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-Jenkins method, Kalman filter, Forecasting


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) × (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:

\[ i) \forall t \in \mathbb{R}, IE(X_t) \]
\[ ii) \forall t \in \mathbb{R}, \text{Var}(X_t) = \sigma^2 = \gamma(0) \]
\[ iii) \forall t \in \mathbb{R}, \forall h \in \mathbb{Z}, \text{Cov}(X_t, X_t-h) = \gamma(h) \text{ (depends only to } h) \]

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

### 2.1.2 White noise

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

\[ IE(\varepsilon_t) = 0, \forall t \in \mathbb{R} \]
\[ \text{Var}(\varepsilon_t) = \sigma^2, \forall t \in \mathbb{R} \]
\[ \text{Cov}(\varepsilon_t, \varepsilon_s) = 0, \text{ if } t \neq s \]

### 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 centralized and stationary process, thus we get the following decomposition:

\[ X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + V_t, t \in \mathbb{R} \]

where :

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

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

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

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

### 2.1.5 Linear series

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

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

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

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

### 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 shifted 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 the partial correlation coefficient between \(X_t\) and \(X_{t-h}\), i.e the correlation between \(X_t\) and \(X_{t-h}\) shiitted by \(h\) 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 \((X_t)_{t \in \mathbb{R}}\) is defined as an ARMA\((p,q)\) process if it is stationary and if and only if, for \(t \in \mathbb{R}\), it satisfies the following difference equation

\[ \phi(B)X_t = \theta(B)\varepsilon_t \tag{1} \]

Where \(\mu\) is mean of the process, since \(B\) is the delay operator such as \(BX_t = X_{t-1}\) and for all the entire \(k, B^k X_t = X_{t-k}\), 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 \((\varepsilon_t)_{t \in \mathbb{R}}\) is a white noise processes centred with a finite variance \(\sigma^2\).

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

\[ \phi(B)X_t = \varepsilon_t \tag{2} \]
If \( p = 0 \), \((X_t)_{t\in\mathbb{Z}}\) is a process of MA\((q)\):
\[
X_t = \theta(B)\varepsilon_t
\]  
(3)

### 2.2.2 ARIMA Processes

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

\[
\phi(B)(1 - B)^d X_t = \theta(B)\varepsilon_t
\]  
(4)

where

\[
\begin{align*}
\nabla X_t &= X_t - X_{t-1} = (1 - B)X_t \\
\nabla^d X_t &= (1 - B)^d X_t
\end{align*}
\]

### 2.2.3 SARMA Processes

A SARMA model is defined as

\[
\Phi_p(B^s)X_t = \Theta_q(B^s)\varepsilon_t
\]  
(5)

where

\[
\Phi_p(B^s)X_t = X_t - \phi_1 X_{t-s} - \phi_2 X_{t-2s} - \ldots - \phi_{s_1} X_{t-s_1}
\]

And

\[
\Theta_q(B^s)\varepsilon_t = \varepsilon_t - \theta_1 \varepsilon_{t-s} - \theta_2 \varepsilon_{t-2s} - \ldots - \theta_{s_1} \varepsilon_{t-s_1} - \Theta_q \varepsilon_t
\]

### 2.2.4 SARIMA Processes

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

\[
\phi(B)(1 - B^{-s_1})\ldots(1 - B^{-s_m})X_t = \theta(B)\varepsilon_t
\]  
(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:

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

or any other way

\[
\phi(B)(1 - B^s)(1 - B)^d X_t = \theta(B)\varepsilon_t
\]  
(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-section you must use, for its 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

\[
\phi_p(B)\Phi_p(B^s)\varepsilon_t = \theta_q(B^s)\Theta_q(B^s)\varepsilon_t
\]  
(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 methodology 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:

\[ \rho_k = \text{corr}(X_t, X_{t-k}), k \in \{1, \ldots, k\} \]

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

\[ r_k = \frac{\sum_{i=k+1}^{n} (X_i - \bar{X})(X_{i-k} - \bar{X})}{\sqrt{\sum_{i=k+1}^{n} (X_i - \bar{X})^2 \sum_{i=k+1}^{n} (X_{i-k} - \bar{X})^2}} \]

where

\[ \bar{X}_1 = \frac{1}{n-k} \sum_{i=k+1}^{n} X_i, \bar{X}_2 = \frac{1}{n-k} \sum_{i=k+1}^{n} X_{i-k} \]

According to the hypothesis \( H_0 : \rho_0 = 0 \), the statistic \( t_r = \frac{|r_k|}{\sqrt{1-r_k^2}} \) follows a law of Student at \( n-2 \) degrees of freedom. If the calculated value of \( t_r \) is greater than the order of quantile \( \alpha/2 \) of a Student’s law at \( n-2 \) degrees of freedom becomes \( t_r > t_{\alpha/2} \), then the hypothesis \( H_0 \) is rejected at the threshold \( \alpha \). The partial autocorrelation function designates all the autocorrelations between the variables \( X_i \) and \( X_{i-k} \), the influence of the variable \( X_{i-k-1} \) being controlled for \( i < k \).

**Table 1: ACF and PACF not seasonal pattern**

<table>
<thead>
<tr>
<th>No.</th>
<th>Model</th>
<th>ACF</th>
<th>PACF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>AR((p))</td>
<td>Downtrend exponentially</td>
<td>Cut off after lag p</td>
</tr>
<tr>
<td>2.</td>
<td>MA((q))</td>
<td>Cut off after lag q</td>
<td>Downtrend exponentially</td>
</tr>
<tr>
<td>3.</td>
<td>ARMA((p,q))</td>
<td>Cut off after lag (q-p)</td>
<td>Cut off after lag (q-p)</td>
</tr>
</tbody>
</table>

**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

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

(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.

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

(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:

$$Y_t = BX_t + V_t, t = 1, 2,...$$

(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

$$Y_t = [1 \ 0 \ldots 0]_{1 \times m} X_{t-1} + W_t$$

(13)

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

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

(14)

Where

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

And
3.1 The data

The monthly data of injuries to road accidents in Skikda from January 2001 to December 2016 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].

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 it 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

\[ \bar{X}_{t+1|t} = AX_{t|t} \], \[ \bar{X}_{t+1|t} = IE(X_{t|t+1}|Y_t) \] (20)

2- State covariance prediction

\[ P_{t+1|t} = AP_{t|t} + Q_t \], \[ P_{t+1|t} = Var(X_{t|t+1}|Y_t) \] (21)

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

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

The Kalman gain can be thought of as the weight given to the most recent observation for updating mean 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

\[ \bar{X}_{t+1|t+1} = \bar{X}_{t+1|t} + K_t e_t \], \[ e_t = X_t - \bar{X}_{t+1|t} \] (23)

5- Covariance

\[ P_{t+1|t+1} = (1 - K_{t+1|t} B) P_{t+1|t} \] (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

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

### 3.2 SARIMA modeling process

![Figure 1: The time plot of IRA](image)

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

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>p-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADF</td>
<td>-1.984</td>
</tr>
<tr>
<td>Critics</td>
<td>1%</td>
</tr>
<tr>
<td>Value</td>
<td>5%</td>
</tr>
</tbody>
</table>
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].

Table 5: The augmented Dickey-Fuller for SDDIRA

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>p-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADF</td>
<td>-13,436</td>
</tr>
<tr>
<td>Critics Value</td>
<td>1%</td>
</tr>
<tr>
<td>Value</td>
<td>5%</td>
</tr>
</tbody>
</table>

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), SARIMA(0,1,1)(0,1,1) and SARIMA(1,2,1,1)(0,1,1,1).

<table>
<thead>
<tr>
<th>Autocorrelation</th>
<th>Partial Correlation</th>
<th>AC</th>
<th>PAC</th>
<th>Q-Stat</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 6: Estimation of three models

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient</th>
<th>Prob</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>-0.3809</td>
<td>0.0000</td>
<td>0.0705</td>
<td>-5.3969</td>
</tr>
<tr>
<td>SAR(12)</td>
<td>0.6448</td>
<td>0.0000</td>
<td>0.0658</td>
<td>9.7887</td>
</tr>
<tr>
<td>MA(1)</td>
<td>-0.4470</td>
<td>0.0000</td>
<td>0.0671</td>
<td>-6.6523</td>
</tr>
<tr>
<td>SMA(12)</td>
<td>0.5324</td>
<td>0.0000</td>
<td>0.0649</td>
<td>8.1943</td>
</tr>
<tr>
<td>AR(12)</td>
<td>1.1097</td>
<td>0.0000</td>
<td>0.0151</td>
<td>73.2488</td>
</tr>
<tr>
<td>MA(1)</td>
<td>-0.4794</td>
<td>0.0000</td>
<td>0.0501</td>
<td>-14.9338</td>
</tr>
<tr>
<td>SMA(12)</td>
<td>-0.8979</td>
<td>0.0000</td>
<td>0.0245</td>
<td>-36.5924</td>
</tr>
</tbody>
</table>

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 (Swarz’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

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA(1,1,0)(1,1,0)</td>
<td>9.4387</td>
<td>9.4739</td>
</tr>
<tr>
<td>SARIMA(0,1,1)(0,1,1)</td>
<td>9.4859</td>
<td>9.5196</td>
</tr>
<tr>
<td>SARIMA(12,1,1)(0,1,1)</td>
<td>8.9221</td>
<td>8.9755</td>
</tr>
</tbody>
</table>

The chosen seasonal ARIMA is of the form:

\[
y_t = -1.10973y_{t-12} + 0.749499 \delta_{t-1} + 0.897996 \delta_{t-12} + \epsilon_t
\]

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).

### 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 Filter 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BJF</strong></td>
<td>78</td>
<td>80</td>
<td>94</td>
<td>94</td>
<td>106</td>
<td>143</td>
</tr>
<tr>
<td><strong>KFF</strong></td>
<td>72.75</td>
<td>71.87</td>
<td>129.17</td>
<td>148.61</td>
<td>134.53</td>
<td>145.59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BJF</strong></td>
<td>143</td>
<td>202</td>
<td>184</td>
<td>116</td>
<td>85</td>
<td>71</td>
</tr>
<tr>
<td><strong>KFF</strong></td>
<td>145.59</td>
<td>182.29</td>
<td>192.26</td>
<td>135.57</td>
<td>84.59</td>
<td>98.91</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>SARIMA Jenkins</th>
<th>Box SARIMA</th>
<th>SARIMA Kalman</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>1.161</td>
<td>0.187</td>
<td></td>
</tr>
<tr>
<td>MAPE</td>
<td>0.944</td>
<td>0.172</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>6.124</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td>Theil’s U</td>
<td>0.035</td>
<td>0.005</td>
<td></td>
</tr>
</tbody>
</table>

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


Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US