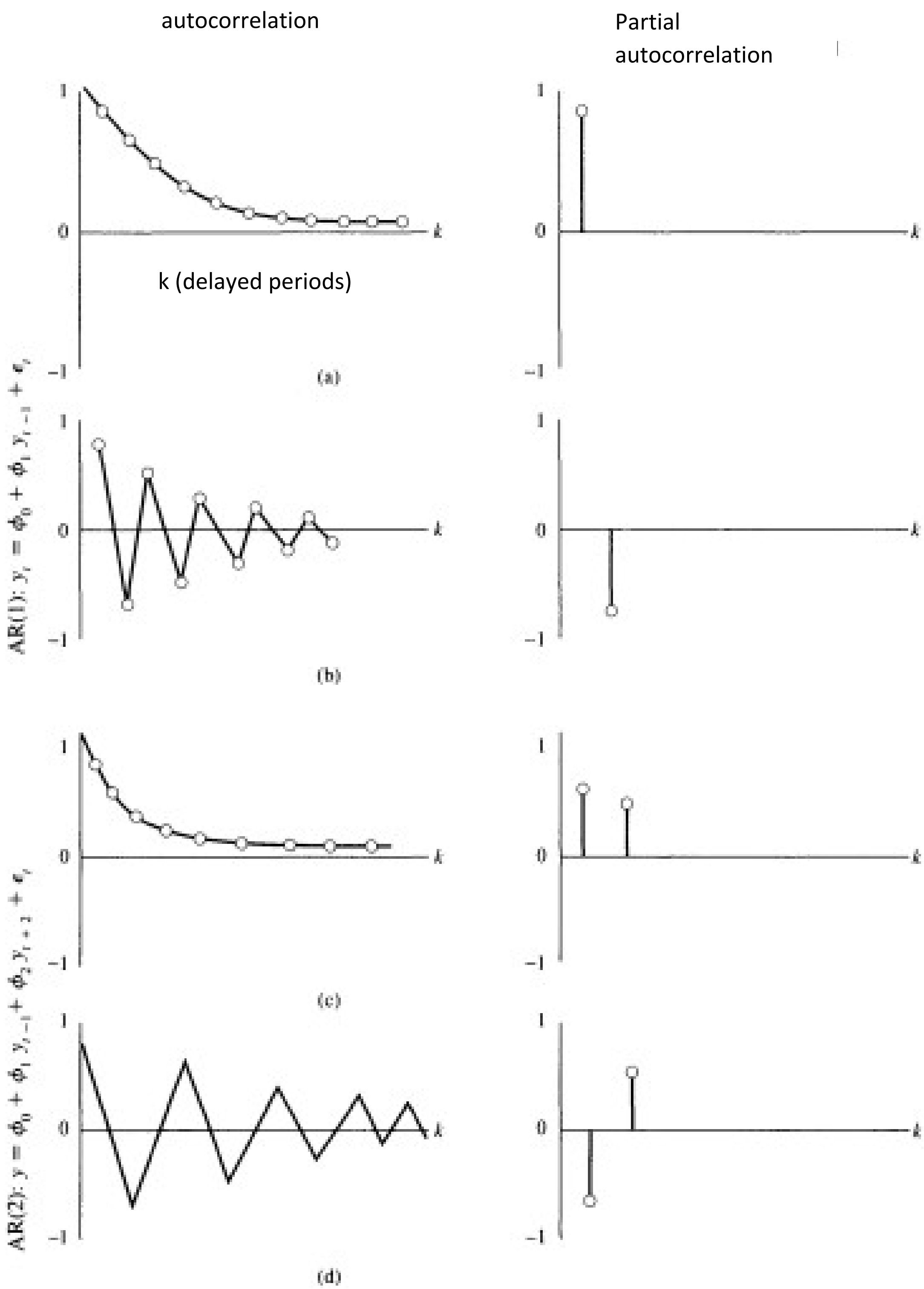

Figure 3.3: *Autocorrelation and partial autocorrelation coefficients of the AR(1) and AR(2). Hanke Reitsch. "Forecasts in business"*

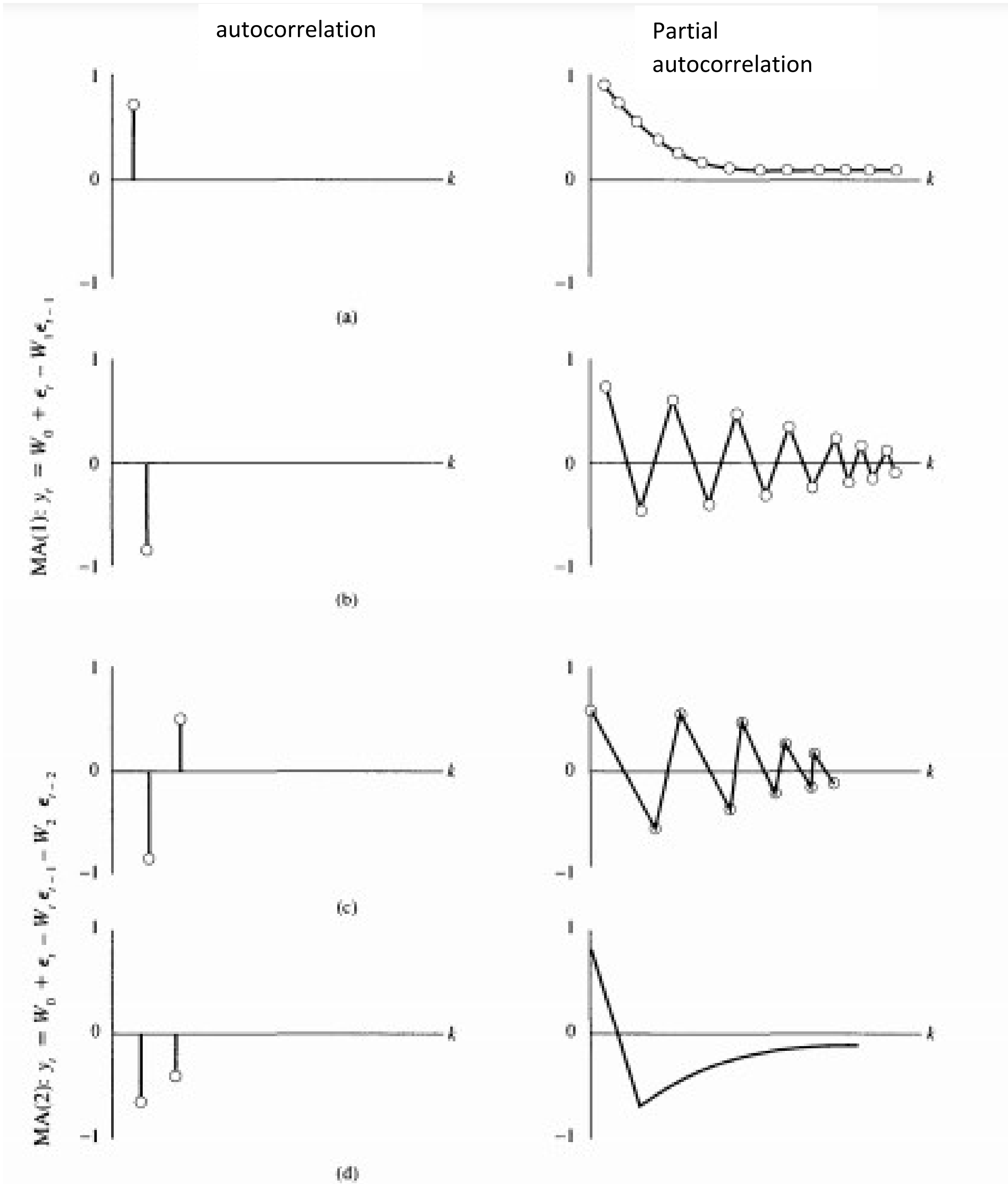

Figure 3.4: *Autocorrelation and partial autocorrelation coefficients of the MA(1) and MA(2) models. Hanke Reitsch. "Forecasts in business"*

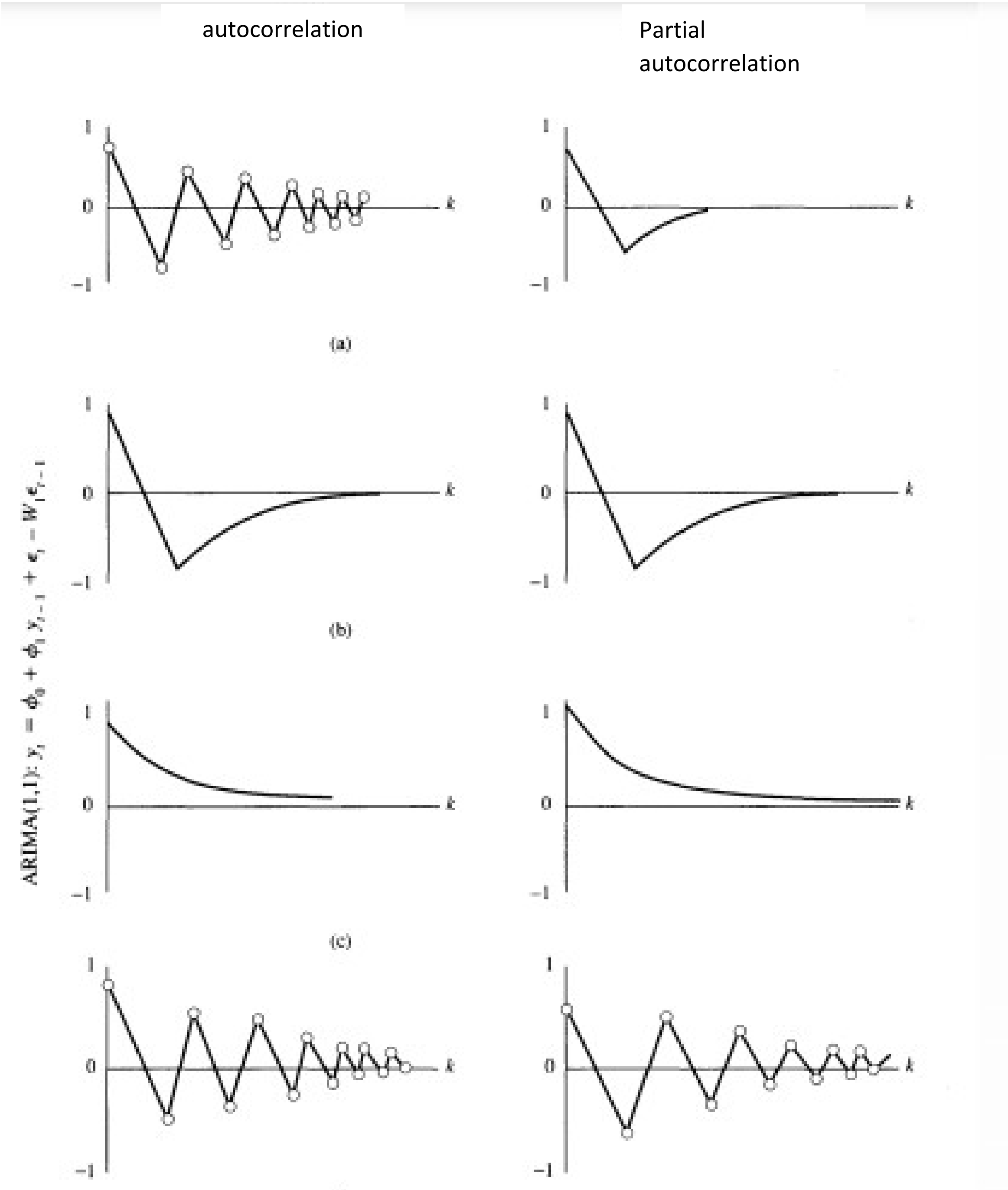

Figure 3.5: *Autocorrelation and Partial Autocorrelation Coefficients of the ARIMA (1,1) Mixed Model. Hanke Reitsch. "Forecasts in business"*

Chapter 3

Materials, Methods and Procedures

3.1. Study materials

3.1.1. Source of observations

The Consumer Price Index is the result of the calculation of the National Household Budget Survey. In this survey, people are asked about which products they usually buy. Under this scheme, the CPI calculation is made by comparing the monthly results of the equestrian against a base year (the base year in Peru was 2009).

The Central Reserve Bank of Peru publishes monthly the level of the Consumer Price Index (CPI) of Metropolitan Lima, this indicator is published on a monthly basis and represents the moving quarter of the CPI levels.

3.1.2. Target population

People who are over 14 years old with monthly income who acquire the basic basket.

3.1.3. Sample

Reports on the level of the Consumer Price Index of Metropolitan Lima comprised in the period January 2001 to March 2020. This series is divided into: January 2001 to December 2017 for the estimate and January 2017 to March 2020 for the validation of the forecast.

3.1.4. Kind of investigation

Non-experimental, Longitudinal, Trending

3.1.5. Analysis unit

Monthly records of the CPI level of Metropolitan Lima.

3.1.6. Study variable

- Variable: CPI
- Type: Quantitative
- Description: Economic index in which the prices of a predetermined set of goods and services determined on the basis of the continuous survey of family budgets, which a number of consumers purchase on a regular basis, and the variation with respect to the price are valued of each, with respect to a previous sample. It measures the changes in the price level of a basket of consumer goods and services purchased by households. This is a percentage that can be positive.

3.1.7. Sampling frame

The 236 monthly records of the CPI level of metropolitan lima from January 2001 to

3.1.8. Software used

The RStudio program was used, which was designed to perform statistical and graphical analysis. This software contains different native packages that allow the information processing to be more effective and efficient.

The packages used are the following:

readxl: This package allows data to be imported from Excel, since the obtained data was stored in the Excel software.

tidyverse: This package allows you to clean and transform the data you have in order to perform a better analysis.

tseries: This package is very useful to perform the Dickey-Fuller test for mean stationarity, using the function adf.test "show the" $p\tau$ if the calculated statistic is outside the critical values it sends a warning.

forecast: his package is very helpful for time series analysis as it provides functions that allow preliminary analysis of the time series as simple and partial autocorrelation graphs. In addition, it allows to build SARIMA models, adjust data and make forecasts.

ggfortify: This package allows you to make time series graphs in a more structured way. For example, it allows you to summarize three graphs in one.

3.2. Time series smoothing method

Smoothing is used very frequently in different time series works to make a quick visual examination of the properties of the data (for example, trend, seasonality, etc.), it is not enough just to observe the graph, to add missing values, and perform a quick forecast of the sample.

There are multiple smoothing methods, but not all of them allow a correct adjustment to the observations of the series. Therefore, it is important to know how to identify which method is the most appropriate for each series.

- **Double moving averages:**

  Unlike simple moving averages, which are used for stationary series, this method takes into account the trend component, whether it is positive or negative. This component is very noticeable in the series of the CPI Levels of Metropolitan Lima. How does it work? The method consists of calculating a set of moving averages and then a second set is calculated as the moving average of the first.

  To determine the moving average horizon, a 12th order horizon will be chosen, which will allow the average to move in ranges of one year.

- **Holt-Winter exponential smoothing:**

This method can not only take into account the trend component, but at the same time it allows to consider the seasonal component, this very beneficial since it was a limitation of the moving averages method, since the nature of the Consumer Price Index is not always increasing, since it also has periods.

How does it work? This smoothing is based on 3 coefficients: α, β and γ. The values that these indices contain will allow us to know what the behavior of the series is, and whether or not its components are outstanding.

3.3. Box-Jenkins methodology SARIMA model

Box and Jenkins presented clearly simple techniques for identifying the appropriate orders of the models, and for the numerical evaluation of the approximate maximum likelihood estimators of the model parameters, which has been very popular in recent years.

These models are widely used for the construction of ARIMA models, but ARIMA models are usually generalized by considering the effect of seasonality, incorporating it as an additional component. In that case, we speak of a SARIMA or SARIMA (p, d, q) (P, D, Q) s model, if you want to specify with their respective orders.

The methodology can be summarized in 4 stages: Identification, prediction, validation and estimation.

- Identification: Graphic analysis, determination if the series is stationary or not, etc.
- Estimation: Estimation of the coefficients of the SARIMA model.
- Validation: Analysis of residues.
- Forecast: Make predictions.

3.4. Process

The procedure to be carried out will be based on the Box-Jenkins methodology.

1. A previous analysis is carried out to identify the components of the time series, the time series graphs and the smoothing methodology are used as support.

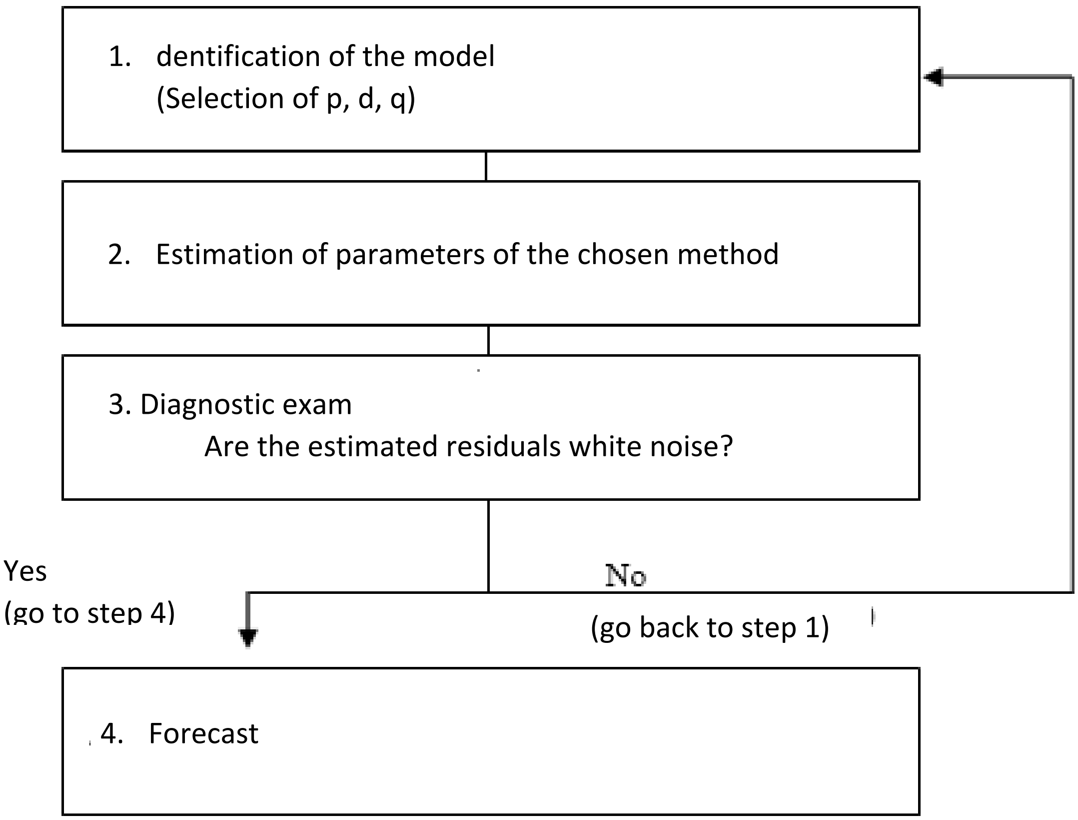

Figure 4.1: *Box-Jenkins Methodology*

2. Use of the estimation sample: This sample is taken forming a time series of fewer observations, to determine the candidate models with the best estimation. The next steps are carried out on this sample, that is, any transformation to the observations, such as differentiation, will be applied directly to the estimation data.

3. Graphical representation: A graphical view of the time series will be very useful to define the presence of stationarity, since this visibility provides a broad panorama and does not make the analysis of the time series cumbersome.

4. Checking for stationarity of the series: In case of non-stationarity in variance, a logarithmic transformation (Box - Cox) will be applied. This transformation allows the variance of the observations to be constant; and in case of visual identification of trend or also possible non-stationarity in the mean, the necessary differences will be applied, in the order and range that is convenient. Additionally, stationarity tests (Dickey - Fuller) will be carried out, which allow ensuring compliance with this necessary property of the time series.

5. Identification of the model: Once the series has been obtained in its stationary form, the model alternatives are determined, that is, examining the order of the autoregressive processes and moving averages of the regular and seasonal components, based on the autocorrelation functions (ACF) and partial autocorrelation (PACF), both in the regular and seasonal parts. Those lags that go beyond the established limit will provide a vision of what parameters the proposed models should have.

6. Estimation of the coefficients of the model: We proceed to the estimation of the parameters of the alternative models, which will form part of the mathematical model that will be proposed.

7. The significance test is applied for each estimated parameter (Wald), this allows a readjustment of the parameters for each model, that is, those coefficients that are significant for the model will be preserved. If u coefficient is not significant, it will be withdrawn.

8. Use of the forecast validation sample: This sample is taken to validate the quality of the forecasts of the candidate series. A graphical analysis will be carried out to corroborate the behavior of the real observations with the adjusted observations.

9. Forecast validation and model selection: The candidate models are compared through the forecast metrics (MAE, RMSE, MAPE), where the one that provides the best results is selected. The selected model will be the one that fits the total data.

10. Diagnosis of residuals: It consists of verifying that the residuals between the predicted and actual values meet the white noise conditions (Ljung Box): normality,

independence, and constant variance.

11. Final Forecast: The selected model will be used to forecast the observations in April, June and July 2020

Chapter 4

Presentation and Analysis of Results

4.1. Preliminary analysis

The inflation index in Metropolitan Lima has maintained a constantly slight growth in most of its periods during the last two decades, although it has also shown periods of considerable changes, caused by various economic, social and conjunctural, external and internal factors.

Although the new pandemic produced by Covid-19 is one of the recent events that has been affecting the economy in our country considerably, there are other events over time that have also been making a difference in its behavior.

In context with the aforementioned, the analysis of the factors that have contributed to these changes could be very varied, in fact, carrying out a previous historical diagnosis, many interesting aspects were found, however, taking into account each one of These factors could lead to the study becoming tedious and taking longer than expected. Therefore, it has been proposed to limit the horizon of the preliminary analysis to the period of the last 20 years (2001-2020).

The aim is to work with a considerable amount of data in order to mitigate the error produced by different changes that have occurred over the years.

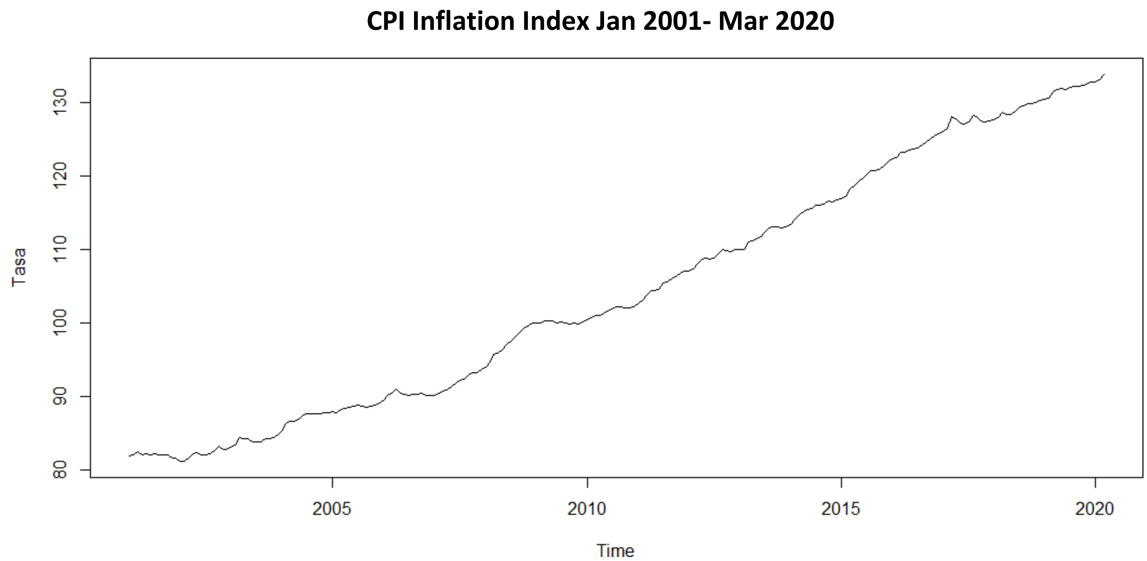

Figure 5.1: *Graph of the CPI series, Jan 2001- Mar 2020*

4.1.1. Trend

In Figure 5.1 we can see that the Consumer Price Index has been having a positive trend since the last years, which means that the price of the basic basket is increasing. This increase means that people have been paying more for the goods and services they purchase; But that is not necessarily bad, since as we see it is a slight growth, which indicates people could be paying more for services because now they are making more money, that is, their demand for products has increased.

Despite the slight growth that the CPI has had, it can be seen that there is a period where the growth slope is higher than in the other periods. This steeply high slope occurred during 2008 and 2009, as a consequence of the international economic crisis at that time, which caused Peru's GDP to fall by 9.8%. In this period, external demand suffered a significant drop, later bringing a decrease in industrial production. In addition, the political and military problems of the US with several oil-producing countries pushed up fuel prices, this increase has generated an increase in the prices of agricultural biofuels such as corn and soybeans, which are imported products in Peru. for the raising of chickens as the production of oil, consequently it encourages the rise of the final products. The consumer group of the CPI that increased the most at that time was food, with a level of 10.29%.

4.1.2. Seasonal

Given the nature of the CPI, in order to observe the effects of seasonality, it is necessary to apply at least one differentiation. In this case, two differentiations were applied: monthly and interannual.

In the interannual differentiation, it is observed that if there is seasonality where it indicates that the CPI grows little in the first quarter of each year, but in the third quarter of the year there is a jump in the level of this indicator. We also see an increase during the years of 2008 and 2009 for the reasons already mentioned above.

Using the monthly differentiation, we see the effects of seasonality every two months, this is greatly affected by the fact that the components of the basic basket respond to different seasons, for example, in transport there is a greater seasonality in the summer season given that the Most people go on vacation and it is Christmas holidays causing people to seek travel.

The food sector has multiple seasons because although some products increase their cost, others decrease it. The real estate component is not usually stationary because there is no station that indicates what date if you pay for the house or what date not, since the cost is normally constant.

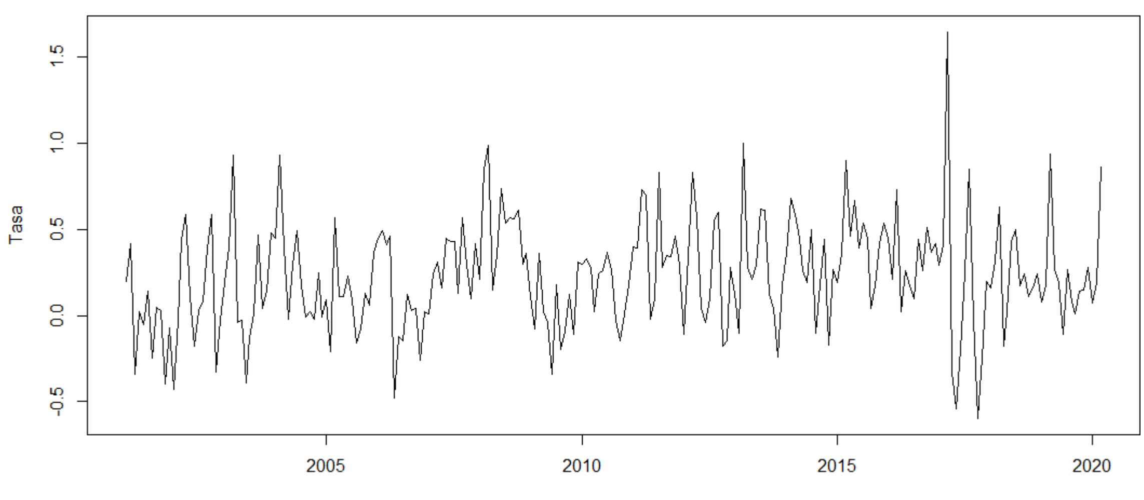

Figure 5.2: *Monthly differentiation of the CPI Level of Metropolitan Lima*

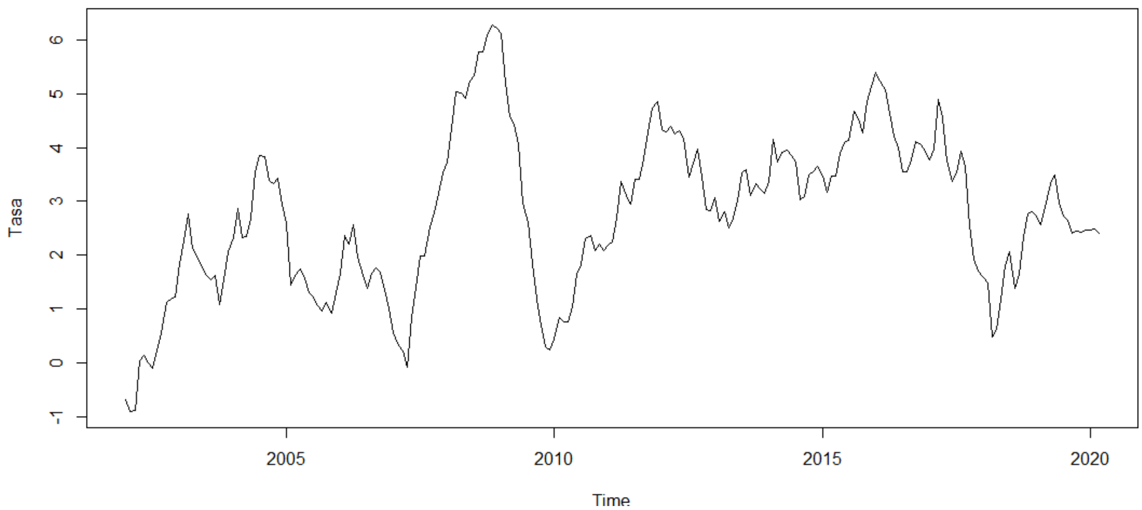

***Figure 5.3**: Interannual differentiation of the CPI Level of Metropolitan Lima*

4.2. Smoothing

In order to explore the behavior of the series at large traits, it was decided to apply smoothing techniques and analyze which could describe the series with better precision. Given that the smoothing methods do not take into account correlation assumptions between the observations, unlike ARIMA, the stationarity property is not relevant in these cases, therefore, the original series was taken to evaluate the method. that best fits the data of the CPI levels. Furthermore, the previous analysis of the components of the time series for CPI will guide the selection of the most suitable smoothing method.

There are different smoothing methods, which allow you to adjust results that predict the values of the time series. Among the methods that are known are: smoothing by moving averages and exponential smoothing. The first allows the series to be adjusted over a period of time using the average of the nth order chosen. If the series responds to the trend component, the smoothing of double moving averages is the most appropriate. The second method is also based on the average of the values of the series, but also uses an autocorrelation mechanism that seeks to adjust the forecasts in the opposite direction to the deviations of the past, that is, it corrects the predictions over the course of the series.

4.2.1. Moving averages

Taking into account the previous analysis, it is known that there is a trend in the series, so when applying a smoothing by moving averages it is necessary to perform a smoothing of double moving averages. In this sense, the calculation will be carried out taking the size equal to 12 as the order of the moving average, that is, the forecasts will be the 12-month averages.

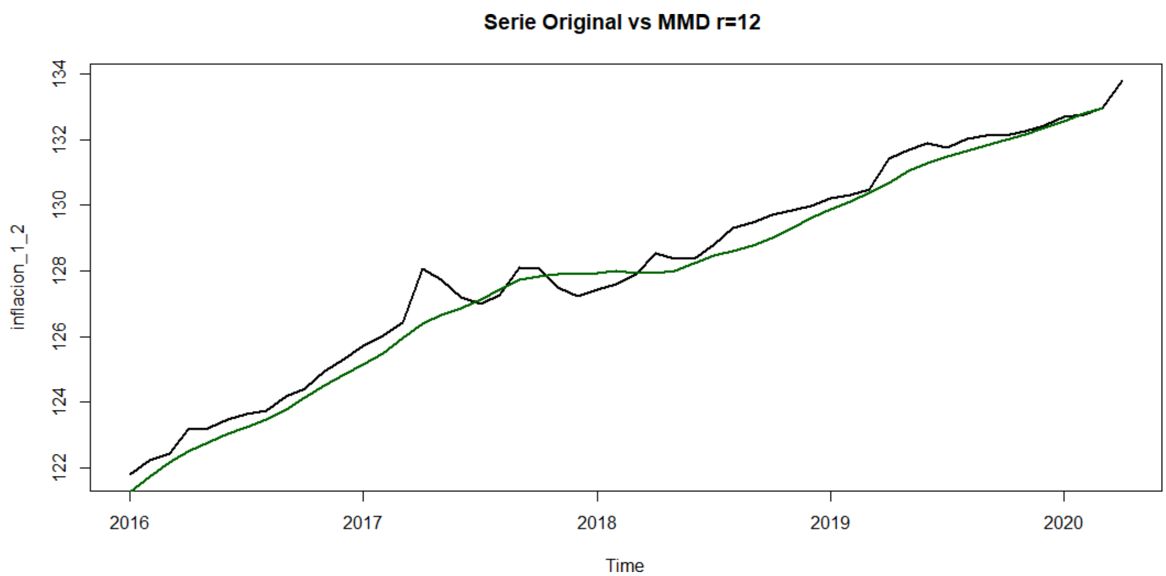

Figure 5.4: *Moving averages of order 12 of the Monthly CPI Level of metropolitan lima*

In Figure 5.4 it can be seen how the 12th order smoothing values, which are represented by the green line, adjust to the time series. As a result of this smoothing, the average value adjustment for the last period is 132.89014 and the trend level is 0.0926010. This indicates that the trend of the series is constantly increasing and thus future values will also be increasing, but the smoothing of double moving averages does not consider the seasonal component, therefore, the generated predictions would not consider this component either. If we used the smoothing of double moving means to obtain the future values, it would generate errors in the prediction.

4.2.2. Holt-Winters exponential smoothing method:

The exponential smoothing method of three components or also called the Holt-Winter method, allows taking into account the trend components, seasonality and randomness. As already mentioned, these components are included within the observations of the Consumer Price Index. It should be noted that the mean value, trend and season of the forecasts respond to the parameters α, β and γ respectively, which are typical of the Holt-Winter method.

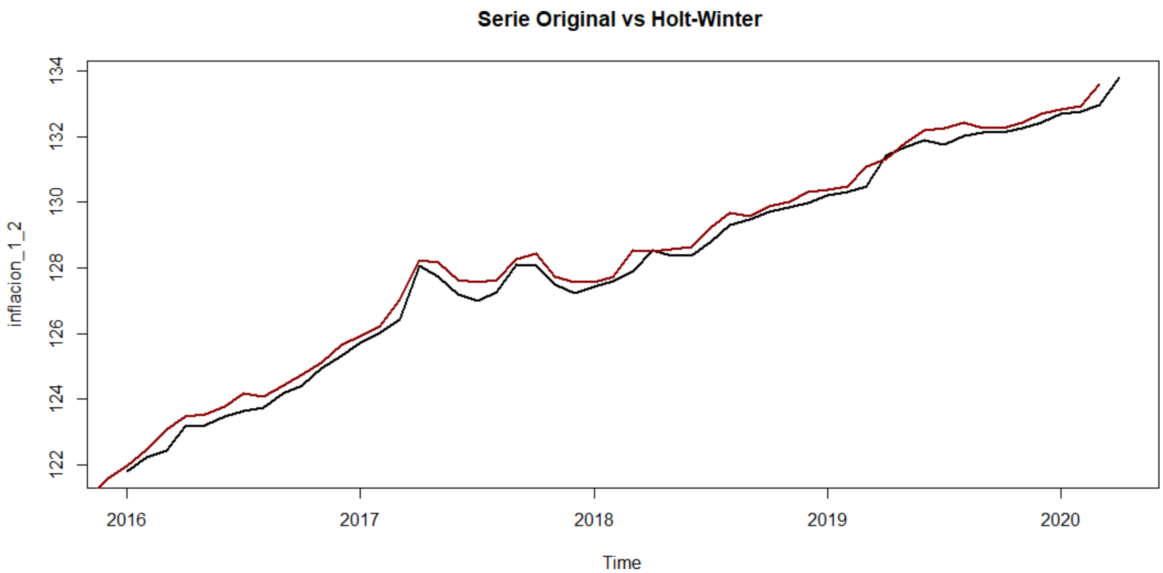

***Figure 5.5**: Holt-Winter*

In Figure 5.5 it can be seen how the exponential smoothing values, represented in the red line, are adjusted to the time series. The parameters of α, β and γ have resulting values of 0.868, 0.0455 and 1 respectively. The value of α (0.868) is relatively high, which indicates that the estimates at the present time are based on recent observations, that is, there is a marked autocorrelation. The value of β (1) indicates that the estimate of the slope of the trend component is updated over time. The high value of γ (0.912) indicates that the seasonal component is based only on very recent observations. As a result, the forecasts under this method for the periods of April, May and June are respectively. 133.7229, 133.845 and 134.069.

Table 1 will show the estimates of Absolute Mean Error (MAE), Mean Square Error (RMSE) and the Percentage Absolute Mean Error (MAPE), for each of the smoothing methods.

	MAE	RMSE	MAPE
Moving averages	0.65	0.68	0.80
HoltWinter	0.25	0.26	0.32

***Table 1:** Errors of MAE, RMSE and MAPE of the smoothing methods for the CPI levels of Metropolitan Lima*

The exponential smoothing also responds to a better fit for the observations of the series, since it shows lower levels of error than those obtained by the smoothing of moving means. Therefore, the best adjustment and forecast obtained for the level of inflation in Metropolitan Lima is obtained using the Holt-Winter smoothing method.

Using the results of the preliminary analysis, the procedures and results will be detailed as a consequence of the different procedures for the adjustment to the time series at the CPI level in Metropolitan Lima.

As indicated in the Box-Jenkins methodology, the first 18 years will be used to propose a model that fits the data, and the last year will be used to validate the model.

4.3. Adjustment of the SARIMA model for the CPI level in Metropolitan Lima:

The SARIMA type model is the most suitable for the time series of the CPI level of metropolitan Lima, as has already been seen in the preliminary analysis.

After carrying out the preliminary and respective smoothing analyzes, the observations are divided into two groups as previously mentioned, where the first group will be the training observations and the second group the validation observations.

4.3.1. Checking for stationarity

Working with the training observations, for the application of the time series models it is necessary to look for the series to be stationary. According to Figure 5.6, a Box-Cox transformation will not be necessary for the variance, since it appears to be constant; This does not happen with the mean, since as observed there is a trend, therefore there is no stationarity in the mean. Therefore, monthly and interannual differentiations will be applied to seek to convert the Metropolitan Lima CPI series into a stationary series.

1. To d=0 y D=1 (∇_t) the contrast is:

$$H_0 : \gamma_1 = 0 \qquad vs \qquad H_1 : \gamma_1 < 0$$

Where γ1 is the coefficient associated with the AR (1) component assumed for the series in this contrast. The p-value for the test statistic obtained is 0.012 <0.05, thereforeH0 can be rejected and it is concluded that the series ∇Z_t does not have a unit root and is stationary.

The series with one differentiation is already stationary, it will be tested whether with two differentiations it is still stationary.

2. To d=1 y D=1 ($\nabla\nabla_{12}Z_t$) the contrast is:

$$H_0 : \gamma_1 = 0 \qquad vs \qquad H_1 : \gamma_1 < 0$$

Where γ1 is the coefficient associated with the AR (1) component assumed for the series in this contrast. The p-value for the test statistic obtained is 0.01 <0.05, therefore H0 is rejected and it is concluded that the series $\nabla\nabla_{12}Z_t$ does not have a unit root and is stationary.

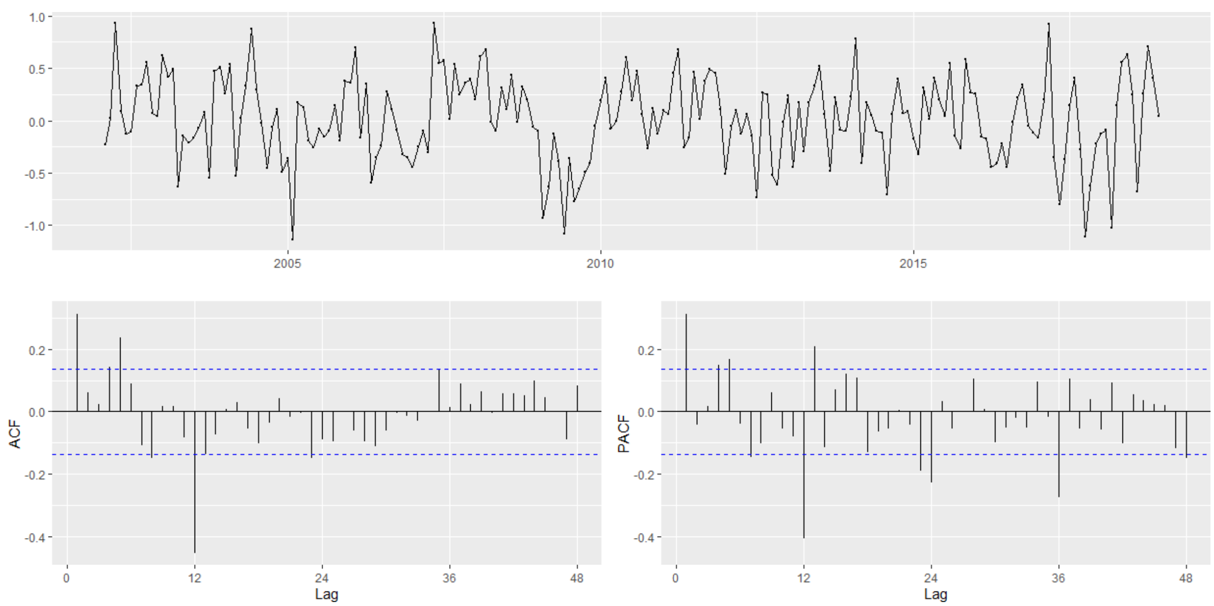

Figure 5.6: *Time graph and graphs of ACF and PACF for the data with 1 non-seasonal and 1 seasonal difference of the CPI level in Metropolitan Lima.*

4.3.2. Construction of the models a first candidate model:

According to the stationarity check, we can observe that by applying a monthly and interannual difference (d = 1 and D = 1) the series becomes stationary.

If we base ourselves on the ACP and PACP graphs, it is observed how several lags are going beyond the limits of the graph, these lags that come out will allow defining the parameter for the components of the model, in this sense, according to the PACP graph (Partial Autocorrelation), the parameters of the AR (and SAR components would be p = 3 and P = 1. Using the same logic in the ACP Autocorrelation graph) the parameters of the MA and SMA component would be q = 1 and Q = 1.

According to the graph, different proposed mathematical models have been defined

- SARIMA(3, 2, 1)(1, 1, 1)$_{12}$:

$$(1-\sum_{k=1}^{3}\phi_k B^k)(1-\sum_{k=1}^{1}\Phi_k^{12}B^k)(1-B^{12})^1(1-B)^2 z_t = (1-\sum_{k=1}^{1}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t$$

$$donde : \varepsilon_t \sim N(0,\sigma^2)$$

After defining the model that has been chosen as a candidate, we will proceed with the estimation of parameters, and additionally the contrast of the Wald Test, of equation 13, will be used to test the significance of the coefficients (estimated parameters) using a level 5% significance. This will allow removing parameters that are not added to the model and simplifying its structure.

Model	Estimated coefficients	Standard error	P- value
SARIMA (3, 2, 1)(1, 1, 1)$_{12}$	$\hat{\varphi}_1 = 0,311$	0.076	0.000 (*)
	$\hat{\varphi}_2 = -0,064$	0.0757	0.049 (.)
	$\hat{\varphi}_3 = 0,0228$	0.074	0.759 (**)
	$\hat{\theta}_1 = -0,968$	0.0407	0.000 (***)
	$\hat{\Phi}_1 = -0,080$	0.084	0.339 (**)
	$\hat{\Theta}_1 = -0,8712$	0.0743	0.000 (*)

Table 2: Estimation of parameters for the candidate models of the series of the CPI level of Metropolitan Lima. Significant parameter: (*), Non-significant parameter: (**)

From these results in table 2 we can say that the proposed models can be reduced in their dimensions, these reductions are defined in the following list:

- SARIMA(2, 2, 0)(1, 1, 1)$_{12}$

4.3.3. Checking the fit of the estimated model:

The following table shows the model with its respective AIC, BIC y $\hat{\sigma}_e^2$

Estimated Models	AIC	BIC	$\hat{\sigma}_e^2$
SARIMA(2, 2, 0)(1, 1, 1)$_{12}$	90.8	100.41	0.106

Table 3: AIC *levels* for the SARIMA models of the CPI level in Metropolitan Lima

4.3.4. Candidate model forecast:

The next step is to evaluate the forecast error, quantified through forecast metrics that measure the difference between the actual values of the series and the predicted value, among the most used are: RMSE, MAE and MAPE. These metrics will compare the predicted rates with the true data of the validation or test sample, that is, from January 2019 to March 2020.

The model's forecasts for the next few months that make up the test sample are shown in Figure 5.7.

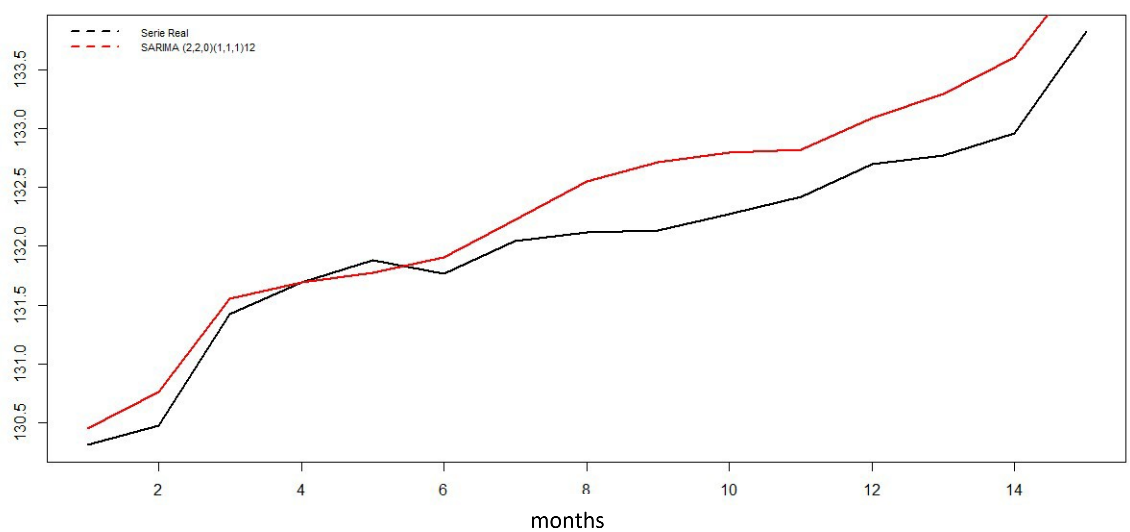

Figure 5.7: Time graph and graphs of the CPI level in Metropolitan Lima, January 2019 - March 2020.

Model	MAE	RMSE	MAPE
$SARIMA(2, 2, 0)(1, 1, 1)_{12}$	0.335	0.391	0.253

Table 4: *Levels of Errors for the predictions of the SARIMA models of the CPI level in Metropolitan Lima*

The SARIMA (2, 2, 0) (1, 1, 1) 12 model is chosen as the best model in comparison to the other two models made since the three measurements of the prediction error evaluated for each model, this model mentioned has the lowest values.

Therefore, to express the model mathematically, the parameters of the model will be defined as follows:

4.3.5. Mathematical form of the model: SARIMA(2,2, 0)(1, 1, 1)$_{12}$:

$$(1-\sum_{k=1}^{2}\phi_k B^k)(1-\sum_{k=1}^{1}\Phi_k^{12}B^k)(1-B^{12})^1(1-B)^2 z_t = (1-\sum_{k=1}^{1}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t \tag{21}$$

4.3.6. Waste diagnosis:

- **Graphic inspection:** Looking at Figure 5.5, the residuals resemble a normal distribution, the constant of the mean is constant and equal to zero, and the graph of the correlogram indicates that the data are stationary. All peaks are now within significance limits, so the residuals appear to be white noise.

- **Ljung Box Test:** Another method to check the white noise of the residuals of the applied model is the Ljung-Box test, which needs to contrast the null hypothesis, H_0: $\rho_1 = \rho_2 = \rho_3 = \ldots = \rho_m = 0$, for this it was obtained a p-value equal to 0.2835 that is greater than 0.05, so that H0 is not rejected, it also shows that the residuals do not have remaining autocorrelations, therefore, it is concluded that statistically the residuals have a white noise behavior.

4.3.7. Forecast

Now we have a SARIMA model that passes the required controls and is ready for forecasting. The model's forecasts for the next three months shown in Figure 5.8 were obtained from the SARIMA (2, 2, 0)(1, 1, 1)$_{12}$ model estimated at 100% of the available observations. The large and rapidly growing prediction intervals show that the CPI level of Metropolitan Lima could start to increase or decrease at any time, while the forecast of the point is trending downward, the prediction intervals allow the data to have a decreasing trend during the forecast period.

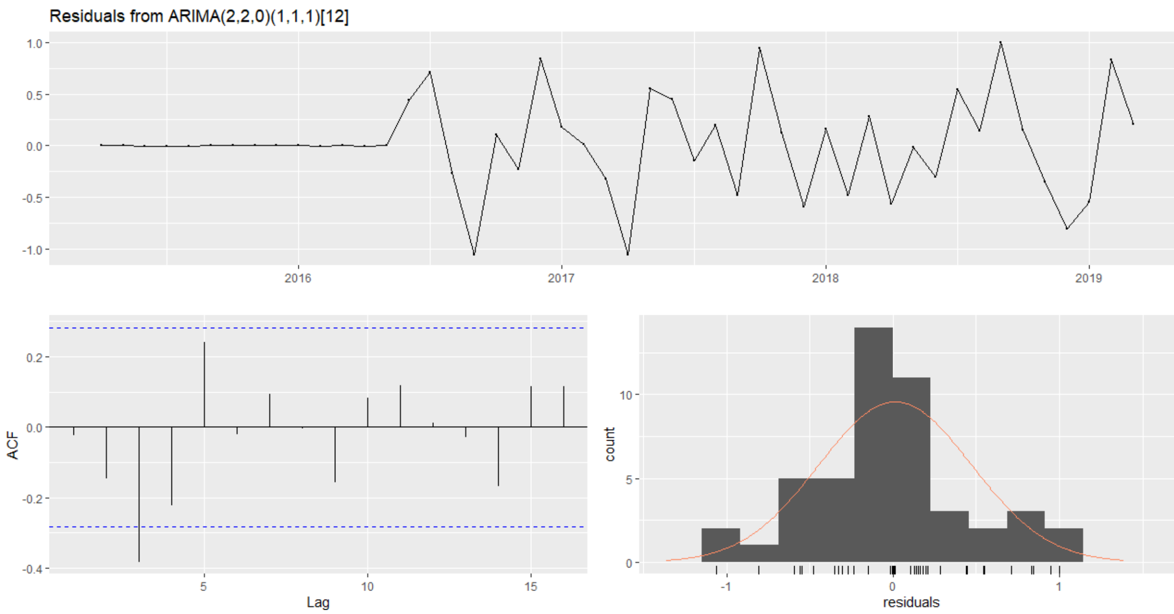

Figure 5.8: *Residues of the SARIMA model (2, 2, 0)(1, 1, 1)12 adjusted for the data of the CPI level in Metropolitan Lima.*

Model	Forecast
April 2020	133.9442
May 2020	134.0025
June 2020	134.1467

Table 5: *Estimated CPI level for April, May and June in Metropolitan Lima proposed by the SARIMA model (2, 2, 0)(1, 1, 1)$_{12}$*

The forecasts shown in Table 5 are those obtained by the first adjusted SARIMA model (2, 2, 0)(1, 1, 1)$_{12}$.

4.3.8. Construction of the candidate models:

By using these parameters, we can propose different SARIMA models that allow adjusting the series of the CPI level of Metropolitan Lima.

4.3.9. Mathematical definition of the candidate models:

According to the graph, different proposed mathematical models have been defined

- SARIMA(1, 1, 2)(1, 1, 1)$_{12}$:

$$(1-\sum_{k=1}^{1}\phi_k B^k)(1-\sum_{k=1}^{1}\Phi_k^{12}B^k)(1-B^{12})^1(1-B)^1 z_t = (1-\sum_{k=1}^{2}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t$$

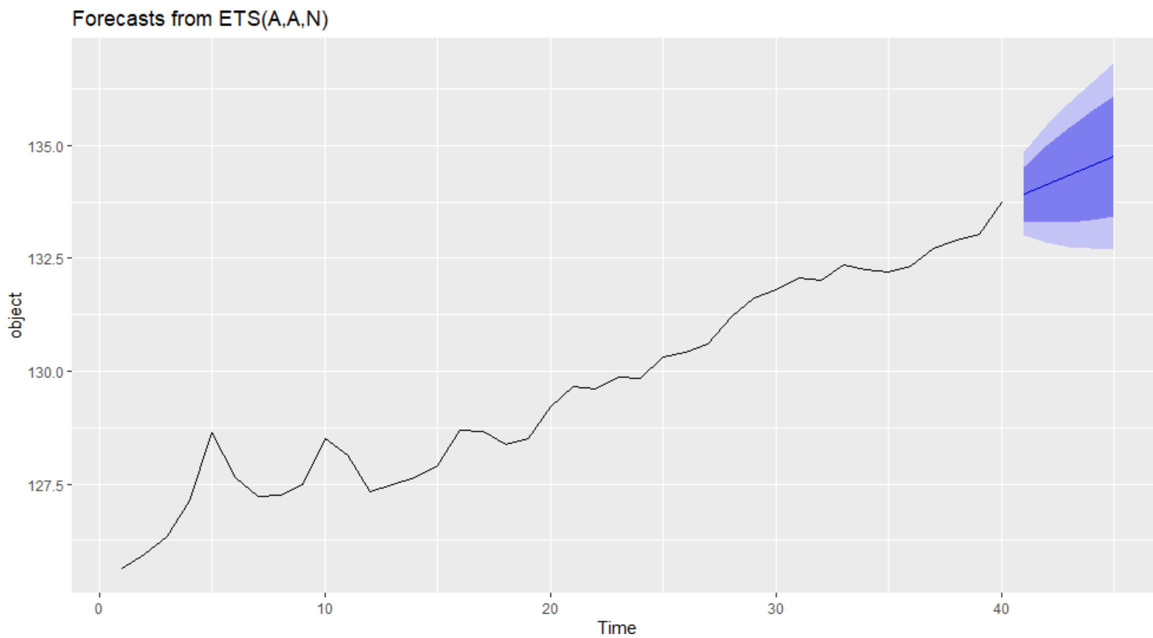

Figure 5.9: Forecast of the CPI level data in Metropolitan Lima using the SARIMA (2, 2, 0)(1, 1, 1)$_{12}$, model, for the months of January-June 2020 (The image shows a zoom starting since January 2018)

- SARIMA(2, 1, 1)(2, 1, 1)$_{12}$

$$(1-\sum_{k=1}^{2}\phi_k B^k)(1-\sum_{k=1}^{2}\Phi_k^{12}B^k)(1-B^{12})^1(1-B)^1 z_t = (1-\sum_{k=1}^{1}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t$$

- SARIMA(2, 1, 0)(2, 1, 2)$_{12}$:

$$(1-\sum_{k=1}^{2}\phi_k B^k)(1-\sum_{k=1}^{2}\Phi_k^{12}B^k)(1-B^{12})^1(1-B)^1 z_t = (1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t$$

- SARIMA(3, 2, 1)(1, 1, 1)$_{12}$:

$$(1-\sum_{k=1}^{3}\phi_k B^k)(1-\sum_{k=1}^{1}\Phi_k^{12}B^k)(1-B^{12})^1(1-B)^2 z_t = (1-\sum_{k=1}^{2}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t$$

- SARIMA(1, 1, 1)(0, 1, 1)$_{12}$:

$$(1-\sum_{k=1}^{1}\phi_k B^k)(1-B^{12})^1(1-B)^2 z_t = (1-\sum_{k=1}^{1}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12}B^k)\varepsilon_t$$

$$donde : \varepsilon_t \sim N(0,\sigma^2)$$

After defining the models that have been chosen as candidates, we will proceed with the estimation of parameters for each of these candidate models, and additionally we will use the contrast of the Wald Test, of equation 15, to test the significance of the coefficients (estimated parameters) using a significance level of 5%. This will allow removing parameters that are not added to the model and simplifying its structure.

Model	Estimated coefficients	Standard error	P-value
SARIMA(1, 1, 2)(1, 1, 1)$_{12}$	$\hat{\varphi}_1 = 0{,}875$	0.106	0.000(*)
	$\hat{\theta}_1 = -0{,}518$	0.129	0.000(*)
	$\hat{\theta}_2 = -0{,}222$	0.085	0.009(*)
	$\hat{\Phi}_1 - 0{,}984$	0.115	0.000(*)
	$\hat{\Theta}_1 - 0{,}843$	0.169	0.000(*)
SARIMA(2, 1, 1)(2, 1, 1)$_{12}$	$\hat{\varphi}_1 = -0{,}596$	0.105	0.000 (*)
	$\hat{\varphi}_2 = 0{,}243$	0.079	0.00214(*)
	$\hat{\theta}_1 = 0{,}930$	0.077	0.000(*)
	$\hat{\Phi}_1 = -0{,}145$	0.096	0.132(**)
	$\hat{\Phi}_2 = -0{,}144$	0.093	0.123(**)
	$\hat{\Theta}_1 - 0{,}77$	0.0819	0.000(*)

Table 6: *Estimation of parameters for the candidate models of the series of the CPI level of Metropolitan Lima. Significant parameter: (*), Non-significant parameter: (**)*

Model	**Estimated coefficients**	**Standard error**	**P-value**
$SARIMA(2, 1, 0)(2, 1, 2)_{12}$	$\hat{\varphi}_1 = 0{,}918$	0.179	0.000 (***)
	$\hat{\varphi}_1 = 0{,}355$	0.0754	0.0000 (*)
	$\hat{\varphi}_2 = -0{,}042$	0.0724	0.553 (*)
	$\hat{\Phi}_1 = -1{,}028$	0.069	0.0000
	$\hat{\Phi}_2 = -0{,}047$	0.0773	0.540
	$\hat{\Theta}_1 = 0{,}117$	0.0534	0.0282
	$\hat{\Theta}_2 = -0{,}810$	0.0646	0.000
$SARIMA(3, 2, 1)(1, 1, 1)_{12}$	$\hat{\varphi}_1 = 0{,}311$	0.076	0.000 (*)
	$\hat{\varphi}_2 = -0{,}064$	0.0757	0.392 (**)
	$\hat{\varphi}_3 = 0{,}0228$	0.074	0.759 (**)
	$\hat{\theta}_1 = -0{,}968$	0.0407	0.000 (*)
	$\hat{\Phi}_1 = -0{,}080$	0.084	0.339 (**)
	$\hat{\Theta}_1 = -0{,}8712$	0.0743	0.000 (*)
$SARIMA(1, 1, 1)(0, 1, 1)_{12}$	$\hat{\varphi}_1 = 0{,}236$	0.210	0.261 (**)
	$\hat{\theta}_1 = 0{,}104$	0.216	0.010 (*)
	$\hat{\Theta}_1 = -0{,}875$	0.059	0.000 (*)

Table 7: *Estimation of parameters for the candidate models of the series of the CPI level of Metropolitan Lima. Significant parameter: (*), Non-significant parameter: (**)*

From table 6 and table 7 it can be said that the proposed models can be reduced in their coefficients, these reductions are defined in the following list:

- $SARIMA(1, 1, 2)(1, 1, 1)_{12}$
- $SARIMA(2, 1, 1)(0, 1, 1)_{12}$
- $SARIMA(1, 1, 0)(1, 1, 2)_{12}$
- $SARIMA(1, 2, 0)(1, 1, 1)_{12}$
- $SARIMA(0, 1, 1)(0, 1, 1)_{12}$

4.3.10. Comparison of the estimated models:

The table shows the models with their respective adjustment statistics.

Estimated models	AIC	BIC	$\hat{\sigma}_e^2$
$SARIMA(1, 1, 2)(1, 1, 1)_{12}$	79.89	96.45	0.077
$SARIMA(2, 1, 1)(0, 1, 1)_{12}$	81.42	97.98	0.077
$SARIMA(1, 1, 0)(1, 1, 2)_{12}$	81.8	98.37	0.078
$SARIMA(1, 2, 0)(1, 1, 1)_{12}$	90.8	100.41	0.106
$SARIMA(0, 1, 1)(0, 1, 1)_{12}$	79.98	89.92	0.078

Table 8: *AIC levels for the SARIMA models of the CPI level in Metropolitan Lima*

4.3.11. Selection of the best model:

The best model should be the one that has the lowest prediction error and best fits the observations. This prediction error is defined as the difference between the real value of the series and the prediction value. The metrics will be used to measure the prediction error: MAE, RMSE and MAPE. For the validation of the different estimated SARIMA models, a test data will be used (CPI level in Metropolitan Lima from January 2019 to March 2020).

The model's forecasts for the months between January 2019 and March 2020 are shown in Figure 5.10.

Male unemployment rate in metropolitan Lima, January 2019- March 2020

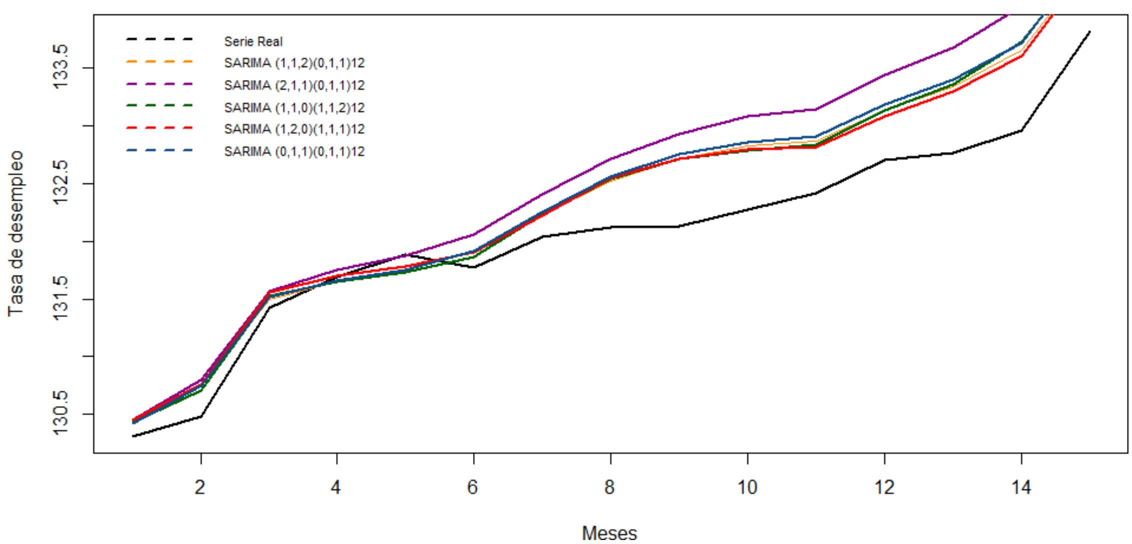

Figure 5.10: *Time graph and graphs of the CPI level in Metropolitan Lima, January 2019- March 2020.*

Model	MAE	RMSE	MAPE
SARIMA(1, 1, 2)(1, 1, 1)$_{12}$	0.348	0.408	0.408
SARIMA(2, 1, 1)(0, 1, 1)$_{12}$	0.529	0.633	0.399
SARIMA(1, 1, 0)(1, 1, 2)$_{12}$	0.355	0.4211	0.268
SARIMA(1, 2, 0)(1, 1, 1)$_{12}$	0.335	0.391	0.253
SARIMA(0, 1, 1)(0, 1, 1)$_{12}$	0.378	0.447	0.285

Table 9*: Levels of Errors for the predictions of the SARIMA models of the CPI level in Metropolitan Lima*

The SARIMA (0, 1, 1)(0, 1, 1)$_{12}$ model is chosen as the best model in comparison with the other models made since the three metrics for measuring the prediction error show good results. One could think of choosing another model with a lower result in errors, but the SARIMA (0, 1, 1)(0, 1, 1)$_{12}$ model also has a good fit in the data as shown in table 9.

Therefore, to express the model mathematically, the parameters of the model will be defined as follows:

Mathematical form of the model SARIMA(0, 1, 1)(0, 1, 1)$_{12}$:

$$(1-B^{12})^1(1-B)^2 z_t = (1-\sum_{k=1}^{1}\theta_k B^k)(1-\sum_{k=1}^{1}\Theta_k^{12} B^k)\varepsilon_t \tag{12}$$

4.3.12. Waste diagnosis:

- **Graphic inspection:** Looking at Figure 5.11, the residuals resemble a normal distribution, the mean constant is constant and equal to zero, and the correlogram plot gives indications that the data are stationary. All peaks are now within significance limits, so the residuals appear to be white noise.

- **Ljung Box test:** Another method to check the white noise of the residuals of the applied model is the Ljung-Box test, which needs to contrast the null hypothesis, $H_0 : \rho_1 = \rho_2 = \rho_3 = ... = \rho_m = 0$, for this it was obtained a p-value equal to 0.2835 that is greater than 0.05, so that H0 is not rejected, it also shows that the residuals do not have remaining autocorrelations, therefore, it is concluded that statistically the residuals have a white noise behavior.

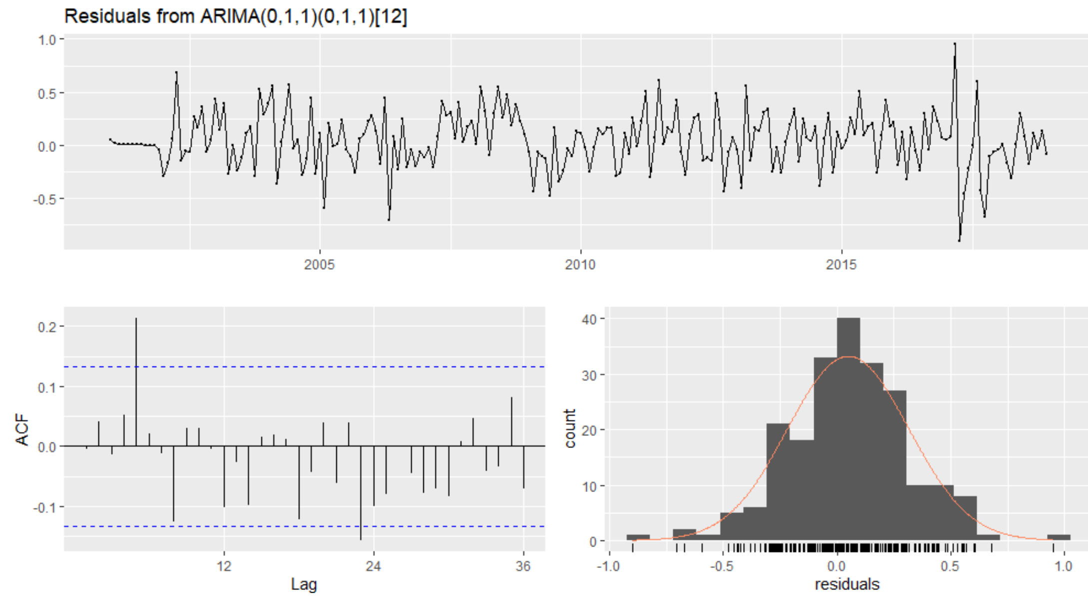

Figure 5.11: *Residues of the SARIMA model (0, 1, 1)(0, 1, 1)$_{12}$ adjusted for the data of the CPI level in Metropolitan Lima.*

4.3.13. Forecast

Now we have a SARIMA model that passes the required controls and is ready for forecasting. The model's forecasts for the next three months are shown in Figure 5.12 and were obtained from the SARIMA(0, 1, 1)(0, 1, 1)12 model estimated at 100% of the available observations. The large and rapidly growing forecast intervals show that the CPI level in Metropolitan Lima could start to rise or fall at any time.

Model	Forecast	95% confidence interval
Abril 2020	134.007	< 133,4728 – 134,542 >
Mayo 2020	134.116	< 133,231 – 135,002 >
Junio 2020	134.242	< 133,110 – 135,375 >
Julio 2020	134.572	< 133,238 – 135,907 >
Agosto 2020	134.846	< 133,337 – 136,356 >

Table 10: *CPI level in Metropolitan Lima estimated for April, May, June, July and August in Metropolitan Lima proposed by the model SARIMA(0, 1, 1)(0, 1, 1)$_{12}$*

From table 10 it can be seen which are the forecasts of the CPI levels of Metropolitan Lima for the months of April, May, June, July and August. As can be seen, the series maintains an increasing trend but with a lower slope than the previous months, this responds to the seasonal component that was identified in the preliminary analysis.

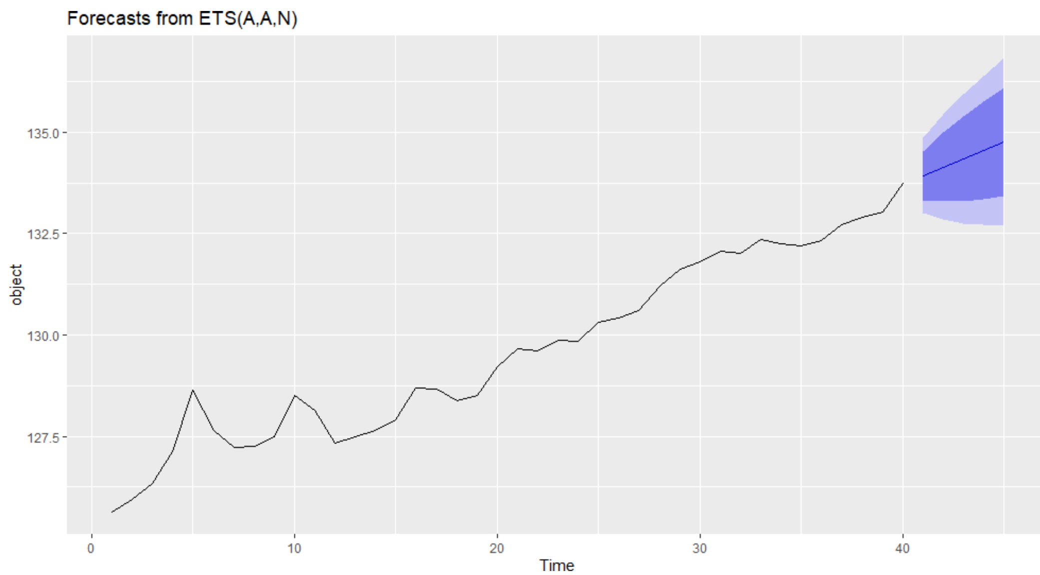

Figure 5.12: *Forecast of the CPI level data in Metropolitan Lima using the SARIMA (0, 1, 1)(0, 1, 1)$_{12}$, model, for the months of April-June 2020*

4.4. Comparison of predicted observations for the second quarter of the year against actual observations

	Real value	Predicted Value
Apr-20	133.96	134.007
May-20	134.23	134.116
Jun-20	133.87	134.242
Jul-20	134.49	134.572
Aug-20	134.35	134.846

Table 11: CPI level in Metropolitan Lima Real vs estimated for April, May, June, July and August

In order to analyze each forecast and real value of the months, it must be remembered that the CPI level of March 133.82 and that of February 132.96, this generated a variation of 0.6%, the highest monthly inflation recorded since 2014, this month was the first month affected by covid-19 and the state of emergency. Now, starting this month, it is seen that the CPI levels have continued to grow.

The levels of Real and Forecast CPI seem to be very similar, therefore we will try to measure the inflation rate generated by these observations.

It can be seen in table 10 and 11 that the real value of the CPI level of Metropolitan Lima for the months of April, May, June, July and August are within the confidence interval of the forecast. Therefore, it can be understood that the impact that the situation caused by Covid-19 may have had on the CPI has been able to be controlled.

To be able to observe in a more meticulous way the little impact that was produced, the percentage variation, the real growth and the expected growth of the CPI levels will be analyzed.

Month analyzed	**Level of Real CPI (%)**	**Predicted CPI level (%)**	**Percentage change (%)**	**Real monthly percentage change (%)**	**Predicted Monthly Percentage Change (%)**
Apr 2020	133.96	134.007	-0.04	0.10	0.14
May 2020	134.23	134.116	0.09	0.20	0.08
Jun 2020	133.87	134.242	-0.28	-0.27	0.09
Jul 2020	134.49	134.572	-0.06	0.46	0.25
Aug 2020	134.35	134.846	-0.37	-0.10	0.20

Table 12: *Impact indicators on the CPI level of Metropolitan Lima*

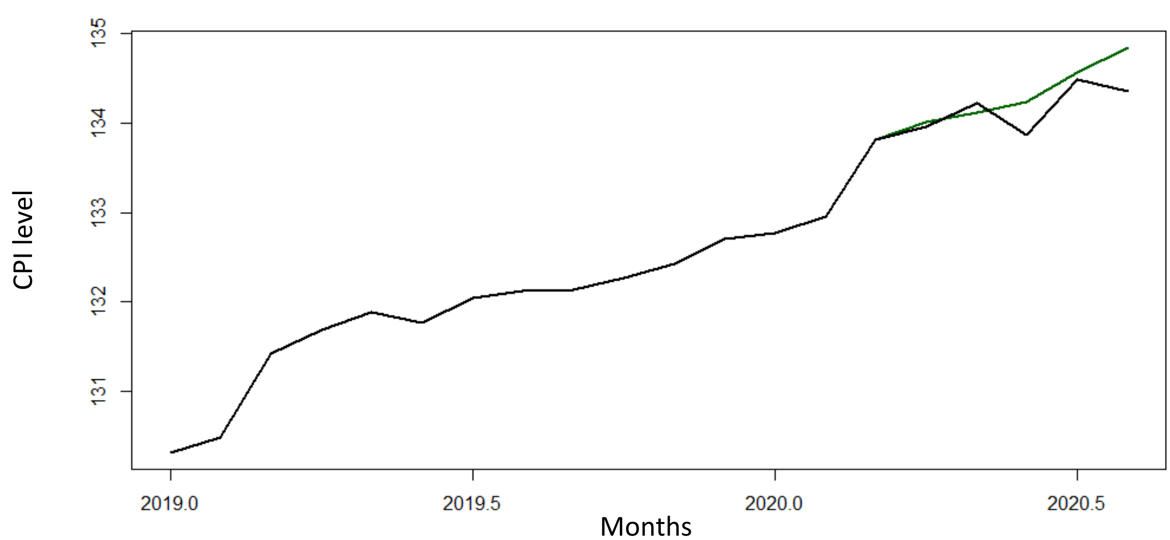

Figure 5.13: *Comparative graph between the real CPI level vs. the predicted CPI level of Metropolitan Lima (January 2019 - August 2020)*

It is observed that the high level of inflation produced in March, tried to be controlled for the month of April, since it is observed that the real impact compared to the forecast was low, this is due to the constant attempt of the government not to reach a crisis financial at the beginning of the quarantine.

The percentage variation indicator allows for support by saying that there has not been a great impact between the real and forecast CPI levels, since table 12 shows a low percentage variation in the months analyzed.

As a result of the economic reactivation that the country had for the month of June, the consumer price index fell 0.27%, this is reflected in the reduction of prices of large consumption groups with that of Food, Beverages. The prices of the Transportation and Communications family were also reduced, as fuel prices decreased. Although other services increased in cost, as is the case of Home Rentals, this did not prevent the economic reactivation from continuing.

The expected and real growth of the CPI level can be compared and it is noted that the behavior of the real CPI has been influenced by various external factors.

CONCLUSIONS

5.1. Resulting SARIMA model without taking into account the impact of Covid-19

According to the results obtained under the different methods, it is determined that the mathematical model SARIMA(0, 1, 1)(0, 1, 1)$_{12}$, is the model that best predicts the monthly information on the CPI levels of Metropolitan Lima..

- The model has a level of differentiation 1 in the seasonal and non-seasonal part, for reasons already described in previous steps, which are caused by the trend. This trend does not allow the series to be stationary. This differentiation helps to make the series stationary.

- The moving averages part of the non-seasonal component has a lag; therefore it can be said that current observations depend on a set of past observations. This explains how the costs of products and services that occurred in the past greatly impact current costs. This also allows us to give a general vision that the next levels of the CPI will still depend a lot on the levels currently being experienced. This is one of the main reasons why the BCRP continually tries to control the percentage variation of the CPI, which is an Inflation Index.

- The seasonal component was also differentiated, since it was observed that when comparing the months of each year on an inter-annual basis against the respective months of the following year or the previous year, the CPI levels increased. This indicates a positive trend in seasonal behavior. In Lima it is seen how the seasons of food, vehicles, recreation, enrollment or other services have been increasing if compared to the previous year.

- The behavior of moving averages of the seasonal component has 1 lag, this only ensures that each observation depends on the history of past errors with one step, this error from the past is kept in the memory of the model to be able to adjust and predict in a better way. efficient.

5.2. Impact of Covid-19 on the Consumer Price Index of Metropolitan Lima

The impact that this inflation indicator has had has been considerable during the beginning of the quarter, since the desperation of people to get basic necessities warned that the supply would decrease and thus the price was set on the rise, but it was possible to control, since although people constantly acquired products to try to be safe from the new coronavirus COVID-19, they also left aside products that they would no longer consume, such as restaurants, transportation, recreation, entertainment, cultural services, etc.

For the economic reactivation, the government supported with different credits or loans to the companies, called **reactive** loans, this project allowed many companies not to go bankrupt and thus could continue offering services without reaching the need to increase the cost of these.

People also received government support through bonds that were offered from time to time. It should be noted that in a stable economy the increase in the CPI can be a beneficial indicator, since d shows that people spend more, but not because the products cost more, but because they can buy more products in savings.

BIBLIOGRAPHY

CENTRAL RESERVE BANK OF PERU (2020). «Inflation Report» Current Program and Macroeconomic Projections. Lima Peru.

CENTRAL RESERVE BANK OF PERU (2019). «Methodological Guide of the Weekly Note» Inflation. Lima Peru.

NATIONAL INSTITUTE OF STATISTICS AND INFORMATICS [INEI] (2019). National survey of family budgets (Enapref) Lima, Peru.

CENTRAL RESERVE BANK OF PERU [BCRP] (2020). «Central Management of Economic Studies, BCRP Data». Lima Peru. (https://estadisticas.bcrp.gob.pe/estadisticas/series/mensuales)

NATIONAL INSTITUTE OF STATISTICS AND INFORMATICS [INEI] (2001). Methodology for calculating the consumer price index of Metropolitan Lima

WALTER ENDERS (2014). Applied Econometric Time Series.4rd Edition. Wiley.

LUIS, A. (2011). Consumer price index as a statistical method to measure inflation in Ecuador College of Economic Sciences, University of Guayaquil.

TATIANA, E. PAULA, A. NATALIA,P. LAURA L. (2016). Analysis, modeling and forecasting of the series of consumer price indices in Medellín 2010-2015 Universidad Pontificia Bolivariana, Medellín.

JONATHAN D. CRYER y KUNG-SIK CHAN (2008). Time Series Analysis with Applications in R, Second Edition, Springer.

TERENCE C. MILLS (2019). Applied Time Series Analysis, A Practical Guide to Modeling and Forecasting Academic Press.

TECHNICAL DIRECTORATE OF ECONOMIC INDICATORS (2010). Methodology for the change of base years 2009 of the Consumer Price Index of Metropolitan Lima. Lima, Peru.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. (2013). Peruvian technical standard ISO 3534-1. Statistics. Vocabulary and symbols. Part 1. General statistical terms and terms used in the calculation of probabilities. Peru

WORLD HEALTH ORGANIZATION [WHO] (2020). « Questions and Answers About Coronavirus Disease» (https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q- a-coronaviruses)

APPENDIX

Appendix A: INEI standard definition for the Consumer Price Index:

It is a statistical indicator that measures the behavior of prices, from one period to another, of a set of products (goods and services) representative of the expenditure of the population of Metropolitan Lima.

We say that the CPI is a Statistical Indicator because for its construction, it uses tools provided by Statistical science, allowing us to estimate the behavior of a population from sample information. Furthermore, it does not show the simple variation of a price, but the average variation of a set of geographically and temporally distributed prices.

It includes a set of goods and services consumed by households, that is, it does not refer exclusively to a particular household or group of households, but is representative of all households that exist in Metropolitan Lima.

This indicator refers to habitual household consumption and consumption patterns in a certain base period, assuming as a hypothesis that these structures are maintained during the validity of the base period.

The basket of goods and services that serves as the basis for calculating the CPI corresponds to an average consumption of the inhabitants of Metropolitan Lima and, therefore, it should not be expected that such an indicator reflects exactly the changes experienced by a household or by a person in particular, since it is unlikely that any consumer will buy everything included in the list and in the same quantities and qualities.

The 8 large consumption groups that make up the family basket for calculating the CPI are:

- Food and drinks

- Dress and footwear
- Home rental fuel and electricity
- Furniture fixtures and home maintenance
- Measured care and maintenance of health and service
- Transport and communication
- Recreation, entertainment, cultural and teaching services
- Other goods and services

Basic Data for the elaboration of CPI

- Define the Reference Population, which shows the behavior of prices.
- Define the Goods and Services consumed, as well as the Household Consumption Structure.
- Define the prices to the Consumer in the base period of said goods and services.
- Perform periodic monitoring of prices.

Appendix B: INEI Standard for the Design of the Sample of the National Survey of National Budgets ENAPREF

B.1. Population

The population was defined as the set of habitual residents in private homes in urban and rural areas throughout the national territory. The population residing in collective dwellings, such as barracks, police stations, convents, boarding schools and hotels, were excluded from the study.

B.2. Sampling frame

The basic sampling frame was made up of the information on the populated centers and conglomerates defined in the CPV 2005 for the rural area and the precensus 2007 for the urban area.

In order to minimize the variances of the estimates, the populated centers of each department were stratified according to their population size, which makes it possible to ensure that all the sub-populations of the department are adequately represented in the sample.

Sample design in order to reduce the sample variance generated by the cluster sampling, these were stratified from the socioeconomic point of view and ordered in the sampling frame.

The clusters were ordered within the populated centers according to this indicator. With this, an implicit stratification of the frame is obtained when the sample is systematically selected with probability proportional to size.

Secondary Sampling Unit (USM) The USM in both urban and rural areas were the private dwellings existing in said conglomerates or AER. The selection was made by systematic sampling with random start.

The type of sampling developed resulted in a probabilistic sample, of areas, stratified, and independent in each department.

Appendix C: ESS Guidelines on Seasonal Adjustment

The recommendations contained in the ESS guidelines on seasonal adjustment are not exclusively restricted to the seasonal adjustment process, but cover other aspects such as the pre-treatment of the series, the revision policy and the quality of adjustment.

Appendix D: Standards of Statistical Definitions, Vocabulary and Symbols ISO-3534-1

- **Population**: Totality of the elements taken into consideration.
- **Sampling unit or sampling unit:** One of the individual parts into which a population is divided.
- **Sample**: Subset of a population composed of one or more sampling units.
- **Observed value:** Value obtained from a property associated with an element of a sample.
- **Hypothesis**: Claim about a population.
- **Null hypothesis (H0)**: Hypothesis to be evaluated by means of a statistical test.
- **Alternative hypothesis: (H1)**: An assertion that considers a set or a subset of all possible admissible probability distributions that do not belong to the null hypothesis.

- **p- value**: Probability of obtaining the value of the test statistic or any other value at least as unfavorable for the null hypothesis.
- **Descriptive statistics:** Graphical, numerical or other summary representation of the observed values.
- **Random variable:** Function defined in the sample space where the values of the function are ordered k-tuples of real numbers.
- **Sample mean:** Arithmetic mean sum of the random variables of a random sample divided by the number of elements in the sum.
- **Sample variance**: Sum of the squares of the deviations of the random variables of a random sample with respect to its sample mean, divided by the number of terms in the sum minus 1
- **Sample space**: Set of all possible outcomes.
- **Statistical test:** Significance test, a procedure to decide whether a null hypothesis should be rejected in favor of an alternative hypothesis.

Appendix E: GUIDANCE STANDARDS ON STATISTICAL TECHNIQUES ISO-10017

E.1. Time series analysis

Time series analysis is a family of methods for studying a collection of observations made sequentially in time. It is used for analytical techniques in applications such as:

- The search for "lag" patterns by statistically analyzing how an observation correlates with the immediately preceding observation, and repeating this for each successive separation period.
- A scatter plot, which helps to evaluate the relationship between two variables, by graphing one variable on the x-axis and the corresponding value of the other on the y-axis.
- The use of statistical tools to predict future observations or to understand which causal factors have contributed the most to the variations in a time series.

E.2. Hypothesis testing

Hypothesis testing is a statistical procedure to determine, with a prescribed level of risk, whether a data set is compatible with a given hypothesis. The hypothesis can be related to an assumption of a particular statistical distribution or model, or it can be related to a value of some parameter of a distribution.

The procedure for a hypothesis test involves evaluating the evidence to decide whether a given hypothesis regarding a statistical model or parameter should be rejected or not.

Hypothesis testing allows you to make a statement about some parameter in the data set, with a known level of confidence. This being the case, it can be helpful in making decisions that depend on the parameter.

E.3. Sampling

Sampling is a systematic statistical method to obtain information about some characteristic of a population by studying a representative fraction of the population (that is, the sample). There are several sampling techniques that can be used (such as simple random sampling, stratified sampling, systematic sampling, sequential sampling, skip-lot sampling, etc.), and the selection of techniques is determined by the purpose of the sampling and the conditions under which it will be carried out.

Appendix F: Limitations

F.1. Time series analysis

When modeling a process to understand its causes and effects, great skill is required to select the most appropriate model and to use diagnostic tools to improve it. In addition, you must have a large number of observations since the omission of a small observation or group of observations can bring a significant consequence in the model.

F.2. Hypothesis testing

To ensure the validity of the conclusions reached from the hypothesis tests, it is essential that the basic statistical assumptions are adequately satisfied. Furthermore, the level of confidence with which the conclusion can be made is governed by the size of the data set. At a theoretical level, there is a debate as to how a hypothesis test can be used to make valid inferences.

F.3. Sampling

In constructing a sampling plan, attention should be paid to decisions regarding sample size, sampling frequency, sample selection, the basis for subgroups, and various other aspects of sampling methodology.

Sampling requires that the sample be selected free of bias (that is, the sample is representative of the population from which it was drawn). Failure to do this will result in a poor estimate of the characteristics of the population. In the case of acceptance sampling, unrepresentative samples may result in the unnecessary rejection of lots of acceptable quality, or the improper acceptance of lots of unacceptable quality.

Even with samples free of bias, the information derived from samples is subject to a certain degree of error. The magnitude of this error can be reduced by taking a larger sample size, but it cannot be eliminated. Depending on the specific question and the context of the sampling, the sample size required to achieve the desired level of confidence and precision may be too large to be of practical value.

Appendix G: Coding in R

```
inflacion <- read_xlsx("IPC.xlsx", col_types = "text")

IPC <- inflacion  %> %
  mutate(Tasa = as.double(Tasa))

length(IPC$Tasa)-12*18

Train_T <- IPC[1:216,]  %>%
  mutate(Tasa = as.double(Tasa))

Test_T <- IPC[217:(length(IPC$Tasa)-5),] %> %
  mutate(Tasa = as.double(Tasa))

ipc  <-  IPC[217:(length(IPC$Tasa)),]  %>%
  mutate(Tasa = as.double(Tasa))

IPC  <-  IPC[1:(length(IPC$Tasa)-5),]  %>%
  mutate(Tasa = as.double(Tasa))

par(mfrow=c(1,1))
IPC  <- ts(IPC$Tasa,start = c(2001,1),frequency = 12)
IPC
plot(IPC)
```

```
        # Unemployment train series


Serie_Train_T   <- ts(Train_T$Tasa,start = c(2001,1),frequency = 12)Serie_Train_T


# Unemployment Test series

Serie_Test_T   <- ts(Test_T$Tasa,start = c(2019,1),frequency = 12)

Serie_Test_T par(mfrow=c(1,1))


ipc   <- ts(ipc$Tasa,start = c(2019,1),frequency = 12)ipc

plot(ipc)


plot(IPC , main="Nivel de IPC en Lima Metropolitana, 2001-2020",ylab="Tasa de desempleo",
        xlab="",plot.type = "single",col="
             black")


lines(decompose(Serie_Train_T)$trend,col=2)library(tseries) #install.packages("tseries")

adf.test(Serie_Train_T,alternative = "stationary",k=0)adf.test(Serie_Train_T)


##Model 1.

library(fpp2)


# Exploratory Analysis, to specify the SARIMA Model par(mfrow=c(1,1))


plot(Serie_Train_T,col="blue", type="o")

acf(Serie_Train_T,lag.max=48,na.action=na.pass,col="darkmagenta",ci
     =0.95,main="fac de Tasa de desempleo")

pacf(Serie_Train_T,lag.max=48,na.action=na.pass,col="darkmagenta",
     ci=0.95,main="facP de Tasa de desempleo")        # Series difference


diff.des <-autoplot(diff(Serie_Train_T), ts.linetype = "dashed", ts
```

```
    .colour = "blue") diff.des adf.test((Serie_Train_T))          #Test of stationarity
adf.test(diff(Serie_Train_T,lag = 1)) adf.test(diff(Serie_Train_T,lag = 1) %>% diff(lag=12))
# A second differentiation is applied

adf.test(diff(diff(Serie_Train_T,lag = 1))) #Nice, this is

# Series Difference Graphics       par(mfrow=c(3,1)) autoplot(diff(diff(Serie_Train_T,lag=1)))
autoplot(acf(diff(Serie_Train_T,lag = 1) %>% diff(lag=12), lag.max
    =48))
#acf(diff(diff(Series_Unemployment)))

autoplot(pacf(diff(Serie_Train_T,lag = 1) %>% diff(lag=12), lag.max
    =48))
#acf(diff(diff(Series_Unemployment)))
diff(diff(Serie_Train_T,lag = 1)) %>% diff(lag=12) %>% ggtsdisplay(lag.max=48)
diff(Serie_Train_T,lag = 1) %>% diff(lag=12) %>% ggtsdisplay(lag.max=48)

fit5 <- Arima(Serie_Train_T, order=c(1,1,2), seasonal=list(order =c(0,1,1),period=12))
coeftest(fit5)#(1,1,2)(0,1,1)
checkresiduals(fit4)summary(fit5)
fit6 <- Arima(Serie_Train_T, order=c(2,1,1), seasonal=list(order =c(0,1,1),period=12))
coeftest(fit6) checkresiduals(fit6)summary(fit6)
fit7 <- Arima(Serie_Train_T, order=c(1,1,0), seasonal=list(order =c(1,1,2),period=12))

coeftest(fit7) checkresiduals(fit7)summary(fit7)

fit8 <- Arima(Serie_Train_T, order=c(1,2,0), seasonal=list(order =c(1,1,1),period=12))
coeftest(fit8) checkresiduals(fit8)summary(fit8)

fit9 <- Arima(Serie_Train_T, order=c(0,1,1), seasonal=list(order =c(0,1,1),period=12))
```

```r
summary(fit9)

coeftest(fit8) checkresiduals(fit8)

# Forecast

Pronos1 <- fit5 %> % forecast(h=15) Pronos2 <- fit6 %> % forecast(h=15) Pronos3 <- fit7 %> %
forecast(h=15)Pronos4 <- fit8 %>% forecast(h=15)

# EVALUATION

Box.test(r1, lag = 1, type = c("Box-Pierce", "Ljung-Box"), fitdf =0)

par(mfrow=c(1,1))

plot(seq(1:length(Serie_Test_T)),type="l",Serie_Test_T,main="Tasa de desempleo de hombres en
      Lima Metropolitana, Abril 2019 - Marzo 2020",lwd=2,xlab="Meses",ylab="Tasa de
      desempleo")
lines(seq(1:length(Pronos1$mean)),Pronos1$mean,col="darkorange")
lines(seq(1:length(Pronos2$mean)),Pronos2$mean,col="darkmagenta",
      lwd=2) lines(seq(1:length(Pronos3$mean)),Pronos3$mean,col="darkgreen",lwd
      =2)
lines(seq(1:length(Pronos4$mean)),Pronos4$mean,col="red",lwd=2)

            legend("topleft",c("Serie Real","SARIMA (1,1,2)(0,1,1)12","SARIMA
              (2,1,1)(0,1,1)12","SARIMA (1,1,0)(1,1,2)12","SARIMA (1,2,0)
                                (1,1,1)12"),
        col=c("black","darkorange","darkmagenta","darkgreen","red"),
              ncol=1,bty="n",y.intersp=0.8,lty=2,lwd=2,cex=0.7)
```

```
fit1  %>% forecast(h=12)  %>% autoplot()

## Calculation of Prediction Quality Indicators     calipre<-function(y,e){ # Prediction Quality
function
    ma<-mean(abs(e), na.rm=TRUE) # MAEms=mean((e)^2, na.rm=TRUE) # MSE mrs=ms^0.5
    mp<-mean(abs(e/y), na.rm=TRUE)*100# MAPEreturn(list(mae=ma,rmse=mrs,mape=mp))
}

#RESIDUAL

r1 = Serie_Test_T - Pronos1$mean r2 = Serie_Test_T - Pronos2$mean r3 = Serie_Test_T -
Pronos3$meanr4 = Serie_Test_T - Pronos4$mean

calipre(Serie_Test_T,r1) calipre(Serie_Test_T,r2) calipre(Serie_Test_T,r3) calipre(Serie_Test_T,r4)

# Forecast 2020

fit1_pron <- Arima(IPC, order=c(1,2,0), seasonal=list(order = c(1,1,1),period=12))
fit1_pron <- Arima(IPC, order=c(0,1,1), seasonal=list(order = c(0,1,1),period=12))
checkresiduals(fit1_pron)

length(IPC)

fit1_pron %>%  forecast(h=5) %>%  autoplot()Pronos1 = fit1_pron  %>% forecast(h=5)
Pronos1
anadir <- Pronos1$mean as.vector(anadir) length(fit1_pron$fitted) length(fit1_pron$fitted)
fit1_pron$fitted[192:231]  %>%  forecast(h=5)  %>%  autoplot()length(fit1_pron$fitted)

#------------------- Forecast Graph-----------------
```

```
serie_vector<-as.vector(Test_T$Tasa)n <- length(serie_vector) serie_vector[(n+1):(n+5)] <- anadir

Serie_Test_T_1   <- ts(serie_vector,start = c(2019,1),frequency =12)
Serie_Test_T_1

plot(ipc,main="Niveles de IPC de Lima Metropolitana
        Enero 2019 - Agosto 2020",col ="darkgreen",lwd=2,xlab="Meses",ylab="Nivel de IPC")
lines(Serie_Test_T_1,col="darkgreen",lwd=2)


plot(Serie_Test_T_1,main="Niveles de IPC de Lima Metropolitana Enero 2019 - Agosto 2020",col
        ="darkgreen",lwd=2,xlab="Meses",
            ylab="Nivel de IPC")


lines(ipc,col="black",lwd=2)
```

Appendix H: Data

Month | Year	2001	2002	2003	2004	2005	2006	2007
January	81.78	81.1	82.95	85.27	87.86	89.53	90.1
February	81.98	81.07	83.34	86.2	87.65	90.02	90.34
March	82.4	81.51	84.27	86.6	88.22	90.43	90.65
April	82.06	82.1	84.23	86.58	88.33	90.89	90.81
May	82.08	82.21	84.2	86.88	88.44	90.41	91.26
June	82.03	82.03	83.81	87.37	88.67	90.29	91.69
July	82.17	82.06	83.68	87.54	88.76	90.14	92.12
August	81.92	82.14	83.69	87.53	88.6	90.26	92.25
September	81.97	82.53	84.16	87.55	88.52	90.29	92.82
October	82	83.12	84.2	87.53	88.65	90.33	93.11
November	81.6	82.79	84.34	87.78	88.71	90.07	93.21
December	81.53	82.76	84.82	87.77	89.08	90.09	93.63

Figure H. 1: *CPI level of Metropolitan Lima (2001-2007)*

Month | Year	2008	2009	2010	2011	2012	2013	2014
January	93.84	99.97	100.4	102.58	106.92	109.99	113.36
February	94.69	99.89	100.73	102.97	107.26	109.89	114.04
March	95.68	100.25	101.01	103.7	108.09	110.89	114.63
April	95.83	100.27	101.03	104.4	108.66	111.17	115.08
May	96.18	100.23	101.27	104.38	108.7	111.38	115.34
June	96.92	99.89	101.53	104.48	108.66	111.67	115.53
July	97.46	100.07	101.9	105.31	108.76	112.29	116.03
August	98.03	99.87	102.17	105.59	109.31	112.9	115.93
September	98.59	99.78	102.14	105.94	109.91	113.02	116.11
October	99.2	99.9	101.99	106.28	109.73	113.06	116.55
November	99.5	99.79	102	106.74	109.58	112.82	116.38
December	99.86	100.1	102.18	107.03	109.86	113	116.65

Figure 8.2: *CPI level of Metropolitan Lima (2001-2007)*

Month | Year	2015	2016	2017	2018	2019	2020
January	116.84	122.23	126.01	127.59	130.31	132.77
February	117.2	122.44	126.42	127.91	130.48	132.96
March	118.1	123.17	128.07	128.54	131.42	133.82
April	118.56	123.19	127.74	128.36	131.69	133.96
May	119.23	123.45	127.2	128.38	131.88	134.23
June	119.62	123.62	127	128.81	131.77	133.87
July	120.16	123.72	127.25	129.31	132.04	134.49
August	120.61	124.16	128.1	129.48	132.12	134.35
September	120.65	124.42	128.08	129.72	132.13	
October	120.82	124.93	127.48	129.83	132.27	
November	121.24	125.3	127.23	129.99	132.42	
December	121.78	125.72	127.43	130.23	132.7	

Figure 8.3: CPI level of Metropolitan Lima (2001-2007)