Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 1673–1681
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5783

Forecasting the numbers of cardiac diseases patients by using Box-Jenkins model in time series analysis

Sabah Hasan Jasim Alsaedi[a,*]

[a]Mathematic Department - College of Basic Education, Misan University, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

The aim of this study is analysis time series with using (Box and Jenkins) method by identification , estimation, diagnosis, checking of model ,forecasting to find the beast forecasting model to the number of patient with cardiac in Misan province by using the monthly data of the period (2005-2016) by using SPSS version (26).The result of data analysis show that the proper and suitable model is Autoregression of order ARIMA (1,1,0) .According to this model the study forecast the numbers of patients with cardiac diseases the next years in monthly , so the forecasting values represented the scours time series data that deal to the efficiency of the model.

Keywords: Forecasting, Cardiac diseases, Box Jenkins , Time series analysis

1. Introduction

Cardiac diseases is an abnormal functioning of the heart or generate term for different of heart condition [1]. Risk factors as diabetics, hypertension, cholesterol, Obesity, lack of exercises, smoking, increased age, family history [2]. Time series analysis is a statistical tool that was deals with times series data. Time series data refers to in a series of particular period's .The data was divided into three types times data, cross section data and pooled data. Times series predicting used, for observed relationship between data and forecasting predict future data this called lead-time. The purpose for prediction data points by production control and planning in industrial, medical economic and other field [3].Some model for times series prediction such as the box-Jenkins or Autoregression integrate moving average (ARIMA) model [4].

Time Series Models: A time series is a set of values of a particular variable that occur over a period in a certain pattern. The most common patterns are increasing or decreasing trend, cycle,

*Corresponding author
Email address: sabah.h.alsaedi@gmail.com (Sabah Hasan Jasim Alsaedi)

Received: May 2021 *Accepted:* October 2021

seasonality, and irregular fluctuations (5). Time series event as a function of its past values, analysis identify the pattern with the assumption that the pattern will persist in the future by using Box -Jenkins.

Time series analysis: It is a set of observations of apparent values $(x_1\ ,\ x_2\ ,x_3,\ \cdots,x_t)$. at specific times of time$(t_1\ ,\ t_2\ ,t_3,\ \cdots,t_k)$. and the interval may be equal or unequal. If it is equal, it expressed as

$$x_t = F\,(t) + \ \varepsilon_t \qquad t = 0\ ,\ \pm 1\ , \pm 2\ ,\ \pm 3\ ,\ \cdots\ , m$$

Where is:
x_t: Value of string variables
t_k: Time periods
$F(t)$: Regular part
ε_t: White Noise process
Box Jenkins' methodology is based on four parts:

A-Autoregression model (AR):

The autoregressive model represented the relationship between the current and past values of the time period, and it is used in various fields, including describing a particular phenomenon, whether it is natural or economic. And this approach to analyzing time series models is to access the mathematical model that represents the data , where the autoregressive model is one of the important models to achieve this goal, when the current value of the time series is a function of its value for the previous period in addition to some errors, the models formed from this process are called Autoregression models [6, 7].

Autoregressive model (AR): It was been symbolized by a symbol (P) which is an integer as:

$$x_t = \ \emptyset_\circ + \ \emptyset_1 x_{t-1} + \ \emptyset_2 x_{t-2} + \emptyset_3 x_{t-3}\ + \cdots + \emptyset_p x_{t-p}\ \ + \varepsilon_t$$

Or

$$\emptyset_p\,(b)\,x_t = \ \emptyset_\circ + \ \varepsilon_t$$

Where is:
x_t: Value of string variables
$\emptyset_p$: Model parameters
$\emptyset_\circ$: Constant
p : Model grade
ε_t : Random errors

2- Moving Average Model (MA): - The value of the functional segment can be obtained in real time for the current and previous periods, and the model from this process is called the moving averages model. It was been used as Moving Average Model in a series times instead Auto regression Model (AR) and refer to Moving Average model (MA) (q) as the following (8):

$$x_t = \ \emptyset_\circ + (\ 1 - \vartheta_1\ b - \ \vartheta_2 b^2 - \vartheta_3 b^2 - \ \cdots - \ \vartheta_q b^q)\varepsilon_t$$

Finally formula:

$$x_t = \ \emptyset_\circ + \varepsilon_t - \vartheta_1 \varepsilon_{t-1} - \ \vartheta_2 \varepsilon_{t-2}\ - \vartheta_3 \varepsilon_{t-3} - \ \cdots - \ \vartheta_q \varepsilon_{t-q}$$

Where is:
$\emptyset_i$: Parameters of the model for moving media**.**

$$-1 < \emptyset < 1 \qquad\qquad i = 1, 2, \cdots\ ,\ q$$

q : Model grade.

3-Auto Regression -Moving Average Mixed Model (ARMA): The two previous models can be combined into a single model called ARMA, and the new model becomes the following relationship: In some cases of time series mixed between two model Auto regression (P) and Moving Average Model (q) and its called Auto regression moving average mixed model(p, q) ARAMA as the following (8): as Mixed Autoregression Moving Average Model (ARMA) equation

$$x_t = \emptyset_\circ + \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + \varepsilon_t - \vartheta_1 \varepsilon_{t-1} - \cdots - \vartheta_q \varepsilon_{t-q}$$

By Using Recoil factor (b) as:
Whereas,

$$\emptyset_{\mathrm{p}}(\mathrm{b})\, \mathrm{x_t} = \emptyset_\circ + \emptyset_{\mathrm{q}}(\mathrm{b})\, \varepsilon_{\mathrm{t}}$$

is $\emptyset_p(b)$ A polynomial in (b) for the parameters of the autoregressive model
is **:** $\emptyset_q(b)$ A polynomial in (b) for the parameters of the moving media model

4- Auto regression integrated moving average model (ARIMA) : Time series models are the most widely used models, as it is not possible to derive all models from them, whether autoregressive or moving averages or mixed, and these models consist of three parts, the first part of which represents the autoregressive model and moving averages into a model (ARIMA) .To the integrated autoregressive model, where p represents the autoregressive rank d the number of differences, the integration q, the moving average model q (8) .Some time series models may be unstable on their own but become stable after a lot of transformations or differences, so the model that expresses this process will differ from the original model , as it must to include those transformations or differences autoregressive $p-(AR)$ model, which is usually used in the process of predictions for the time series, and the other part represents .The moving average model $(q) - MA$ and the third part represents the differences that the chain requires in order to be stable Therefore it expresses mixed models that where: ARIMA (p, d, q)
P- (AR): is the autoregressive model
q- (MA): is the order of the moving average model
d : is the number of differences that make the series stable
The formula of Autoregression Integrated Average Models (ARIMA) and stable with different with order (9-10) as;

$$x_t = \emptyset_\circ + \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + \cdots + \mathrm{d}x_{t-p-d} + \varepsilon_t - \vartheta_1 \varepsilon_{t-1} + \cdots + \vartheta_q \varepsilon_{t-q}$$

Non-Stationary times series: By using Box- Jenkins model as
Box-Jenkins Model: The process of building the model for the ready-made time series requires a great effort as it considered a method (interactive methods) and consist of many stages:
1-Identification
2- Estimation
3-Diagnostics Checking
4- Forecasting
These stages are very important in box-Jenkins model but identification and estimation is very important stages if wrong in theses stages lead to incorrect results.
1-Identifiaction: The most important step in building a time series model by observing the drawing of the original data and the eigenvector and partial correlations. But if it is stable in the mean and variance, it must the stability times series by taken $(d-1)$ the $(d-2)$ but usually stable after $(d-1)$

and $(d-2)$ by using ACf (Autocorrelation model)and PACF (partial correlation model), If there is a decrease in the ACF at the expense of PACF, it is considered to be the appropriate model (p,d,q) . The middle element d is investigated before p and q to adjacent determine if the process is stationary or not and make it . Stationary or not and make it stationary before determine the value p and q .Auto-correlation component as a memory of the process for preceding observations .The letters p or value p is the number of Auto regression in ARIAM as (p,d,q) model . as
$P=\ 2, ARIMA\ (1,1,0)$ is
The most tools in identification process is Auto-correlation function and partial Auto- correlation function:

1-Auto correlation function (PACF): It is one of the important statistical methods in knowing the stability of the time series with the constant arithmetic mean and variance, and it has a major role in diagnosing and determining the type of model(9). The following autocorrelation coefficient is estimated as the following:

$$\gamma_k = \frac{\sum_{t=1}^{n-d} (x_t - \overline{x}\ )(x_{t+k} - \overline{x})}{\sum_{t=1}^{n} (x_t - \overline{x})^2}$$

x_t: *Value of string variables*

$\overline{x}$: *The arithmetic mean and the equal* $\qquad \overline{x} = \frac{\sum_{t-1}^{n} x_t}{n}$

n : *Sample size*

2-Partial Auto Correlation function: The partial correlation coefficient is a measure of the degree of relationship between pedestrians t and $x_{t+m}$and stable others x_{t+1}, x_{t+2}, ..., x_{t+m-1} , and that a function of partial correlation (PACF)is no less important than the function of autocorrelation(ACF) in an important function in analyzing time series data and it is used in diagnosing the model and determining its degree and in examining and fitting the model through the random residual test as(9):

$$\gamma_{kk} = \frac{\gamma_k - \sum_{i=1}^{k-1} \gamma_{k-1,i} \cdot\ \gamma_{k-i}}{1 - \sum_{=1}^{k-1} \gamma_{k-1,i}\ \cdot\ \gamma_i} \qquad\qquad k = 1,2,3,\ \ldots.$$

B- Estimation model parameter:

Estimation the value of model consist of estimation parameters from an Auto-Regression or from a moving average model as parameter differ significantly from zero and all significant parameters. It must be include in model or all auto regression parameters ϑ between -1 and 1 if there is 2 parameters $p=2$ as:

$$\vartheta_1 + \ \emptyset_2\ < 1 \ and\ \vartheta_2 + \ \emptyset_1\ < 1$$

C-Diagnosis of checking stage: At this stage, the model is been chosen to know its suitability to represent the data of the studied phenomenon and use it to obtain future predictions. It directs many tests that can be used for this purpose and which depend in its calculations on the protective factor to detect whether there is any factor other than randomness within these residuals, as we assume these remainders will be random and devoid of any other influence such as tests as Mean squared error (MSE) and Mean Absolute Error (MAE) .

D-Forecasting stage: One of the main objectives of time series analysis is one of the main objectives in the analysis of time series is predicting its future values. This stage is been considered the last stage of Box Jenkins, where prediction is the choice of choosing the future model of the time series. It is correct and accurate if a valid model is completed a proper pass and passing the examination and diagnosis stage. After determining the models of the (p,d,q) model, which confirmed that it is

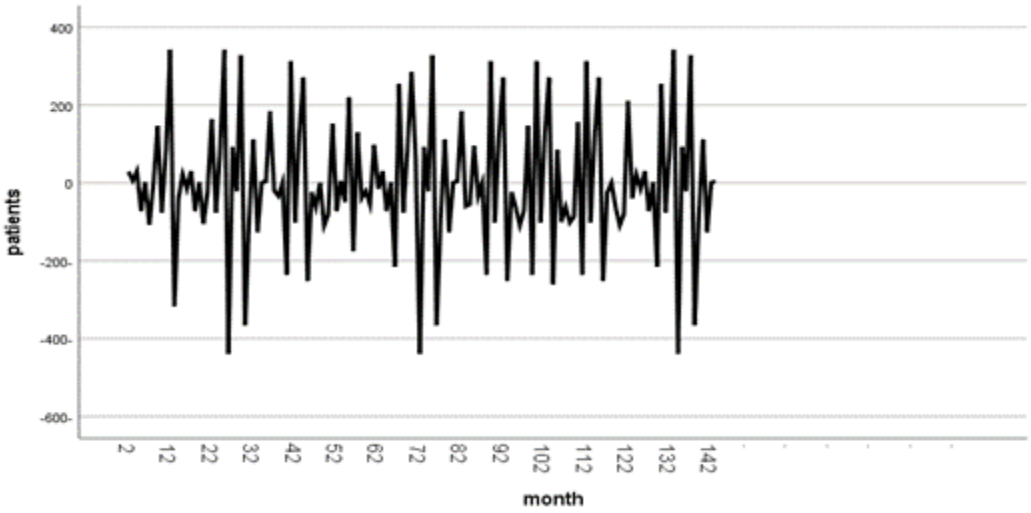

Figure 1: Time series value for Cardiac diseases patients

the best model according to the previous tests, then it is used in prediction by putting the present and past values of the variable $(x_{t)}$, and the residuals e_t as an estimation value for the error term to the right of the function, to get the first predicted future value, which is called prediction for one future period, and the second future value can be obtained, (x_{t+1})As estimation values for the error term in the right of the function in order to get the first predicted future value, which is called prediction for one future period. The second future value (x_{t+2}) can be obtained by placing the first future value (x_{t+1}) that was connected to in the first step to predict in the model equation to be predicted for future periods (9).

Applied practical:
Data were been obtained from the Cardiology Center in Misan province. The main objective of time series analysis is to build the best prediction model, determine the number of its features, estimate them, and ensure the appropriateness of choosing the model for the data. The drawing of the time series is the first step in the analysis process, where it is possible through the drawing to identify some of the characteristics of the series in a preliminary way, and the distinction is made.

Data analysis using Box's Jenkins model:
The main objective of the time series analysis is to build the best prediction model, determine the number of its features and estimate them, as well as ensure that the model test is suitable for the data as in Figure 1.The time series is drawn, which is the first step in the analyzes, where the series is drawn with its first images on some of its properties, where we notice the difference in the form of fluctuations and that the series has a general trend, which indicates the stability of the series. Distinguishing between stable and unstable time series is done by using autocorrelation coefficients, whose value is close to zero, and converting the unstable series to a stable series is done using the method of differences.
By taking the first difference and converting it to a stable series, we take the first difference and then extract the coefficients and plot them as indicated and PACF in Figure 2, which indicates and Figure 3, as ACF which represents by observing the two figures, we find that the time series is stable by identification of model (Diagnosis model) and degree with ACF and PACF methods and finally in figure 4 the diagnosis model is (1,1,0)ARIMA as the forma :

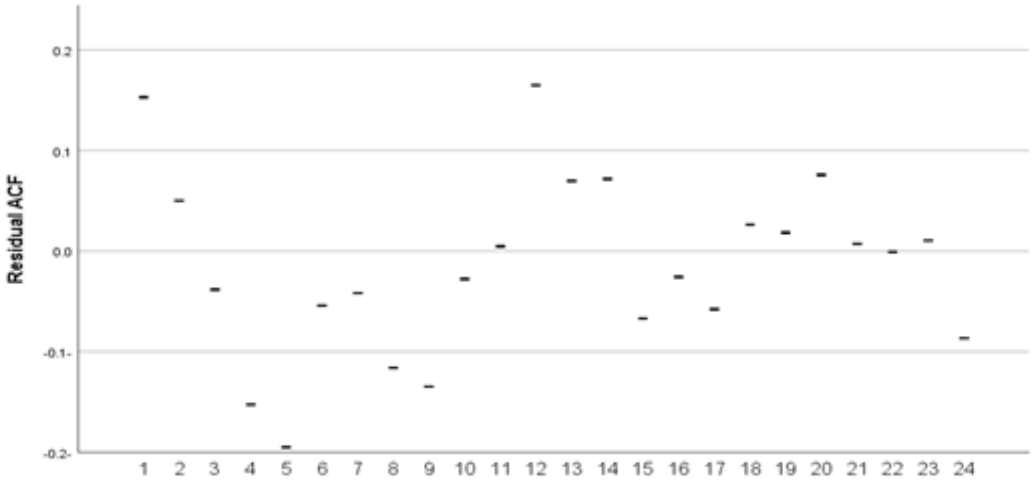

Figure 2: Residual ACF in series time analysis

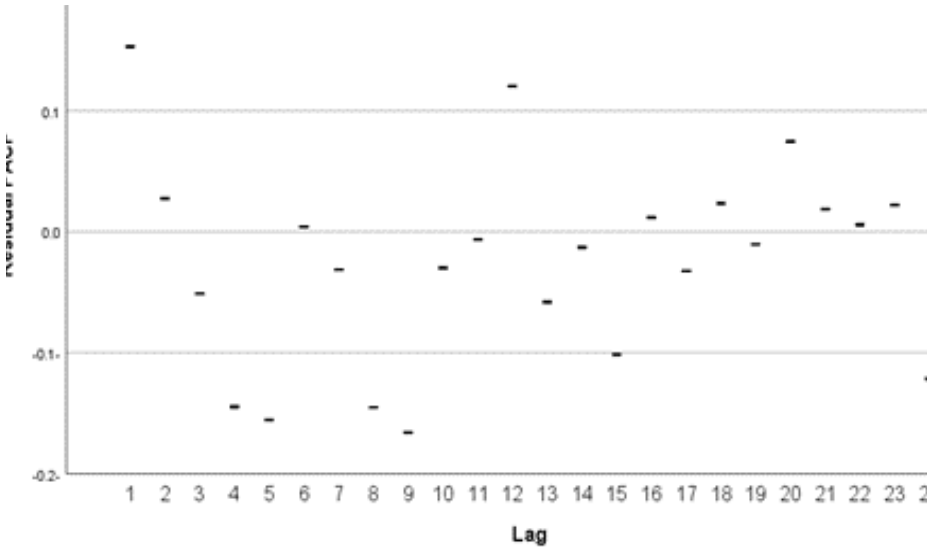

Figure 3: Residual PACF in series time analysis

Diagnostic checking of model and Estimation: After obtaining the stable time series, the model ARIMA was been determined by its degree, depending on the behavior of the autocorrelation and partial autocorrelation functions and use the less RMSE .
RMSE: Average residual squares.
MAPE: The average ratios of the absolute values of the remainders.
From figure 4 the best ARIMA model is (2, 1, 0) this less value in RMSE by use SPSS version(26). In table 2 and figure 5 show forecasting and predictive value of cardiac patients.

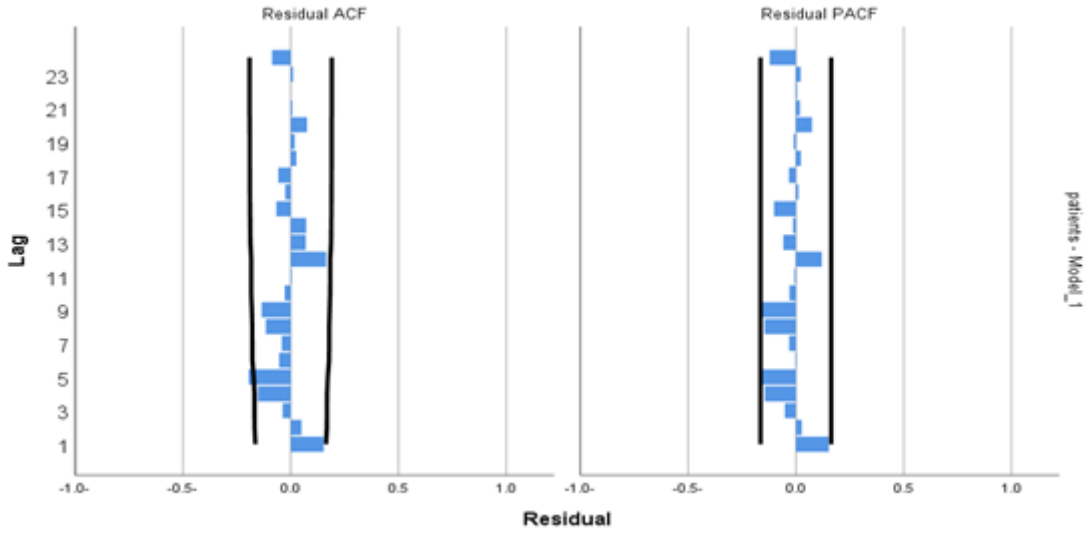

Figure 4: Autocorrelation and partial Autocorrelation

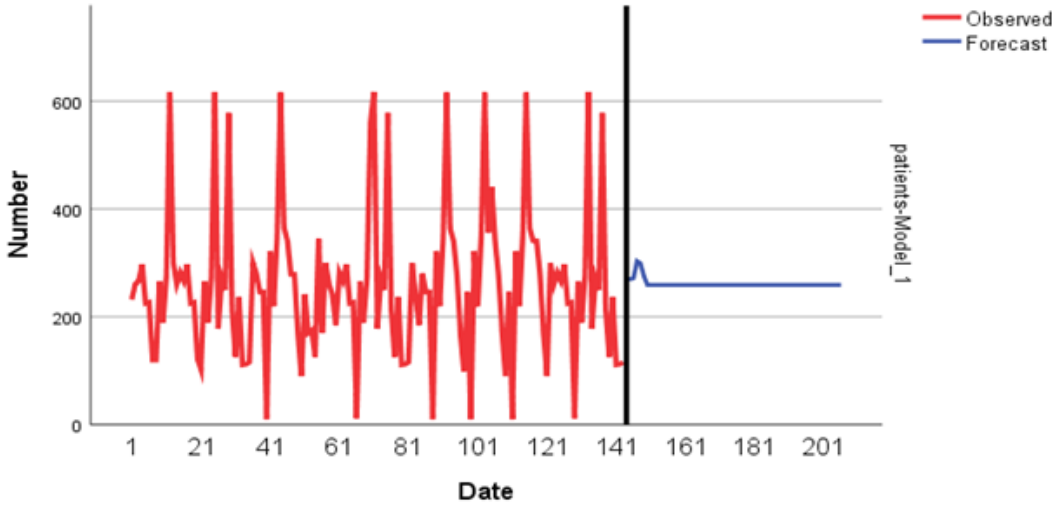

Figure 5: Predicted value of series time value with months

Table 1: ARIMA Model Parameters

					Estimate	SE	t	Sig.
patients-Model_1	patients	No Transformation	Constant		254.759	30.169	8.444	.000
			AR	Lag 1	.191	.085	2.256	.026
				Lag 2	.078	.085	.910	.364
	month	No Transformation	Numerator	Lag 0	.040	.363	.110	.913

Table 2: Forecasting value of cardiac patients number

Period	Forecast	Limits Lower	Limits Upper
144	177	-76-	431
145	270	17	524
146	270	16	524
147	272	18	525
148	304	50	557
149	300	46	553
150	276	22	529
151	259	-4-	522
152	259	-4-	522
153	259	-4-	522
154	259	-4-	522
155	259	-4-	522
156	259	-4-	522
153	259	-4-	522
154	259	-4-	522
155	259	-4-	522
156	259	-4-	522
157	259	-4-	522
158	259	-4-	522
159	259	-4-	522
160	259	-4-	522
161	259	-4-	522
162	259	-4-	522
163	259	-4-	522
164	259	-4-	522
165	259	-4-	522
166	259	-4-	522
165	259	-4-	522
166	259	-4-	522
167	259	-4-	522
168	259	-4-	522

Conclusion: Note through the study an instability in the number of people with heart disease due to some important factors, including smoking, nutrition and genetic diseases. Time series stability achieved using the Box Jenkins methodology model after taking the first difference for the data. We found that the appropriate model for the data is an autoregressive model from the validity of the model diagnosed through statistical tests, including autoregressive residual analysis

References

[1] Cardiovascular Disease Foundation. *What is Cardiovascular Disease?*, Available at`http://www.cvdf.org/`. Date Accessed: September 19, 2011.

[2] National Heart, Lung, and Blood Institute. *What Is Heart Disease?*, Available at `https://www.nhlbi.nih.gov/health/health-topics/topics/hdw/`. Date Accessed: September 19, 2011.

[3] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods* (2nd ed.). New York: Springer-Verlag,1991 .

[4] J. Crosbie and C. F. Sharpley, *DMITSA: A simplified interrupted time-series analysis program. Behavior Research Methods*, Instrum. Comput., 21(6) (1989) 639-642.

[5] Bowerman, L. Bruce , R. T. O'Connell and Anne B. Koehler, *Forecasting, Time Series, and Regression*, 4th ed. Belmont, CA: Thomson Brooks/Cole, 2005.

[6] R. l. Anderson, *Distribution of the series Analysis Correlation Coeffficient*, Ann, Mat. Statistic, 13(1942) 113-129 .

[7] R. Kaiser and A. Maravall , *Notes on Time Series Analysis ARIMA Models and Signal Extraction*, Benco de Esponaservicio Estudios , 2001.

[8] R. H. SHUMWAY,*Applied Statistical Time Series Analysis*, First Edition, prentice Hall New Jersey, USA, 1998, P.537.

[9] G.E.P Box and G.M. Jenkins, *Time Series Analysis Forecasting and control*, Holden Day, London , 1976 .

[10] R. Kaiser and A. Maravall,*Notes on Time series Analysis, ARIMA models and signal Extraction*, Banco de España. Servicio de Estudios, 2001.