The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Comparison Between SARIMA and SARIMAX Time Series Models with Application on Groundwater in Sulaymaniyah

Bakhan Hoshyar Qadir[1], Monem Aziz Mohammed[2]

[1,2]Department of statistic and informatics, College of Administration and Economics, University of Sulaimani, Sulaymaniyah, Iraq.

Email: hoshyarbaxa@yahoo.com[1], monem.mohammed@univsul.edu.iq[2]

Abstract:

Groundwater is one of the common essential water resources for billions of people, especially for many developing countries in Asia. Indeed, climate variation is one factor in the quantity and quality of groundwater resources in the world. The study used time-series data, it can be used to understand the past as well as predict the future.

Additionally, were taken climate index (Rainfall) to show and understand ground water level is affected by external factors and show the relationship between them. For this purpose, using groundwater level data during (7) years period contained 89 observation of data; beginning from (Jan 2013) through (May 2020) in the center of Sulaymaniyah city. Additionally, the climate variability data and groundwater data are monthly through a duration time.

The objective of this study Fitting the suitable Seasonal Autoregressive Integrated Moving Average (SARIMA) and Seasonal Autoregressive Integrated Moving Average with Explanatory variable (SARIMAX) model to studies the relationship between the climate variation and groundwater level. Finally, after added climate variability (Rainfall) to the SARIMA model, study showed the ground water level is affected by external factors. while, coefficient of external factor is positive and significant at %5 level of significant. This showed ARIMAX (0,1,0) x (1,0,1)12 with AIC (456.744) is a best model.

Keywords: SARIMA, SARIMAX, groundwater level, climate variability.

الملخص:

المياه الجوفية هي أحد موارد المياه الأساسية المشترك المياه الجوفية هي أحد موارد المياه الأساسية المشتركة لمليارات البشر، وخاصة للعديد من البلدان النامية في آسيا. في الواقع ، يعد تغير المناخ أحد العوامل في كمية ونوعية موارد المياه الجوفية في العالم. استخدمت الدراسة بيانات السلاسل الزمنية ، يمكن استخدامه لفهم الماضي وكذلك التنبؤ بالمستقبل .بالإضافة إلى ذلك تم أخذ مؤشر المناخ (هطول الأمطار) لإظهار وفهم تأثر منسوب المياه الجوفية بالعوامل الخارجية وإظهار العلاقة بينهما. لهذا الغرض، فإن استخدام بيانات مستوى المياه الجوفية خلال فترة (7) سنوات يحتوي على(89) ملاحظة للبيانات. بداية من (يناير 2013) حتى (مايو 2020) وسط مدينة السليمانية. بالإضافة إلى ذلك ، فإن بيانات تقلبات المناخ وبيانات المياه الجوفية الشهرية خلال فترة زمنية .الهدف من هذه الدراسة ملاءمة النموزج الخاص بالمتوسط المتحرك الموسمي المتكامل الانحدار الذاتي (SARIMA) والنموزج الخاص بالمتوسط المتحرك الانحدار الذاتي الموسمي المتكامل مع المتغير التوضيحي (SARIMAX) لدراسة العلاقة بين التغير المناخي ومستوى المياه الجوفية. أخيرًا ، بعد إضافة عامل تقلبات المناخ (هطول الأمطار) إلى نموذج SARIMA ، أظهرت الدراسة أن

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

مستوى المياه الجوفية تتأثر بالعوامل الخارجية. في حين أن معامل العامل الخارجي موجب ومعنوي عند مستوى 5٪ من المعنوية. أظهر أن(0 ARIMAX ،1،1) × (0،0، 12 (1هو الافضل مع قيمة (456.744) AIC.

الكلمات المفتاحية: النموزج الخاص بالمتوسط المتحرك الموسمى المتكامل الأنحدار الذاتي، النموزج الخاص بالمتوسط المتحرك الانحدار الذاتي الموسمي المتكامل مع المتغير التوضيحي، مستوى المياه الجوفية، التغير المناخ.

پوختە:

ئاوی ژێر زەوی یەکێکە لەسەرچاوە ئاوییە بنەڕەتییە باوەکان بۆ ملیارەها کەس ، بەتایبەتی بۆ زۆرێک لەوڵاتانی پێشکەوتووی ئاسیا. بێگومان گۆڕانی کەشوهەوا فاکتەرێکە لە چەندێتی و چۆنیەتی ی سەرچاوەکانی ئاوی ژێر زەوی لە جیهاندا . توێژینەوەکە داتای زنجیرەی کاتی بەکارهێناوە، دەتوانرێت بۆ تێگەیشتن لە ڕابردوو و هەروەها پێشبینی کردنی داهاتوو بەکاربهێنرێت

هەروەها ئیندێکسی کەش و هەوا (باران) وەرگیرابۆ پیشاندان و تێگەیشتن لە ئاستی ئاوی ژێر زەوی، کاریگەری فاکتەری دەرەکی و نیشاندانی پەیوەندی نێوانیان. بۆ ئەم مەبەستەش بەکارهێنانی داتای ئاستی ئاوی ژێر زەوی لەماوەی (7) ساڵدا 89 چاودێری داتای لەخۆگرتوە ؛ سەرەتا لە (ڕێبەندانی2013) هەوە بۆ (جۆزەردانی 2020) لە ناوەندی شاری سلێمانی . لەگەڵ ئەوەش، داتای جیاوازی ئاووهەوا و داتای مانگانە ی ئاوی ژێر زەوی لە ماوەی ئەو ماوەدایە.

ئامانجی ئەم لێکۆڵینەوەیە هەڵبژاردنی گونجاوترین مۆدێلی زنجیرەکانی کاتی ساریماو (SARIMA) زنجیرەکانی کاتی ساریماکسە (SARIMAX) بۆ لێکۆڵینەوە لە پەیوەندی نێوان گۆڕاوی ئاووهەوا و ئاستی ئاوی ژێر زەوی. دواجار دوای ئەوەی کە گۆڕانکاری کەشوهەوای زیادکراو (بارانبارین) بۆ سەر مۆدێلی SARIMA ، لێکۆڵینەوەکە نیشانی دا ئاستی ئاوی ژێر زەوی کاریگەری فاکتەری دەرەکی هەیە . لەکاتێکدا، هاوکۆلکەی دەرەکی و ئەرێنی و بەرچاوە لە ئاستی 5٪ ی بەرچاو. نموزەجی SARIMAX (0,1,0)x(1,0,1)12 لەگەڵ AIC (456.744) باشترین مۆدێلە .

کلیلە وشە: زنجیرەکانی کارتی ساریما، زنجیرەکانی کارتی ساریماکس، ئاستی ئاوی زەوی، گۆڕاوی ئاو و هەوا.

1. **Introduction:**

Gound water is an almost universal source of generally high-quality freshwater. These characteristics promote its overall development, scaled and localized to demand, obviating substantial infrastructure needs. Globally, groundwater is the cause of one-third of all freshwater withdrawals, supplying an estimated 36%, 42%, and 27% of the water used for domestic, agricultural, and industrial purposes respectively. In many environments, natural groundwater discharges sustain baseflow to rivers, lakes, and wetlands during low or no rainfall periods. In addition, the statistical tools could be analyzed these problems, especially when time is a significant factor in them. Time series analysis is one of the powerful statistical tools used to forecast the groundwater level and study the relationship between climate variation and groundwater level.

1.1 Objective of the Study

Fitting the suitable Seasonal Autoregressive Integrated Moving Average with Explanatory (SARIMAX) model to studies the relationship between the climate variation and groundwater level and compare two methods. These are SARIMA and SARIMAX to choose high accuracy of the forecasting model. can be used this model to develop procedures for forecasting groundwater levels.

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

2. Literature Review

(Arunraj, N. S., et al., 2016) In the study, develop a Seasonal Autoregressive Integrated Moving Average with external variables (SARIMAX) model which tries to account for all the effects due to the demand influencing factors, to forecast the daily sales of perishable foods in a retail store. Concerning performance measures, it is found that the proposed SARIMAX model improves the traditional Seasonal Autoregressive Integrated Moving Average (SARIMA) model. [3]

(Chiu, L. Y., et al., 2019) This study aims to improve a model for predicting the possible increase in the whitefly population in greenhouses applying autoregressive integrated moving average (ARIMA) and ARIMA with exogenous variables (ARIMAX). The proposed ARIMAX model can be used to assist farmers in decision-making for pesticide application scheduling. This research proves that the best model for predicting the incidence of whiteflies was ARIMAX, with temperature and humidity included as the exogenous factors. [8]

(Farih, L. N., et al., 2019) This thesis aims to estimate the total departure of ship customers in the main port of Makassar working the ARIMAX method with the effects of calendar variations. Moreover, the ARIMAX method is a system that can be used when there are exogenous variables, where in this example, the exogenous variable is in the form of a variable dummy which is the Eid holidays. Finally, these forecasting outcomes show that the ARIMAX (2,1,0) (0,0,1)12 method has a relatively small accuracy with the MAPE value of 14.08%. [11]

(Ling, A. S. C., et al., 2019) This study proposes to develop an Autoregressive Integrated Moving Average with external variables (ARIMAX) model, which tries to account for the effects due to the climatic influencing factors, to forecast the weekly cocoa black pod disease incidence. A study found that The ARIMAX models performed well with lower error as compared to the ARIMA models. Key findings indicate that maximum temperature and relative humidity have a significant correlation with black pod incidence and are suggested as indicators in forecasting the cocoa black pod incidence. [15]

(Tadesse, K. B., & Dinka, M. O., 2017) In this study, the Waterval River flood was forecasted by the SARIMA model. Monthly flows from 1960 to 2016 were done for modeling and forecasting. Based on seasonally differenced correlogram characteristics, many SARIMA models were evaluated. Their parameters were optimized, and a diagnostic checkup of estimates was made applying white noise and heteroscedasticity tests. Lastly, based on minimum Akaike Information (AI) and Hannan–Quinn (HQ) criteria, SARIMA (3, 0, 2) x (3, 1, 3)12 model chosen for Waterval River flow forecasting. [18]

(Gikungu, S. W., et al., 2015) In this research, Seasonal Autoregressive Integrated Moving Average (SARIMA) model is developed to forecast Kenya's inflation measure using quarterly data from 1981 to 2013 obtained from KNBS. SARIMA (0,1,0) (0,0,1)4 was identified as the most suitable model. The predictive ability tests RMSE=0.2871, MAPE=3.9456, and MAE= 0.2369 showed that the model was fitting for forecasting the inflation rate in Kenya.[12]

(Cools, M. et al., 2009) In this study, daily traffic counts are explained and forecasted by different modeling philosophies, namely the ARIMAX and SARIMA(X) modeling approaches. Particular emphasis is put on investigating the seasonality in the daily traffic data and on the identification and

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

comparison of holiday effects at different site locations. Finally, results revealed that the ARIMAX and SARIMAX modeling approaches are practical frameworks for identifying and quantifying possible influencing effects..[9]

3. Materials and Methods

3.1 Time series analysis

Time series analysis is done for mainly two reasons:

1. Understanding the behavior of the process by studying its records to model and identify the main parameters that influence the time series and identify its components.
2. Forecasting the future values of the series using an adequate model that has been trained on past values.

The initial action in the analysis of each time series is to plot the data. If there are apparent discontinuities in the series, such as an immediate change of level, it may be advisable to investigate the series by first breaking it into homogeneous segments. If there are external observations, they should be studied carefully to check whether there is any justification for discarding them (for instance, if an observation has been recorded of any other process by blunder) [13],[17].

3.1.1 time series

A time series is a collection of measures introduced sequentially into time. It is mathematically defined as a collection of vectors in which the index parameter (T) is the time-space or a set of observations (indexed by time). The variable treat as a random variable.

These observations can be as different as numbers, labels, colors, and many others. Furthermore, these measurements may be made continuously within time or be taken at a discrete set of time duration. By convention, these two kinds of series are named continuous and discrete-time series, respectively, even though the regular variable may be discrete or continuous in each case [1],[7].

3.1.2 components of a time series

The signal in time series data usually is divided into four components: Trend, seasonal, cyclical, and irregular. Each of these components describes a different mechanism by which past values of a time series may be related to the present value [4].

discuss each of these components:

1. Trend (T)
2. Seasonal Variation (S)
3. Cyclical Variation (C)
4. Irregular Variation (I)

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

3.1.3 Attribute of Time Series

- **Stationary Time Series**

The basis for any time series analysis is stationary time series. It is essential for stationary time series that we can develop models and forecasts. However, it is the nonstationary time series that is most interesting in many applications, especially in business and economics. Similarly, when processes are left alone in industrial applications, they are expected to show nonstationary behavior simply following the second law of thermodynamics. Therefore, while in real-life applications, it happens only under specific situations, the stationary time series play a vital role as the foundation for time series analysis [5].

- **Non-Stationary Time Series**

They could have nonconstant means μ, time-varying second such as nonconstant variance σ^2, or both properties.

Many applications with nonstationary data use different methods (d) from Non-Stationary to Stationary process as follows.

$\nabla Y_t = Y_t - Y_{t-1}$. If that is the case, we can then model the changes, make forecasts about the future values of these changes, and build models and create forecasts of the original nonstationary time series [19].

By applying one of (lnY, $\sqrt{Y}$) methods, the series is converted from nonstationary around variance to a stationary around variance. And by taking differences the series is converted form non stationary around mean to a stationary around mean.

3.4 TESTS FOR NONSTATIONARITY

There are objective tests that may be conducted to determine whether a series is nonstationary. The series could be nonstationary because of random walk, drift, or trend. In order to test for non-stationary, the Augmented Dickey-Fuller (ADF) test is used where it test for a unit root in a time series sample. Given [20]

$$\Delta Y_t = \beta_0 + \alpha_t + \beta_1 Y_{t-1} + \sum_{i=1}^{p} \lambda_i \Delta Y_{t-i} + \varepsilon_t \qquad \text{.....(3.1)}$$

Where a random walk, $\alpha_t = \alpha_{t-1} + \alpha\varepsilon_t$ is allowed.

3.5 Box-Jenkins Models for Forecasting

Box and Jenkins popularized a three-stage method to select an appropriate model to estimate and forecast a univariate time series.

A measurement of the sample (ACF) and (PACF) to those of different theoretical (ARMA) processes may suggest several plausible models. Then, in the estimation stage, each of the tentative

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

models is fit, and the various($\emptyset_p$ and θ_q) coefficients are examined. In this next stage, the goal is to select a stationary and parsimonious model with a good fit. Finally, the third stage involves diagnostic checking to ensure that the residuals from the estimated model mimic a white-noise process [10].

3.6 Autocorrelation Function (ACF)

Weakly stationary time series, Y_t, has finite variance with constant mean and variance over time t. Hence we can write

$$E(Y_t) = \mu_t = \mu$$

Similarly, the sample variance can be calculated using

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2 \qquad \text{……..(3.2)}$$

Furthermore, their correlation is their covariances scaled for their standard deviations.

$$Corr(Y_t, Y_{t\mp k}) = \frac{Cov(Y_t, Y_{t\mp k})}{\sqrt{\sigma^2_{Y_t}}\sqrt{\sigma^2_{Y_{t\mp k}}}} \qquad \text{…….(3.3)}$$

We can find the covariance between observations of k lags apart from $Cov(Y_{t\mp k}, Y_t) = E[(Y_{t\mp k} - \mu)(Y_t - \mu)]$, also called autocovariance since we are dealing with the same data set. Once again, owing to stationarity we have $Cov(Y_{1\mp k}, Y_t) = Cov(Y_{2\mp k}, Y_t)$, getting the autocovariance only a function of the time lag k. Therefore, we describe the autocovariance function as

$$\gamma(k) = E[(Y_{t\mp k} - \mu)(Y_t - \mu)] \qquad \text{……..(3.4)}$$

Note that the variance of the time series is γ(0).

It is an important measurement to analyze time-series observations for correlation observations between the series at different times. We denoted by: ρ_k, (k =0, ±1, ±2, …), which depend only on the lag of k.

The (ACF) plays a very crucial role in the description of time series models as it summarizes as a function of (k) whereby correlated the observations that are (k) lags apart are. Of course, we cannot know the actual value of (ACF) in actual life, but instead, we will consider it from the data at hand working [5],[6].

$$\hat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(Y_{t\mp k} - \bar{Y})(Y_t - \bar{Y}) \qquad \text{…….(3.5)}$$

And

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} \qquad \text{…….(3.6)}$$

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Now, let us have (n) observations such that $(Y_1, Y_2, \dots, Y_n)$, therefore the covariance matrix (Γ_n) is:

$$\therefore \Gamma_n = \sigma_y^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma_y^2 \rho_n$$

Or $\rho_n = \frac{\Gamma_n}{\sigma_y^2} = \frac{\Gamma_n}{\Gamma_0}, \quad (n = 0,1,2, \dots)$

In general, we have $$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\gamma_0} = \frac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t\mp k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \qquad \text{.....(3.7)}$$

3.7 Partial Autocorrelation Function (PACF)

A Partial Autocorrelation Function is a tool that exploits the fact that, whereas an AR(p) process has an autocorrelation function that is infinite in extent, the partial autocorrelations are zero beyond lag p.

The partial autocorrelations can be described in terms of p nonzero functions of the autocorrelations. Denote by (ϕ_{kk}) the $k'th$ coefficient in an autoregressive representation of order (k), so that ($\emptyset_{kk}$) is the last coefficient [6].

$$\rho_k = \emptyset_{k1}\rho_{k-1} + \cdots + \emptyset_{k(k-1)}\rho_1 + \emptyset_{kk}\rho_0 \qquad \text{.....(3.8)}$$

which may be written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \emptyset_{k1} \\ \emptyset_{k2} \\ \emptyset_{k3} \\ \vdots \\ \emptyset_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_k \end{bmatrix}$$

Or

$$\Gamma_k \emptyset_k = \rho_k$$

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

3.8 Autoregressive Model

A stochastic model that can be very useful in representing specific practically occurring series is the autoregressive model. In this model, the popular value of the process is expressed as a finite, linear aggregate of previous values of the process and a random shock a_t. Let us denote the values of a process at equally spaced times [6] $t, t-1, t-2, \ldots$ by $y_t, y_{t-1}, y_{t-2}, \ldots$

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + a_t \qquad \text{.....(3.9)}$$

If we define an autoregressive operator of order p in terms of the backward shift operator (B) by

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \qquad \text{.....(3.10)}$$

3.9 Moving Average Model

The autoregressive model expresses the variation (y_t) of the process as a finite weighted sum of (p) previous deviations $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ of the process, plus a random shock (a_t) . Here we take (y_t) , linearly dependent on a finite number (q) of previous (a's). Thus,

$$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad \text{.....(3.11)}$$

Is called a moving average (MA) process of order (q). The name ''moving average'' is somewhat misleading because the weights 1, $-\theta_1, -\theta_2, \ldots, -\theta_q$, which multiply the (a's), need not total unity nor need they be positive.

If we define a moving average operator of order (q) by

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ the polynomial function of order (q) in B

the moving average model may be written economically as

$$y_t = \theta(B) a_t$$

It contains (q+2) unknown parameters ($\mu, \theta_1, \ldots, \theta_q, \sigma_a^2$), which in practice have to be estimated from the data [5],[6].

3.10 Mixed Autoregressive--Moving Average Models (ARMA)

To achieve higher flexibility in fitting actual time series, it is sometimes advantageous to include both autoregressive also moving average terms in the model. This leads to the mixed autoregressive--moving average (ARMA) model:

$$y_t = \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad \text{.....(3.12)}$$

Or $\emptyset(B) y_t = \theta(B) a_t$

is called the mixed autoregressive--moving average process of order (p, q), which we abbreviate as ARMA (p, q).

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Now writing

$$y_t = \emptyset^{-1}(B)\theta(B)a_t$$

$$= \frac{\theta(B)}{\emptyset(B)}a_t = \frac{1-\theta_1 B-\cdots-\theta_q B^q}{1-\emptyset_1 B-\cdots-\emptyset_p B^p}$$

The model employs ($p + q + 2$) unknown parameters μ, ϕ_1, … , ϕ_p, θ_1, … , θ_q, σ^2_a, that are estimated from the data. This model may also be written in the form of the linear filter as $y_t = \phi^{-1}(B)\theta(B)a_t = \psi(B)a_t$, with $\psi(B) = \phi^{-1}(B)\theta(B)$ [5],[6].

3.11 Non-Seasonal Autoregressive Integrated Moving Average Model

If we join differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for Auto Regressive Integrated Moving Average (in this context, "integration" is the reverse of differencing). The ARIMA class of models is a crucial forecasting tool and is the basis of many fundamental ideas in time-series analysis. The full model can be written as [7],[14]:

$$y'_t = c + \emptyset_1 y'_{t-1} + \cdots + \emptyset_p y'_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad \text{……..(3.13)}$$

Where y'_t is the differenced series (it may have been differenced more than once).

$$y'_t = y_t - y_{t-1} \qquad \text{……..(3.14)}$$

Equation above can be written in backshift notation as :

$$\left(1-\emptyset_1 B-\cdots-\emptyset_p B^p\right)(1-B)^d Y_t = c + (1+\theta_1 B+\cdots+\theta_q B^q) \qquad \text{…….(3.15)}$$

3.12 Seasonal Autoregressive Integrated Moving Average Model (SARIMA)

A seasonal ARIMA model is formed by adding additional seasonal terms in the ARIMA models we have seen so far. It is written as follows:

SARIMA (p, d, q) $(P,D,Q)_m$

(p, d, q) : non seasonal part of the model

$(P,D,Q)_m$: seasonal part of the model

The seasonal ARIMA(p, d, q)(P, D, Q)s model can be most succinctly expressed applying the backward shift operator:

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^d Y_t = \Theta_Q(B^s)\theta_q(B)a_t \qquad \text{……..(3.16)}$$

where Φ_P , ϕ_p, Θ_Q, and θ_q are polynomials of orders P, p, Q, and q, respectively. For stationarity to exist, both the regular and the seasonal autoregressive parameters need to lie within the bounds of stationarity. That is,

$-1 < \Phi_P , \phi_p < +1$

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Autoregressive processes whose parameter estimates remain within these bounds are invertible [16],[20].

3.13 ARIMAX

Autoregressive Integrated Moving Average with external variables (ARIMAX) model can be observed as a multiple regression model with one or more autoregressive (AR) terms and/or one or more moving average (MA) terms.

The general ARIMAX models are as follows:

$$Y_t = \beta X_t + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \quad \text{……..(3.17)}$$

The general form of ARIMAX (p, d, q) model for one explanatory variable has the following condensed form in lag operator notation

$$Y_t = \beta X_t + \emptyset(B)^{-1}\theta(B)a_t \quad \text{……..(3.18)}$$

The model can also be written as:

$$Y_t = \frac{\beta}{\emptyset(B)} X_t + \frac{\theta(B)}{\emptyset(B)} a_t$$

For more than one explanatory variable, the mathematical form of ARIMAX model has the form:

$$Y_t = \beta X_t + \beta_1 X_{1,t} + \cdots + \beta_j X_{j,t} + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$
……(3.19)

The first step in building an (ARIMAX) model consists of identifying a suitable (ARIMA) model for the endogenous variable [2],[15].

3.14 SARIMAX model

(SARIMAX) the structure is a popular regression model type for time series forecasting, which is linear in the parameters, enabling linear regression techniques for estimating those parameters. (SARIMAX) type models fall within the category of multivariate regression models.

The (SARIMAX) model is a (SARIMA) model with external variables.

The general SARIMAX model equation:

$$Y_t = \beta X_t + \beta_1 X_{1,t} + \cdots + \beta_j X_{j,t} + \left(\frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D} a_t\right) \quad \text{…..……(3.20)}$$

where $\emptyset_1$, $\emptyset_2$, …. $\emptyset_p$, Φ_1 , Φ_2 ,…, Φ_P , θ_1 , θ_2, …. θ_q and Θ_1 , Θ_2, …. Θ_Q are the weights for the non-seasonal and seasonal autoregressive terms and moving average terms [3],[9].

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

4.1 Data Description:

An **SARMAX model** is an S**ARIMA model with an exogenous variable, for this purpose, we should collected two part of data to analysis SARIMAX model forecasting:**

- Monthly ground water level data from a well was elevation (880m) above sea level and depth (58m), to analysis SARIMA model. We collected the data from the Directorate of Sulaymaniyah Water.
- Monthly Rainfall data, was collected from the Sulaymaniyah Directorate of Meteorology and Seismology, climate variability to add exogenous variable to SARIMA model and build SARIMAX time series model

The data during (7) years period contained 89 observation of data; beginning from (Jan 2013) through (May 2020) in the center of Sulaymaniyah city. the climate variability data and groundwater data are monthly through a duration time.

4.2 Applications:

This study can be done by using SARIMAX and SARIMA time series model. The first step build SARIMA model.

- **SARIMA model:**

The time series are display observations on the y-axis against equally spaced time intervals on the x-axis. They are used to evaluate patterns, knowledge of the general trend, and behaviors in data over time. The time series plot of monthly Groundwater level in Sulaymaniyah city is displayed in figure 1 below:

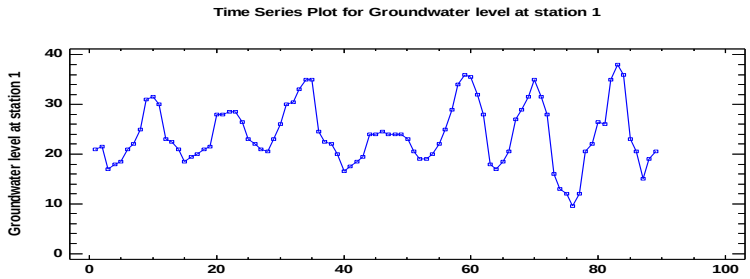

Figure 1: Monthly plot of time series Groundwater level in Sulaymaniyah city

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 1: shows the results of ADF of the data of the time series of Groundwater level

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.972437	0.1459
Test critical values:	1% level	-4.065702	
	5% level	-3.461686	
	10% level	-3.157121	

Table (1) explain that the p-value of the Dickey-Fuller test equals (0.1459)and it is greater than (0.05). This result indicates that the data of the time series of monthly Groundwater level is not random and demonstrates these results by examining the autocorrelation and partial autocorrelation functions as shown below.

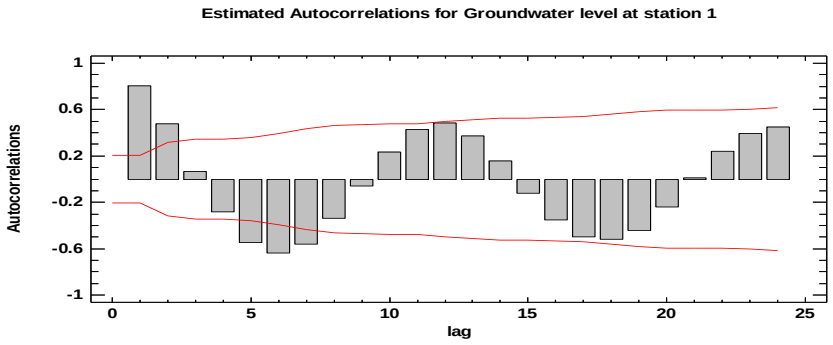

Figure 2: Autocorrelations function for the monthly Groundwater level in Sulaymaniyah city

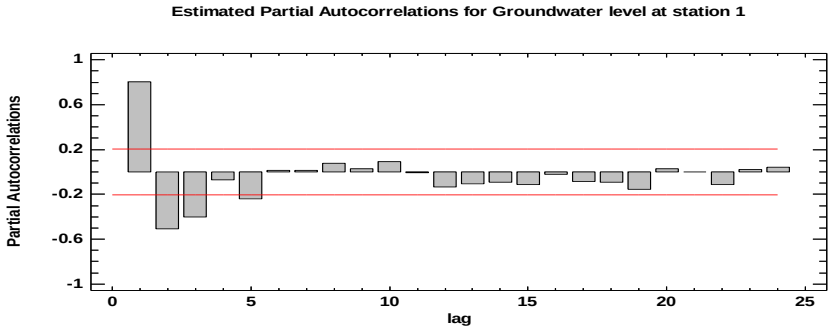

Figure 3: Partial Autocorrelations function for the monthly Groundwater level in Sulaymaniyah city

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

All the above results and plots support that the time series data is not random at the level, which needs to be transformed to a random series. Therefore, we used many transformations, and we found that the most suitable transformation is by differencing the series. We note that the time series for the first differenced series in figure 4 indicates that the series is stationary.

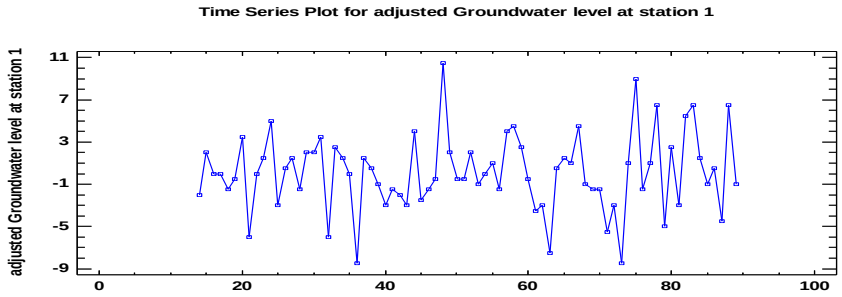

Figure 4: Time series plot of the first difference of monthly Groundwater level in Sulaymaniyah city

Table 2: shows the results of ADF of the data of the time series of Groundwater level

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.943665	0.0006
Test critical values:	1% level	-4.066981	
	5% level	-3.462292	
	10% level	-3.157475	

Table (1) explain that the p-value of the Dickey-Fuller test equals (0.0006) and it is less than (0.05). This result indicates that the data of the time series of monthly Groundwater level is random.

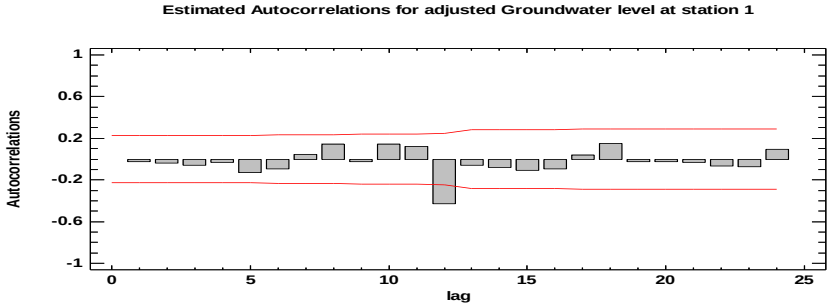

Figure 5 : Autocorrelation Function for the first – differenced series of the monthly Groundwater level in Sulaymaniyah city

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

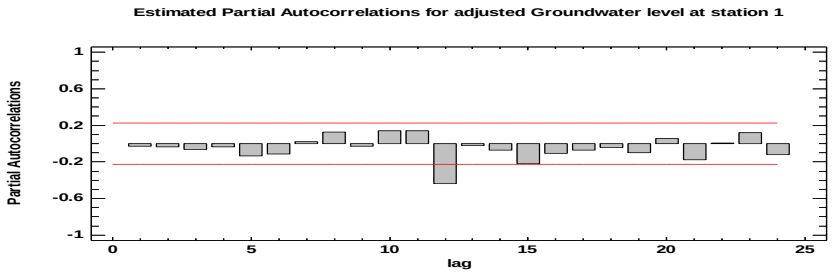

Figure 6 : Partial Autocorrelation Function for the first – differenced series of the monthly Groundwater level in Sulaymaniyah city

The results above demonstrate the success of differencing the time series data of the monthly Groundwater level at the Directorate of Sulaymaniyah Water. Thus, the series becomes stationary.

4.3 Model Identification

We use the ACF and PACF plots to identify the order of the ARIMA model. The plots of ACF and PACF for the first differencing order of log for the monthly Groundwater level are shown in Figure 1. For the first difference log series and seasonal order difference, the ACF cut-off is lag 1 and PACF cut-off is lag 1.

Table 3 : SARIMA Models Criteria for the monthly Groundwater level

Model	AIC
ARIMA(0,1,0)x(1,0,1)12	1.94015
ARIMA(1,1,0)x(1,0,1)12	1.9674
ARIMA(0,1,1)x(1,0,1)12	1.96832
ARIMA(1,1,1)x(1,0,1)12	1.99995
ARIMA(1,0,0)x(0,1,1)12	2.06346

The performance of seasonal-ARIMA models is shown in Table 3. We found that ARIMA (0,1,0) (1,0,1)12 has the smallest value of AIC (1.94015) among all the other models that shows the best performance for a prediction that can be obtained for the monthly groundwater level at the Directorate of Sulaymaniyah Water.

4.4 Parameters Estimation:

Since we concluded in the previous section that the SARIMA (0,1,0)x(1,0,1)12 model is the best model with the smallest value of AIC criteria, the parameters had been estimated using maximum likelihood it is the best and most appropriate method of estimation. The results of the parameter estimation of the model are shown in table (4) below.

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 4 : Parameter estimation of SARIMA (0,1,0)x(1,0,1)12 Model Estimate model coefficients

Parameter	Estimate	Stand. Error	T	P-value
SAR(1)	1.2764	0.0464593	27.4735	0.000000
SMA(1)	1.23016	0.0685782	17.9381	0.000000

It is shown in table (4) that the p-value for the parameters SAR (1) and SMA (1) coefficients are less than $\alpha = 0.05$. As it is show for this model, the AIC criteria are the smallest values among the other models. Thus, the final model is SARIMA (0,1,0) x (1,0,1)12

4.5 Forecasting

After getting the final model **SARIMA (0,1,0) x (1,0,1)12** of the data of the monthly ground water level at the Directorate of Sulaymaniyah Water that has been expressed above, the researcher used it for forecasting future ground water level.

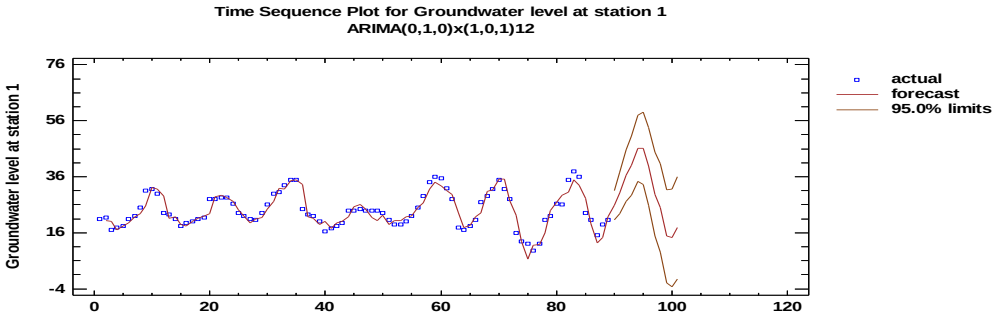

Figure 7: Plot of the data and the forecasts with 95% confidence interval are represented

Figure 7 shows the result that the behavior of forecasted values is the same as original series of ground water level at the Directorate of Sulaymaniyah Water. The result of the forecasted values in table (3) for the year 2020-2021 for 12 months are all between the upper and lower boundaries of the 95% confidence intervals. This confirms that the forecasting is very efficient.

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 5: Forecast future value with the lower and upper 95% confidence interval

Period	Forecast	Lower 95.0% Limit	Upper 95.0% Limit
Jun-2020	25.8687	20.6074	31.1299
Jul-2020	30.2368	22.7963	37.6773
Aug-2020	36.4517	27.339	45.5644
Sep-2020	40.0459	29.5234	50.5683
Oct-2020	46.085	34.3205	57.8494
Nov-2020	46.0992	33.2119	58.9865
Dec-2020	39.7645	25.8446	53.6844
Jan-2021	29.7996	14.9186	44.6806
Feb-2021	24.9871	9.20346	40.7708
Mar-2021	14.7788	-1.85864	31.4163
Apr-2021	14.2539	-3.19558	31.7035
May-2021	17.7817	-0.443729	36.0071

Table 3 shows that the quantities of monthly ground water level at the Directorate of Sulaymaniyah Water in 2020 – 2021 for 12 months have been forecasted. It is also shown from these results that the forecasted values are all between the upper and lower boundaries of the 95% confidence intervals. This supports that the forecasting is efficient.

- **SARIMAX model**

After getting the final model **SARIMA (0,1,0) x (1,0,1)12** of the data of the monthly ground water level at the Directorate of Sulaymaniyah Water we should consider the influencing external factors such as, (rainfall) is displayed in figure 8 below

DOI: http://dx.doi.org/10.25098/5.2.17

The Scientific Journal of Cihan University – Sulaimaniya **PP: 30-48**
Volume (5), Issue (2), December 2021
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

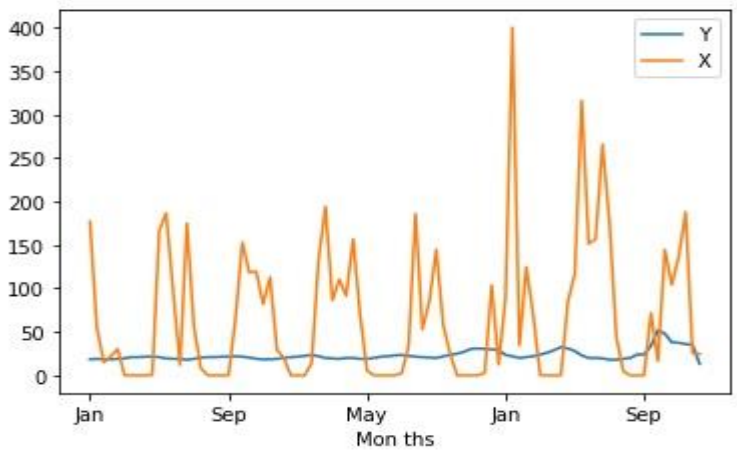

Figure 8: The monthly average rainfall(mm) recorded

Table 6: SARIMAX Results

SARIMAX Results			
Dep. Variable:	Y	**No. Observations:**	89
Model:	SARIMAX(0, 1, 0)x(1, 0, [1], 12)	**Log Likelihood**	-224.372
Date:	Sat, 26 Jun 2021	**AIC**	456.744
Time:	15:29:02	**BIC**	466.654
Sample:	0	**HQIC**	460.737

Table 7: fitted SARIMAX models

	coef	std err	z	P>|z|
X1	0.0059	0.002	2.461	0.014
ar.S.L12	0.9159	0.076	12.116	0.000
ma.S.L12	-0.5605	0.158	-3.546	0.000
sigma2	8.6498	1.344	6.437	0.000

Table (4) and (5) shows Rainfall an important factor in effecting the groundwater level while, coefficient of Rainfall is positive (0.0059) are significant at %5 level of significant. This showed

SARIMAX (0,1,0) x (1,0,1)12 with AIC (456.744) is a best model and all of coefficient are significant at %5 level of significant.

DOI: http://dx.doi.org/10.25098/5.2.17

5. Conclusion and Recommendations

5.1 Conclusion

This research studies the relationship between the climate index and the groundwater level of the Sulaymaniyah city, in order to forecast the groundwater level in the studied area by using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Seasonal Autoregressive Integrated Moving Average with Explanatory (SARIMAX). Add climate indices (rainfall) were used, along with the groundwater level data from station during the period 2013–2020 to develop the forecast model and verify it with the data of 2021. the first step before built the suitable model is to check stationary for data by using Augmented Dicky-Fuller Test (ADF Test). After that Identification of AR and MA terms requires the model builder to examine the autocorrelation coefficient function (ACF) and the partial autocorrelation coefficient function (PACF). The possible model was then selected using AIC statistics. Diagnostic Checking was done to consider the white noise characteristic of estimated residuals by using the statistics of Box and Ljung (Q-statistic). The simulated results of the monthly groundwater level in 2021 of the wells have a confidence interval of around 95%. To conclude, the results show that there is a relationship between the groundwater level and the climate index. while, coefficient of SARIMAX (0,1,0) x (1,0,1)12 are significant at %5 level of significant.

5.2 Recommendations:

can be used this model to develop procedures for forecasting groundwater levels, which can then be used to better manage the groundwater resources in my country.

6. References:

[1] Adhikari, R., & Agrawal, R. K. (2013). An introductory study on time series modeling and forecasting. arXiv preprint arXiv:1302.6613.

[2]Andrews, B. H., Dean, M. D., Swain, R., & Cole, C. (2013). Building ARIMA and ARIMAX models for predicting long-term disability benefit application rates in the public/private sectors. Society of Actuaries, 1-54.

[3]Arunraj, N. S., Ahrens, D., & Fernandes, M. (2016). Application of SARIMAX model to forecast daily sales in food retail industry. International Journal of Operations Research and Information Systems (IJORIS), 7(2), 1-21.

[4]Becketti, S. (2013). Introduction to time series using Stata (Vol. 4905). College Station, TX: Stata Press.

[5] Bisgaard, S., & Kulahci, M. (2011). Time series analysis and forecasting by example. John Wiley & Sons.

[6] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

[7] Chatfield, C. (2000). Time-series forecasting. CRC press.

DOI: http://dx.doi.org/10.25098/5.2.17

[8] Chiu, L. Y., Rustia, D. J. A., Lu, C. Y., & Lin, T. T. (2019). Modelling and Forecasting of Greenhouse Whitefly Incidence Using Time-Series and ARIMAX Analysis. IFAC-PapersOnLine, 52(30), 196-201.

[9]Cools, M., Moons, E., & Wets, G. (2009). Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations. Transportation research record, 2136(1), 57-66.

[10]Enders, W. (2015) Applied Econometric Time Series. John Wiley & Sons.

[11]Farih, L. N., Fauziah, I., & Wijaya, M. Y. (2019). Prediction of The Number of Ship Passengers in The Port of Makassar using ARIMAX Method in The Presence of Calender Variation. InPrime: Indonesian Journal of Pure and Applied Mathematics, 1(1).

[12] Gikungu, S. W., Waititu, A. G., & Kihoro, J. M. (2015). Forecasting inflation rate in Kenya using SARIMA model. American Journal of Theoretical and Applied Statistics, 4(1), 15-18.

[13] Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2003). Market response models: Econometric and time series analysis (Vol. 12). Springer Science & Business Media.

[14] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[15] Ling, A. S. C., Darmesah, G., Chong, K. P., & Ho, C. M. (2019). Application of ARIMAX Model to Forecast Weekly Cocoa Black Pod Disease Incidence. Available at: https://scholar.google.com

[16] Metcalfe, A. V., & Cowpertwait, P. S. (2009). Introductory time series with R (p. 2). Springer-Verlag New York.

[17] Pal, A., & Prakash, P. K. S. (2017). Practical time series analysis: master time series data processing, visualization, and modeling using python. Packt Publishing Ltd.

[18]Tadesse, K. B., & Dinka, M. O. (2017). Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa. Journal of water and land development, 35(1), 229-236.

[19] William W.S. Wei. (2006) Time series analysis Univariate and Multivariate Methods. Pearson, Addison Wesley.

[20]Yaffee, R. A., & McGee, M. (2000). An introduction to time series analysis and forecasting: with applications of SAS and SPSS. Elsevier.

DOI: http://dx.doi.org/10.25098/5.2.17