On Modelling seasonal ARIMA series: Comparison, Application and Forecast (Number of Injured in Road Accidents in Northeast Algeria)

FARIDA MERABET [A] , HALIM ZEGHDOUDI [B]
Department of Mathematic and Informatics, ENSET, Skikda 21000, ALGERIA
LaPS laboratory, Badji-Mokhtar University BP12, Annaba 23000, Alger, ALGERIA

Abstract: - This paper study and modeless a number of road accidental injuries in the region of Skikda (northeast Algeria) according to Box- Jenkins method using EViews software using series data from January 2001 to December 2016. Also, Kalman filter method is given. To this end, Kalman filter method is used for short term prediction and parametric identification purpose. The other side, a comparative study is given to compare between the two methods by de following criteria: Mean absolute percentage error (MAPE), root mean square percentage error (RMSPE) and the Theils's U statistic. This application used Eviews 5.0 and SPSS 26 software's.

Key-Words: - $(S)ARIMA$ models, Box-Jenkin method, Kalman filter, Forecasting

Received: January 5, 2020. Revised: April 13, 2020. Re-revised: May 2, 2020. Accepted: May 20, 2020. Published: June 4, 2020.

1 Introduction

The study of time series or chronological time corresponds to the statistical analysis of regularly spaced observations over time. It has been used in astronomy (in the periodicity of sunspots, 1906), meteorology (time series regression of sea level on weather, 1963) in signal theory ("Noise in FM receivers", 1963), in biology ("the autocorrelation curves of Schizophrenia brain waves and the power spectrum", 1960), in economics ("time series analysis of imports, exports and other economic variables", 1971). A chronological series is a series of observations generated sequentially over time, in addition to be continuous if all instances of observations are continuous, and be discrete if all these moments are discrete. Subsequently, only discrete time series or observations will be considered in equidistant time intervals. In this study, we have proposed a modelization seasonal series of the number of road accidents at Skikda region, according to the method $(S)ARIMA$ - Box-Jenkins. The second section presents a theoretical reminder of the integrated linear processes $ARIMA$, and the seasonal processes $ARIMA$ [5, 6, 10, 11], then we discussed the Box-Jenkins methodology [1, 5, 12] which is considered as one of the most commonly used ARMA process processing methods (especially under R, SAS, SPSS, EViews, ...) due to its simplicity in determining the appropriate $ARIMA$ model for modelization of time series [2, 9]. This method suggests four main steps: identification, estimation, validation and model prediction [1, 5, 12, 13]. Whilst, the third section was devoted to an empirical application of the models $SARIMA$ on a number of road accident injuries in Skikda region using the Box-Jenkins methodology providing a best model for this series, and hence we the appropriate model $SARIMA(12,1,1)\times(0,1,1)_{12}$ was found. Several previous authors (Ette Harrison Etuk et Nathaniel Ojekudo [8], Ette Harrison Etuk et Tariq Mahgoub Mohamed [9], Gerolimetto, M [11], Suhartono [15], Xiaosheng Li, Chunliu Ma, Haike Lei et Haixia Li [20], Xujun Zhang et Yuanyuan Pang [21]) have investigate the modilization of seasonal series by using $SARIMA$ process.

2 Problem Formulation

The idea of the problem is to present the $(S)ARIMA$ models, Box- Jenkins methodology, then give the presentation of the state space of these models to apply the Kalman filter.

2.1 Seasonal time series model

2.1.1 Stationary process

A process (X_t) is a stationary process of second order (weakly stationary), based on the check of the following conditions:

$i/\forall t \in \mathbb{Z}, IE(X_t)$

$ii/\forall t \in \mathbb{Z}, Var(X_t) = \sigma^2 = \gamma(0)$

$iii/\forall t \in \mathbb{Z}, \forall h \in \mathbb{Z}, Cov(X_t, X_{t-h}) = \gamma(h)(depends only to h)$

$\gamma(h)$ is autocovariance order h of (X_t).

2.1.2 White noise

(ε_t) is a strong white noise when if only if, (ε_t) are randomly distributed independent and identically variables (i.i.d). (ε_t) is a weak white noise process, if only and if the following conditions are fulfilled:

$$IE(\varepsilon_t) = 0, \forall t \in \mathbb{Z}$$

$$Var(\varepsilon_t) = \sigma^2, \forall t \in \mathbb{Z}$$

$$Cov(\varepsilon_t, \varepsilon_s) = 0, if\ t \neq s$$

2.1.3 Stationarity test

Overall, we naturally begins by the question on the stationarity of the study series, and we insist on the fact that the proposed models only allow to model stationary series. This test promotes to test the hypothesis of stationarity of series (absence of unit root) using two methods, as proposed by Dickey and Fuller (1979) and Philips and Person (1988).

2.1.4 Wold Theory

If (X_t) is centrilized and stationary process, thus we get the following decomposition:

$$X_t = \sum_{j=0}^{+\infty} \psi_j \varepsilon_{t-j} + V_t, t \in \mathbb{Z}$$

where :

$1.\psi_0 = 1$ et $\sum_{j=0}^{+\infty} \psi_j^2 < \infty$

$2.(\varepsilon_t)_{t\in\mathbb{Z}}$ is a white noise of $(X_t)_{t\in\mathbb{Z}}$

$3.(V_t)_{t\in\mathbb{Z}}$ is a deterministic process

$4.Cov(\varepsilon_t, V_t) = 0, \forall t, s \in \mathbb{Z}$

2.1.5 Linear series

A series $(X_t)_{t\in\mathbb{Z}}$ is called linear when it can be written as follows:

$$X_t = \mu + \sum_{j=-\infty}^{+\infty} \psi_j \varepsilon_{t-j}$$

Where $\varepsilon_t \sim NN(0, \sigma^2), \psi_0 = 1$ and the sequence (ψ_j) is absolutely summable, i.e $\sum_{j=-\infty}^{+\infty} |\psi_j| < \infty$. A series (X_t) is called linear and causal if:

$$X_t = \mu + \sum_{j=0}^{+\infty} \psi_j \varepsilon_{t-j}, \psi_j = 0 if\ j < 0$$

2.1.6 The autocorrelation function ACF and the partial autocorrelation PACF

1.The autocorrelation function (ACF) is the function denoted $\rho_X(h)$ measuring the correlation of the series with itself shitted by h periods:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

Values of $\rho_X(h)$ are values in the interval $[-1,1]$, and $\rho_X(0) = 1$.

2. The partial delay autocorrelation h is defined as as the partial correlation coefficient between X_t and X_{t-1}, i.e the correlation between X_t and X_{t-h}, the influence of the other variables shifted of h period ($X_{t-1}, X_{t-2}, ..., X_{t-h-1}$) having and removed.

2.2 $(S)ARIMA$ Processes

2.2.1 $ARMA$ Processes

A second order process $(X_t)_{t\in\mathbb{Z}}$ is defined as an $ARMA(p,q)$ process if it is stationary and if and only if, for $t \in \mathbb{Z}$, it satisfies the following difference equation

$$\phi(B)X_t = \theta(B)\varepsilon_t \quad (1)$$

Where μ is mean of the process, since B is the delay operator such as $BX_t = X_{t-1}$ and for all the entire k, $B^k X_t = X_{t-k}$ k, where:

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p, \theta(z) = 1 - \theta_1 z - ... - \theta_q z^q$$

are two polynomials and $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a white noise processes centred with a finite variance σ^2.

If $q = 0$, $(X_t)_{t\in\mathbb{Z}}$ becomes an $AR(p)$ process:

$$\phi(B)X_t = \varepsilon_t \quad (2)$$

If $p=0$, $(X_t)_{t\in\square}$ is a process of $MA(q)$:

$$X_t = \theta(B)\varepsilon_t \quad (3)$$

2.2.2 *ARIMA* **Processes**

A second order process $(X_t)_{t\in\square}$ is defined an $ARIMA$ process (p,d,q), but the process $((1-B)^d X_t)_{t\in\square}$ is an $ARIMA$ process.

$$\phi(\mathrm{B})(1-B)^d X_t = \theta(B)\varepsilon_t \quad (4)$$

where

$$\begin{cases}\nabla X_t = X_t - X_{t-1} = (1-B)X_t \\ \nabla^d X_t = (1-B)^d X_t\end{cases}$$

2.2.3 *SARMA* **Processes**

A SARMA model is defined as

$$\Phi_P(\mathrm{B}^s)X_t = \Theta_Q(\mathrm{B}^s)\varepsilon_t \quad (5)$$

Where

$$\Phi_P(\mathrm{B}^s)X_t = X_t - \phi_s X_{t-s} - \phi_{2s} X_{t-2s} - \ldots - \phi_{Ps} X_{t-Ps}$$

And

$$\Theta_Q(\mathrm{B}^s)\varepsilon_t = \varepsilon_t - \theta_s \varepsilon_{t-s} - \theta_{2s}\varepsilon_{t-2s} - \ldots - \theta_{Qs}\varepsilon_{t-Qs}$$

2.2.4 *SARIMA* **Processes**

$SARIMA$ process may be appeared as a generalization models of $ARIMA$, including a seasonal part. Generally, as $s_1,\ldots,s_n,n$ integers, then a process $(X_t)_{t\in\square}$ is a $SARIMA(p,d,q)$ - integrated autoregressive seasonal moving average process – following the below equation

$$\phi(B)(1-B^{s1})\ldots(1-B^{sn})X_t = \theta(B)\varepsilon_t \quad (6)$$

for all $t\geq 0$, $\phi_0=1,\theta_0=1$ where, $\phi(B),\theta(B)$ are polynomials whose roots are of modulus higher than 1. This form includes the $ARIMA$ models as it is enough to take $n=d$ and $s_1=\ldots=s_n=1$. However, the two most used forms are as follows:

$$\phi(B)(1-\mathrm{B}^s)X_t = \theta(B)\varepsilon_t \text{, pour } t\geq 0 \quad (7)$$

or any other way

$$\phi(B)(1-\mathrm{B}^s)(1-\mathrm{B})^d X_t = \theta(B)\varepsilon_t \quad (8)$$

For any $t\geq 0$ where only one seasonal factor (s) intervenes, either applied to an $ARMA$ process in the first case, or applied to an $ARIMA$ process in the second case.When including a sub-subsection you must use, forits heading, small letters, 11pt, left justified, bold, Times New Roman as here.

2.2.5 Seasonal multiplier processes

It's a mix of non-seasonal and seasonal patterns according to the following form

$$\phi_p(B)\Phi_P(\mathrm{B}^s)\nabla^d\nabla_s^D X_t = \theta_q(B)\Theta_Q(\mathrm{B}^s)\varepsilon \quad (9)$$

where :

p : Normal autoregressive degree.

P : Seasonal autoregressive degree.

Q :Order of the mean seasonal mobile.

d : Integration degree.

D : order of the seasonal difference.

s : Season duration.

The below indicated model is as follows:

$SARIMA(P,D,Q)\times(p,d,q)_s$

2.3 Box-Jenkins Method

The Box-Jenkins methology makes it possible to determine the $ARIMA(p,d,q)$ model accordingly to the modelling of a time series, as well as the behavior of a time series. This methodology suggests four steps: The identification, estimation, validation and prediction of the model [2, 4, 5, 7]. We will now present them in detail:

2.3.1 Identification (ACF, PACF)

The identification step is based on the theoretical information of the $ARMA(p,q)$ processes recognized in model a time series of data indicated by two statistical characteristics: or aurocorrelation function ACF and Partial autocorrelation function PACF. Since any seasonal component is supposed to be eliminated, the identification consists of specifying the three parameters p,d,q of the $ARIMA(p,d,q)$ model.

The stationarity of the model is first tested. Graphical study, correlograms and Augmented Dickey-Fuller test are performed in turn. If the series is not stationary, it should be transformed

(usually by differentiation) to obtain a stationary series.

The order of integration d is obtained by the number of times that the initial series has been differentiated to obtain stationarity. According to Augmented Dickey-Fuller test, correlograms analysis is used to determine it. Based on a stationary series, the autocorrelation function is analyzed by the set of autocorrelations:

$$\rho_k = corr(X_t, X_{t-k}), k \in \{1,...,k\}$$

k being the maximum allowable offset for the autocorrelation coefficient having a sense. This autocorrelation coefficient of order k, ρ_k, can be estimated by

$$r_k = \frac{\sum_{t=k+1}^{n} (X_t - \overline{X}_1)(X_{t-k} - \overline{X}_2)}{\sqrt{\sum_{t=k+1}^{n} (X_t - \overline{X}_1)^2 \sum_{t=k+1}^{n} (X_{t-k} - \overline{X}_2)^2}}$$

where

$$\overline{X}_1 = \frac{1}{n-k} \sum_{t=k+1}^{n} X_t, \overline{X}_2 = \frac{1}{n-k} \sum_{t=k+1}^{n} X_{t-k}$$

According to the hypothesis $H_0 : \rho_0 = 0$, the statistic $t_c = \frac{|r_k|}{\sqrt{1-r_k^2}}$ follows a law of Student at $n-2$ degrees of freedom. If the calculated value of t_c is greater than the order of quantile $\alpha/2$ of a Student's law at $n-2$ degrees of freedom becomes $t_c > t_{n-2}^{\alpha/2}$, then the hypothesis H_0 is rejected at the threshold α. The partial autocorrelation function designates all the autocorrelations between the variables X_t and X_{t-k}, the influence of the variable X_{t-k-i} being controlled for $i < k$.

Table 1: ACF and PACF not seasonal pattern

No.	Model	ACF	PACF
1.	$AR(p)$	Downtrend exponentially	Cut off after lag p
2.	$MA(q)$	Cut off after lag q	Downtrend exponentially
3.	$ARMA(p,q)$	Cut off after lag (q-p)	Cut off after lag (q-p)

Table 2: ACF and PACF seasonal pattern

No.	Model	ACF	PACF
1.	$SAR(P)$	Decrease exponentially on seasonal lag	Cut off after lag Ps
2.	$SMA(Q)$	Cut off after lag Qs	Decrease exponentially on seasonal lag
3.	$SARMA(P,Q)$	Cut off on seasonal lag	Cut off on seasonal lag

2.3.2 Estimation of $ARIMA$ model parameters

The software of computer science as Eviews leads to estimate the coefficient of identified models in the previous mentioned step. The p, q and d parameters are obviously specified. In theory, the estimation of $ARIMA(p,d,q)$ model parameters when p, d, q are supposed to be known may be determined following various methods in temporal domains.

- At less ordinary square when $q = 0$, we can use an equation called Yule - Walker equation. By replacing the autocorrelations by their estimators, we can find the estimators OLS of the model through solving the equations of Yule-Walker.
- Maximum likelihood approach.

2.3.3 The diagnosis of an $ARIMA$ model

In this part, it is necessary to carry out a set of checks.

- **Parameter test (coefficients)**

Among the estimated $ARMA$ processes, only those whose coefficients have a t Student > $1,96$ (for a risk of 5% and for a sufficiently large sample size: $T > 30$).

- **Choice of the model**

We choose the model that minimizes standard and information criteria. The selection model will then be used for the forecast.

- **Residual test**

The values of the autocorrelation and partial autocorrelation functions of the residual series must

all be nil. If the self-correlations of order 1 and 2 differ significantly from zero, then the specification of the model becomes surely not suitable. It may happen, however, that one or two higher- order autocorrelations randomly exceed the limits of the 95% confidence interval. The characteristics of the residues must correspond to those of a white noise. The Q statistic of Box and Jung, still known as modified by Box and Pierce statistics is commonly used to test the white noise hypothesis for being retained for the residual series. It is defined by

$$Q = n(n+2)\sum_{k=1}^{K}\frac{r_k^2(\varepsilon_t)}{n-k} \tag{10}$$

And follows the law $\chi^2_{k-(p+q)}$ according to the hypothesis, i.e,
$H_0 : \rho_1 = 0, \rho_2 = 0, ..., \rho_k = 0$
hypothesis absence of the autocorrelation.

2.4 The Kalman filter

The Kalman filter is an algorithm that provides an efficient recursive solution to the least-squares problem .The KF allows a unified approach to prediction and estimation for all processes that can be given by state space representation [3, 23]. The classical Kalman recursions were introduced by Rudolph E. Kalman 1960 [13, 14, 16, 22].
The Kalman filter method has many advantages. It relies only on the recursive method and does not require all historical data. It can be used to deal with not only the stationary and non- stationary random processes, but also time-varying and non-time-varying systems [9, 16, 17].
A state space model is generally represented in two equations: first equation is called the state equation which finds the state X_{t+1} in time $(t+1)$ using the previous state X_t and a noise term as shown below.

$$X_{t+1} = AX_t + W_t, t = 1,2,... \tag{11}$$

Where A is sequences of $n \times n$ matrices and W_t is the process disturbance $\sim N(0, Q_t)$. The second equation is the observation equation which expresses the $m-$dimensional observation Y_t as a function of a $n-$dimensional state variable X_t and noise. Thus:

$$Y_t = BX_t + V_t, t = 1,2,... \tag{12}$$

Where V_t is the measurement noise $\sim N(0, R_t)$ and B is a sequence of $m \times n$ matrices, the processes $\{W_t\}$ and $\{V_t\}$ are uncorrelated. When a time series consists of daily, monthly, or quarterly observations, the presence of seasonal effects should be investigated. Hence, adding seasonal components to equation (12) [18, 22].

2.4.1 State space Representation of $(S)ARIMA$ models

The observation equation of the state space represents $ARMA(p,q)$ is

$$Y_t = \begin{bmatrix} 1 & 0 \cdots 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-i-1} \end{bmatrix} + W_t \tag{13}$$

The state equation of $ARMA(p,q)$ is:

$$\begin{bmatrix} X_{t+1} \\ X_t \\ \vdots \\ X_{t-i} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{i-1} & \phi_i \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-i-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{i-1} \end{bmatrix} \varepsilon_t \tag{14}$$

Where

$$A = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{i-1} & \phi_i \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$$

And

$$W_t = \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{i-1} \end{bmatrix} \varepsilon_t , \quad Q_t = \sigma^2 \begin{bmatrix} 1 & \theta_1 & \cdots & \theta_1 \\ \theta_1 & \theta_1^2 & \cdots & \theta_1\theta_{i-1} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_1 & \theta_1\theta_{i-1} & \cdots & \theta_{i-1}^2 \end{bmatrix}$$

This leads to an $ARMA(p,q)$ model, for which $i = \max(p, q+1)$, $\phi_i = 0$, for $i > p$ and $\theta_i = 0$, for $i > q+1$ [5, 6, 14].

Thus before applying the Kalman filter we need to estimate the matrices A, B, W and Q_t.

$SARIMA$ model can be dealt with by constructing $ARMA$ models for the stationary differenced series $Y_t = (1-B)^d (1-B^s)^D X_t$ and placing the non stationary variables such as X_{t-i} and $(1-B)^d X_{t-i}$ in the state vector. Y_t is a seasonal $ARMA(p^*, q^*)$ process with $p^* = p + sP$ and $q^* = q + sQ$.

The seasonal $ARIMA(1,1,1)\times(0,1,1)_s$ model with a nonzero mean term can be written as

$$\begin{aligned} &(1-\phi B)(1-B)^d (1-B^s)^D Z_t \\ &= c + (1-\theta B)(1-\Theta B^s)\varepsilon_t \end{aligned} \tag{15}$$

we denote the stationary and seasonality differenced flow series, $Y_t = (1-B)^d (1-B^s)^D Z_t$, the model can be rewritten as

$$Y_t = c + \phi Y_{t-1} - \theta\varepsilon_{t-1} - \Theta\varepsilon_{t-s} + \theta\Theta\varepsilon_{t-s-1} + \varepsilon_t \tag{16}$$

And in vector as

$$\begin{aligned} Y_t = &\begin{bmatrix} 1 & Y_{t-1} & -\varepsilon_{t-1} & -\varepsilon_{t-s} \end{bmatrix} \begin{bmatrix} c & \phi & \theta & \Theta \end{bmatrix}^T \\ &+ \theta\Theta\varepsilon_{t-s-1} + \varepsilon_t \end{aligned} \tag{17}$$

Thus the model has the following state space representation :

Observation equation:

$$Y_t = AX_{t-1} + Cs_t + \varepsilon_t \tag{18}$$

where Cs is a seasonal components.

State transition equation:

$$X_t = X_{t-1} + V_t \tag{19}$$

2.4.2 Kalman filter prediction algorithm

The KF is constituted essentially by a set of five mathematical equations that implement a predictor – corrector- type estimator that is optimal in the sense that is minimizes the estimated error covariance, when some presumed conditions are not. Those equations are recursive and present the main advantage to provide the prediction error accurately.

The first three equations are the predictor equations (before observation at time t+1 is available).

1- State mean prediction

$$\hat{X}_{t+1|t} = A\hat{X}_{t|t} , \quad \hat{X}_{t+1|t} = IE(X_{t+1}|Y_t) \tag{20}$$

2- State covariance prediction

$$P_{t+1|t} = AP_{t|t}A^T + Q_t , \quad P_{t+1|t} = Var(X_{t+1}|Y_t) \tag{21}$$

3- The Kalman gain (denoted by K) was calculated using:

$$K_t = P_{t+1|t}B^T (BP_{t+1|t}B^T + R_t)^{-1} \tag{22}$$

The Kalman gain can be thought of as the weight given to the most recent observation for updating mea and covariance of the state.

The two corrector equations update mean vector and covariance matrix after observation at time t+1 is available.

4- State update

$$\hat{X}_{t+1|t+1} = \hat{X}_{t+1|t} + K_t e_t , \quad e_t = X_t - \hat{X}_{t|t} \tag{23}$$

5- Covariance

$$P_{t+1|t+1} = (1 - K_{t+1}B)P_{t+1|t} \tag{24}$$

The unknown variance parameters in the state space model are estimated by the maximum likelihood estimation via the Kalman filter prediction error decomposition initialized with the exact Kalman filter [8, 9, 16, 24].

The unknown variance parameters in the state space model are estimated by the maximum likelihood estimation via the Kalman filter prediction error decomposition initialized with the exact initial Kalman filter [21, 22, 24].

3 Problem Solution

3.1 The data

The monthly data of injuries to road accidents in Skikda from January 2001 to December 2016 are examined in this study shall be herein IRA.

Table 3: Monthly Data of IRA

Date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nob	Dec
2001	16	12	20	13	12	25	33	13	39	23	28	22
2002	16	23	8	17	5	11	14	19	9	13	12	14
2003	8	4	10	18	6	19	34	31	16	7	21	19
2004	11	14	17	16	3	21	32	27	22	12	16	29
2005	21	12	22	13	8	21	43	46	48	24	7	14
2006	20	26	24	34	27	49	71	58	47	25	23	18
2007	37	17	21	32	25	23	39	63	46	35	51	26
2008	27	27	42	30	32	65	109	71	34	54	33	40
2009	39	79	35	54	50	95	184	115	38	55	69	64
2010	32	32	68	37	47	59	85	86	62	64	51	49
2011	48	27	52	52	80	102	203	118	83	76	75	84
2012	95	50	86	148	84	194	195	200	81	91	62	86
2013	90	81	116	150	77	125	195	212	106	123	61	106
2014	93	87	84	109	114	186	221	193	143	109	141	94
2015	60	59	107	106	107	173	221	185	130	92	78	108
2016	103	90	117	121	78	109	220	200	123	72	69	83

3.2 SARIMA modeling process

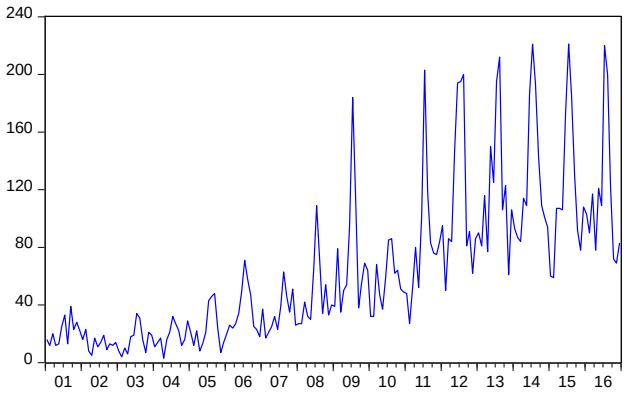


Figure 1: The time plot of IRA

Figure 1 shows the monthly number of road accidental injuries in Skikda from January 2001 to December 2016 is presented in Table 3. Over the past 16 years, the annual numbers varied greatly in Skikda from a low 3 in May 2004 to a high of 221 in Jul 2014 and 2015.

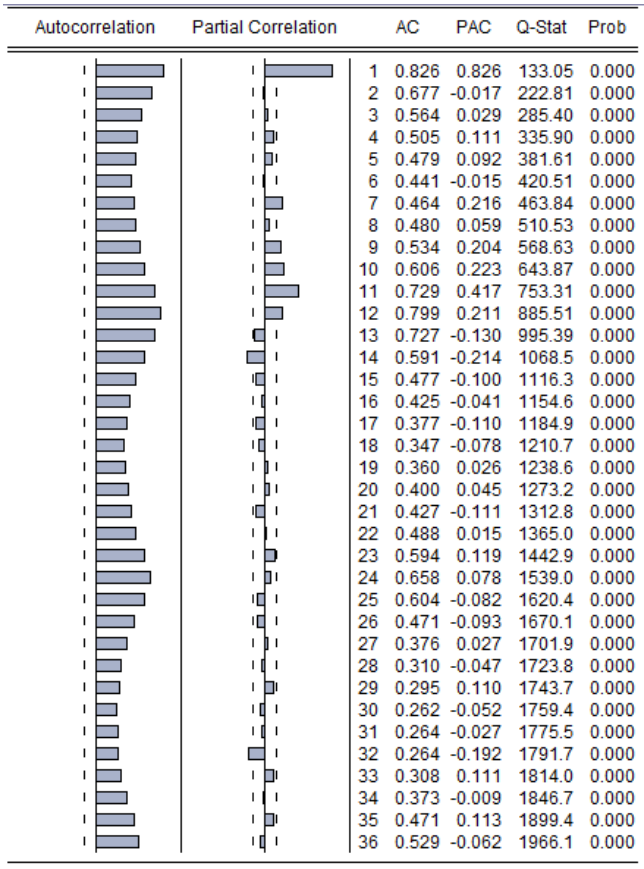
The time plot of Figure 1 shows a generally positive trend depicting relative depreciation of monthly road accidental injuries then the Figure 1 presents a non stationary mean, so it was necessary to stabilize the mean of monthly number by first-order trend difference.

Note that from 2008 to 2016 are years with large number of individuals with the injuries, and graphical description of monthly cases of the injuries with high values occurring in summer from 2008 to 2016, in spring and autumn from 2011 to 2016, and finally in winter from 2009 to 2016.This recurring pattern is an indication of a seasonal effect.

The ACF plot (Figure 2) has the shape typical for seasonal time series, it has a recurrent pattern: there are significant peaks at the seasonal frequencies (lag 12, 24, 36, etc) which decay slowly and the p-value of the Ljung-Box statistic fall below the 0,05, these observations imply that the standardized residuals are correlated, and the Table 4 shows that the p-value is greater than 0,05 and |value of ADF| = 1,984<|the critical value|. Thus we accept the null hypothesis at 1%, 5% level of significance that the time series is unit root non stationary [1, 29].

Table 4: The augmented Dickey-Fuller for IRA

		Test statistic	p-VALUE
ADF		-1,984	0,605
Critics Value	1%	-4,009	
	5%	-3,434	



		AC	PAC	Q-Stat	Prob
1		0.826	0.826	133.05	0.000
2		0.677	-0.017	222.81	0.000
3		0.564	0.029	285.40	0.000
4		0.505	0.111	335.90	0.000
5		0.479	0.092	381.61	0.000
6		0.441	-0.015	420.51	0.000
7		0.464	0.216	463.84	0.000
8		0.480	0.059	510.53	0.000
9		0.534	0.204	568.63	0.000
10		0.606	0.223	643.87	0.000
11		0.729	0.417	753.31	0.000
12		0.799	0.211	885.51	0.000
13		0.727	-0.130	995.39	0.000
14		0.591	-0.214	1068.5	0.000
15		0.477	-0.100	1116.3	0.000
16		0.425	-0.041	1154.6	0.000
17		0.377	-0.110	1184.9	0.000
18		0.347	-0.078	1210.7	0.000
19		0.360	0.026	1238.6	0.000
20		0.400	0.045	1273.2	0.000
21		0.427	-0.111	1312.8	0.000
22		0.488	0.015	1365.0	0.000
23		0.594	0.119	1442.9	0.000
24		0.658	0.078	1539.0	0.000
25		0.604	-0.082	1620.4	0.000
26		0.471	-0.093	1670.1	0.000
27		0.376	0.027	1701.9	0.000
28		0.310	-0.047	1723.8	0.000
29		0.295	0.110	1743.7	0.000
30		0.262	-0.052	1759.4	0.000
31		0.264	-0.027	1775.5	0.000
32		0.264	-0.192	1791.7	0.000
33		0.308	0.111	1814.0	0.000
34		0.373	-0.009	1846.7	0.000
35		0.471	0.113	1899.4	0.000
36		0.529	-0.062	1966.1	0.000

Figure 2: The correlogram of IRA

Therefore, the differencing process is necessary to obtain the stationary data. By taking difference d=1 for non seasonal and D=1 with s=12 for seasonal the data become stationary series (Figure 3) [21, 29].

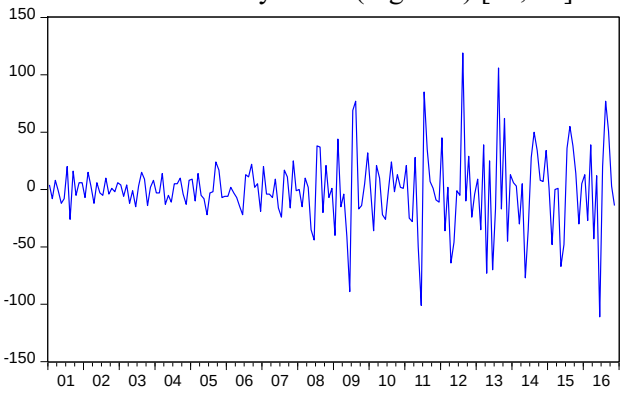


Figure 3: The time plot of SDDIRA

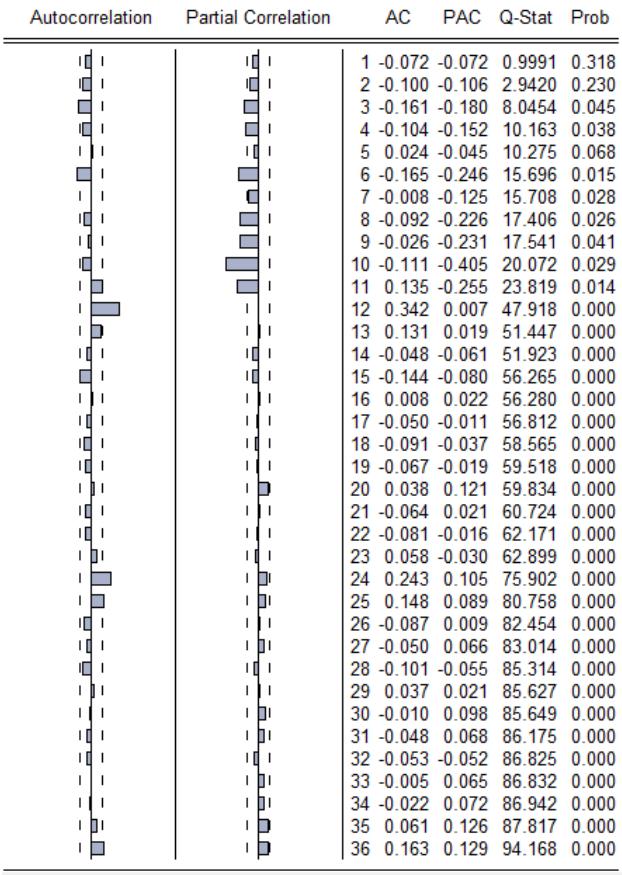
Table 5: The augmented Dickey-Fuller for SDDIRA

		Test statistic	p-VALUE
ADF		-13,436	0,000
Critics Value	1%	-4,009	
	5%	-3,434	

The Table 5 shows |value of ADF|=13,436> |the critical value| and p-value is less than 0,05. It means that it had enough evidence to reject H_0 that the data were stationary to mean [1, 11].

The Figure 4 shows that in the ACF plot there are three significant lags, lag 12, 24 and 36 (every season there is a pick). The partial correlogram indicates an exponentially regression with two or three terms being significant. Based on these patterns, some possible candidate SARIMA models are defined.

Table 6 shows there are three models that have significant parameters at $\alpha = 0$ they were model $SARIMA(1,1,0)(1,1,0)_{12}$, $SARIMA(0,1,1)(0,1,1)_{12}$ and $SARIMA(12,1,1)(0,1,1)_{12}$



		AC	PAC	Q-Stat	Prob
1		-0.072	-0.072	0.9991	0.318
2		-0.100	-0.106	2.9420	0.230
3		-0.161	-0.180	8.0454	0.045
4		-0.104	-0.152	10.163	0.038
5		0.024	-0.045	10.275	0.068
6		-0.165	-0.246	15.696	0.015
7		-0.008	-0.125	15.708	0.028
8		-0.092	-0.226	17.406	0.026
9		-0.026	-0.231	17.541	0.041
10		-0.111	-0.405	20.072	0.029
11		0.135	-0.255	23.819	0.014
12		0.342	0.007	47.918	0.000
13		0.131	0.019	51.447	0.000
14		-0.048	-0.061	51.923	0.000
15		-0.144	-0.080	56.265	0.000
16		0.008	0.022	56.280	0.000
17		-0.050	-0.011	56.812	0.000
18		-0.091	-0.037	58.565	0.000
19		-0.067	-0.019	59.518	0.000
20		0.038	0.121	59.834	0.000
21		-0.064	0.021	60.724	0.000
22		-0.081	-0.016	62.171	0.000
23		0.058	-0.030	62.899	0.000
24		0.243	0.105	75.902	0.000
25		0.148	0.089	80.758	0.000
26		-0.087	0.009	82.454	0.000
27		-0.050	0.066	83.014	0.000
28		-0.101	-0.055	85.314	0.000
29		0.037	0.021	85.627	0.000
30		-0.010	0.098	85.649	0.000
31		-0.048	0.068	86.175	0.000
32		-0.053	-0.052	86.825	0.000
33		-0.005	0.065	86.832	0.000
34		-0.022	0.072	86.942	0.000
35		0.061	0.126	87.817	0.000
36		0.163	0.129	94.168	0.000

Figure 4: The correlogram of SDDIRA

Table 6: Estimation of three models

Model	Coefficient	Prob	SE	t
AR(1)	-0,3809	0,0000	0,0705	-5,3969
SAR(12)	0,6448	0,0000	0,0658	9,7887
MA(1)	-0,4470	0,0000	0,0671	-6,6523
SMA(12)	0,5324	0,0000	0,0649	8,1943
AR(12)	1,1097	0,0000	0,0151	73,2488
MA(1)	-0,4794	0,0000	0,0501	-14,9338
SMA(12)	-0,8979	0,0000	0,0245	-36,5924

However, to determine the best model, it can be selected based on the model that has the smallest AIC (Akaike's Information Criteria) and SBC (Scwartz's Bayesian Criteria) criterion value.
The following table displays the AIC and SBC value calculation for three models.

Table 7: Value AIC and SBC of SARIMA Models

Model	AIC	SBC
$SARIMA(1,1,0)(1,1,0)_{12}$	9,4387	9,4739
$SARIMA(0,1,1)(0,1,1)_{12}$	9,4859	9,5196
$SARIMA(12,1,1)(0,1,1)_{12}$	8,9221	8,9755

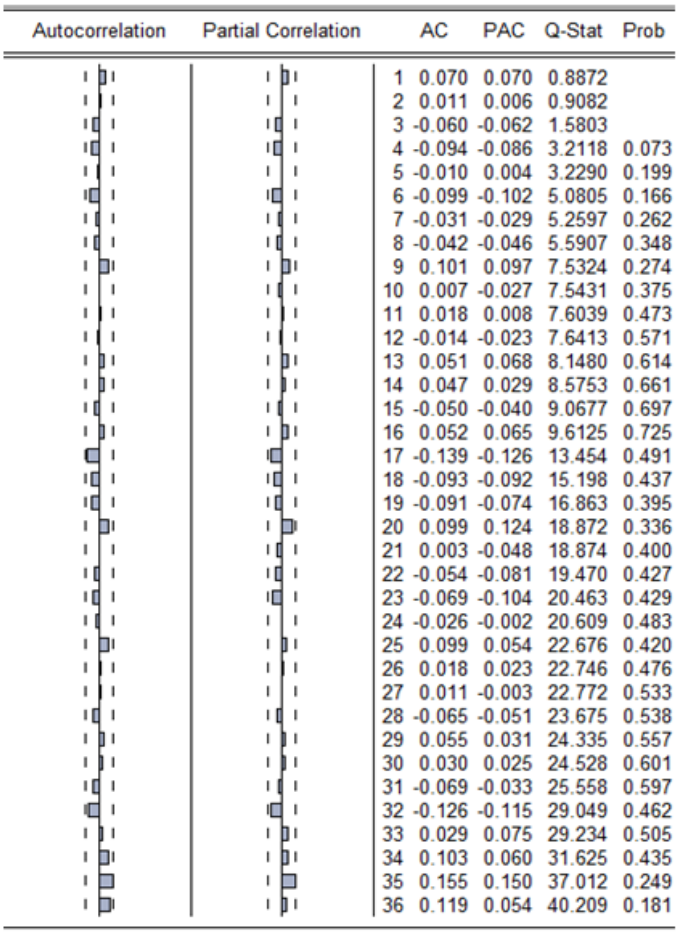
The chosen seasonal ARIMA is of the form:

$$Y_t - 1{,}109730 Y_{t-12} = 0{,}749499\varepsilon_{t-1} + 0{,}897996\varepsilon_{t-12} + \varepsilon_t \quad (23)$$

Where:

$$Y_t = (1-B^{12})(1-B)IRA$$

It should be noted that the coefficients of simple autocorrelation and partial autocorrelation are all within the confidence interval, which means that the residuals of this model indicate a white noise (Figure 5).



	AC	PAC	Q-Stat	Prob
1	0.070	0.070	0.8872	
2	0.011	0.006	0.9082	
3	-0.060	-0.062	1.5803	
4	-0.094	-0.086	3.2118	0.073
5	-0.010	0.004	3.2290	0.199
6	-0.099	-0.102	5.0805	0.166
7	-0.031	-0.029	5.2597	0.262
8	-0.042	-0.046	5.5907	0.348
9	0.101	0.097	7.5324	0.274
10	0.007	-0.027	7.5431	0.375
11	0.018	0.008	7.6039	0.473
12	-0.014	-0.023	7.6413	0.571
13	0.051	0.068	8.1480	0.614
14	0.047	0.029	8.5753	0.661
15	-0.050	-0.040	9.0677	0.697
16	0.052	0.065	9.6125	0.725
17	-0.139	-0.126	13.454	0.491
18	-0.093	-0.092	15.198	0.437
19	-0.091	-0.074	16.863	0.395
20	0.099	0.124	18.872	0.336
21	0.003	-0.048	18.874	0.400
22	-0.054	-0.081	19.470	0.427
23	-0.069	-0.104	20.463	0.429
24	-0.026	-0.002	20.609	0.483
25	0.099	0.054	22.676	0.420
26	0.018	0.023	22.746	0.476
27	0.011	-0.003	22.772	0.533
28	-0.065	-0.051	23.675	0.538
29	0.055	0.031	24.335	0.557
30	0.030	0.025	24.528	0.601
31	-0.069	-0.033	25.558	0.597
32	-0.126	-0.115	29.049	0.462
33	0.029	0.075	29.234	0.505
34	0.103	0.060	31.625	0.435
35	0.155	0.150	37.012	0.249
36	0.119	0.054	40.209	0.181

Figure 5: Residual diagnostic plots

The Figure 6 shows the predicted values and current values for the year 2017.

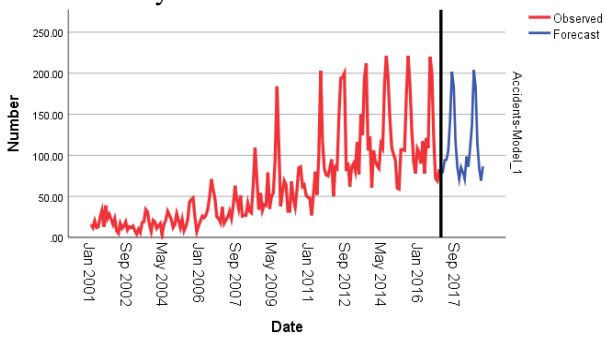


Figure 6: Forecasting with SARIMA model

3.2 Comparison between Box Jenkins method and Kalman Fiter recursions

In this subsection, we give a comparative study between SARIMA Box Jenkins and SARIMA Kalman models.

Table 8: Forecasting data by SARIMA and SARIMA Kalman model

Month	Jan 2017	Feb 2017	Mar 2017	Apr 2017	May 2017	Jun 2017
BJF	78	80	94	94	106	143
KFF	72.75	71.87	129.17	148.61	134.53	145.59
Month	Jun 2017	Jul 2017	Aug 2017	Sep 2017	Oct 2017	Nov 2017
BJF	143	202	184	116	85	71
KFF	145.59	182.29	192.26	135.57	84.59	98.91

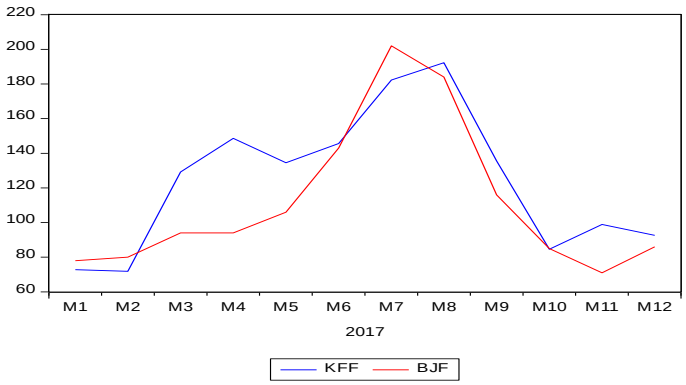


Figure 6: Forecasting with BJ method and KF recursions

Table 7 and figure 7 shows the forecasting performance: MAE, MAPE, RMSE and Theil's Statistics between Box Jenkins method and Kalman Filter recursions.

Table 9: Forecasting performance: MAE, MAPE, RMSE and Theil's Statistics

Method	SARIMA Box Jenkins	SARIMA Kalman
MAE	1.161	0.187
MAPE	0.944	0.172
RMSE	6.124	0.927
Theil's U	0.035	0.005

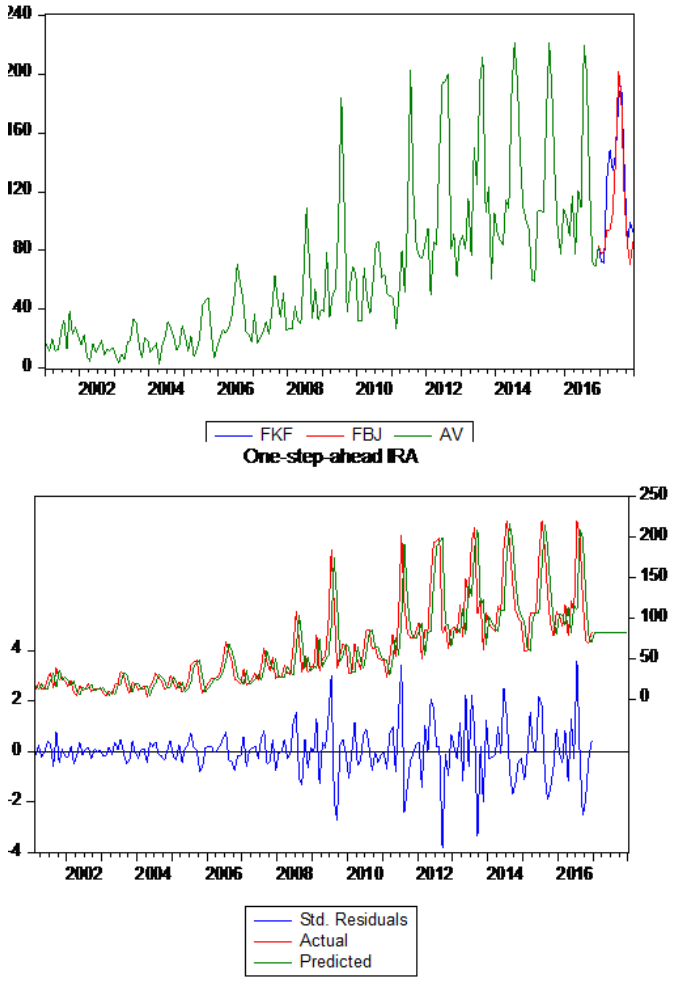


Figure 7: Plots of Actual values versus Box-Jenkins and Kalman recursions

4 Conclusion

This work allowed us to model seasonal time series using ARIMA integrated linear processes and seasonal SARIMA linear processes, using Box-Jenkins techniques: The identification of the model using the autocorrelation and partial autocorrelation functions, the estimation of the parameters by the likelihood method (LM) and in n the data criteria AIC and BIC are proposed to test the quality of the chosen model. Our empirical study focused on a univariate random process and estimated from the model $SARIMA(12,1,1)\times(0,1,1)_{12}$ which is valid and test in the case study. Also, this study shows that the Kalman filter method is efficient for forecasting by input to $(S)ARIMA$ processes.

References:

[1] A.C. Akpanta, I.E, Application of Box-Jenkins Techniques in Modeling and Forecasting Nigeria Crude Oil Prices, *International Journal of Statistics and Applications*, Vol.4, No.6, 2014, pp. 283-291.

[2] Afrifa-Yamoah E, Application of ARIMA Models in Forecasting Monthly Average Surface Temperature of Brong Ahafo Region of Ghana, *International Journal of Statistics and Applications*, Vol. 5, No. 5, 2015, pp. 237-246.

[3] Balakrishman, A.V, *Kalman Filtring Theory,* 2nd. ED. Optimisation Software, Inc, New York, 1987.

[4] Billy M. Williams, M.ASCE and Lester A. Hotel, F.ASCE, Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results, *Journal of Transportation Engineering,* Vol. 129, No. 6, 2003, pp. 664-672.

[5] Box G.E.P.,Jenkins G.M, *Time Series Analysis: Forecasting and Control,* Holden-Day, San Francisco, 1976.

[6] Brockewel, P.J., Davis, R.A. *Introduction to Time Series and Forecasting.* 2nd. ED, Springer Velag. 2002.

[7] Dominic Buchstaller, Jing Liu, Mark French, The Deterministic Interpretation of the Kalman Filter, *International Journal of Control,* Vol.

[8] Dongwei Xu, Yong-dong WANG, Li-min JLA, Yong QIN, Hong-hi DONG, Rea-time road traffic state prediction based on ARIMA and Kalamn filter, *Frontiers of Information Technology and Electronic Engineering,* Vol. 18, No.2, 2017, pp. 267-302.

[9] Durbin J, Koopman S.J, *Time series analysis by state space methods.* Oxford University Press, London, UK, 2001.

[10] Ette Harrison Etuk, Tariq Mahgoub Mohamed, Time Series Analysis of Monthly Rainfall data for the Gadaref rainfall station, Sudan, by SARIMA Methods. *International Journal of Scientific Re-search in Knowledge,* Vol.2, No.7, 2014, pp. 320-327.

[11] Ette Harrison Etuk, Nathaniel Ojekudo, Subset Sarima Modeling: An Alternative Definition and a Case Study, *British Journal of Mathematics Computer Science,* Vol.5, No.4, 2015, pp. 538-552.

[12] Ette Harrison Etuk, Bartholomew Uchendu, Mazi Yellow Dimkpa, A Box- Jenkins Method Subset Simulating Model for Daily Ugx- Ngn Exchange Rate, *Journal of Scientific and Engineering Research*, Vol.3, No.2, 2016, pp 11-15.

[13] Ette Harrison Etuk, Tariq Mahgoub Mohamed, Application of linear stochastic models to monthly streamflow data of Rahad River, Sudan, *International Journal. Hydrology Science and Technology,* Vol.7, No.2, 2017, pp. 197-212.

[14] Haykin, S, *Adaptive Filter Theory*, 4th ED, Prentice-Hall, Englewood Cliffs, Inc, New York, 2002.

[15] Hamilton, J.D, *Time Series Analysis,* Princeton University Press, New Jersey, 1994.

[16] Hindrayanto, I. Koopman, S.J, Ooms, M, Exact maximum likelihood estimation for non-stationary periodic time series models, *Computational Statistics and Data Analysis,* Vol. 54, No. 11, 2010, pp. 2641-2654.

[17] Jibril Y Kajuru, Kamaluddin Abdulkarim, Muhamed M Muhamed, Forecasting performance of ARIMA and SARIMA Models on Monthly Average Temperature of Zaria, Nigeria, *Journal of Science Technology and Education*, Vol.7, No.3, 2019, pp. 205-212.

[18] Kalman, R, A new approach to linear filtring and prediction problem, *Journal of basic Engineering,* Vol. 82, 1960, pp. 35-45.

[19] Koopman, S, Exact Initial Kalman Filtring and Smoothing for Nonstationary Time Series Models, *Journal of the American Statistical Association*, Vol. 92, No.440, 1997, pp. 1630-1638.

[20] Md. Siray Ud Doulah, Forecasting Temperatures in Bangladesh: An application of

SARIMA Models, *International Journal of Statistics and Mathematics*, Vol. 5, No. 1, 2018, pp. 108-118.
[21] M. Milenkovic, N. Bojovic, A Recursive Kalman Filter Approach to Forecasting Railway Passenger Flows, *Iternational Journal of Railiway Technology*, Vol. 3, No. 2, 2014, pp. 39-57.
[22] M. Nirmala and Tariq Mahgoub Mohamed, Seasonal Predictability of Rinfall data using Box- Jenkins models in Kordofan State, Sudan, *Indian Journal of Science and Technology,* Vol. 11, No. 48, 2018, pp. 1-9.
[23] Mustafa M.Ali Alfaki, Shalini Bhawana Masih, Modeling and Forecasting by using Time series ARIMA Models, *International Journal of Engineering Research and Technology*, Vol. 4, No. 3, 2015, pp. 914-918.
[24] Omorophe. Joseph ASEMOTA, State Space Versus, *Sri Lankan Journal of Applied Statistics*, Vol. 17, No. 2, 2016, pp. 87-108.
[25] Tendai Makoni, Talent. D. Murwendo, Romeo Mawonike, Musara Chipumuro, *African Journal of Hospitality, Tourism and Leisur*, Vol. 8, No. 1, 2019, pp. 1-8.
[26] Toyo Pamela Naden, Ette Harrison Etuk, SARIMA Modeling of Nigerian Food Consumer Price Indices, *International Journal of Science and Advenced Inovative Research*, Vol. 2, No. 4, 2017, pp. 56-67.
[27] Shashank Shekhar, Ill Williams, Adaptive Seasonal Time Series Models for Forecasting Short-Term Traffic Flow, *Journal of the Transportation Research Board,* No. 2024, 2007, pp. 116-125.
[28] Suhartono, Time Series Forecasting by using Autoregressive Integrated Moving Average: Subset, Multiplicative or Additive Model. *Journal of Mathematics and Statistics*. Vol.7, No.1, 2011, pp. 20-27.
[29] Susan W, Gikungu, Anthony G, Waititu, John M. Kihoro, Forecasting inflation rate in Kenya using SARIMA model, *American Journal of Theoretical and Applied Statistics,* Vol. 4, No. 1, 2015, pp. 15-18.
[30] Wang, Y, Papageorgiou, M, *Real time freeway* traffic state estimation based on extend Kalman Filer: a case study. *Transprtation Science,* Vol. 42, No. 2, 2007, pp. 167-181.
[31] Xiaosheng Li, Chunliu Ma, Haike Lei Haixia Li, Applications of SARIMA Model in Forecasting Outpatient Amount. Chinese Medical Record English Edition, *Journal of Cardiovascular Magnetic Resonance*. Vol.1, No.3, 2013, pp. 124-128.