International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-5, Issue-3, July 2016

Statistical Analysis of Traffic Accidents time series in Egypt Using Classical Methods and Box and Jenkins Models

Abeer S. Mohamed

***Abstract*— The Trafficaccidentfrequencyhasbeenincreasing inEgyptinthe recentyears for many reasons (human, place, and time). This paper aims to find the best model for the annual traffic accidents statistics in Egypt from 2005 to 2015 and make a prediction of the number of annual traffic accidents likely to occur in future. The analysis of time series data of traffic accidents is presented using the classical statistical methods and Box-Jenkins methodology to build ARIMA model.**

***Index Terms*— Time series analysis, classical statistical methods, forecasting, ARIMA, traffic accident.**

I. INTRODUCTION

Victims due to traffic accidents are more than 5000 of death and 22000 injures with different hurts, annually. Economical loses are 2 % from national total income according to data of Egyptian society for protection from traffic accidents. Traffic accidents are considered the second reason for death in Egypt and 80 % of victims are between 15 and 45 years old (Ali (2009)). For that, this paper analyzes and predicts the future traffic accidents using statistics methods.

Time series models have been the basis for process behavior studies or metrics over a period of time. There are many application areas of time series models such as sales forecasting, weather forecasting, and inventory studies. In decisions that involve factor of uncertainty of the future, time series models have been found one of the most effective methods of forecasting (Makridakis et al, 1998) .

Time series data often have time-dependent moments (e.g. mean, variance, skewness, kurtosis).The mean or variance of many time series increases over time.This is a property is called nonstationarity. Astationary time series have mean, variance, and autocorrelation function that are essentially constant through time.

Among the most important models of time series analysis is the model of ARIMA which has been introduced by Box and Jenkins.The Box and Jenkins model assumes that the time series is stationary. For nonstationary time series, Box and Jenkins recommend differencing of one or more time series to achieve stationarity. This produces an ARIMA model (auto regressive (AR), Integrated (I) and the moving average (MA)).

Ali (2009) decided three main factors (human, place and time) which have the most effect on the traffic accidents in Egypt. Momani (2009) presented the time series analysis rainfall data in Jordanand studied the Box-Jenkins methodology to build ARIMA model for monthly rainfall data taken for Amman airport station for the period from 1922-1999 with a total of 936 readings. ARIMA (1, 0, 0) (0, 1,1)$_{12}$ model was developed.

Tularam and Mahbub (2010) examined a large data set involving more than 50 years of rainfall and temperature where spectral analysis and time series analysis-ARIMA methodology were used to analyze climatic trends and interactions.

Balogun etal (2014) analyzed a data set collected from Nigerian traffic accidents using time series approach. The data collected spanned the period between 1989 to 2008. They found that the best model was AR (1) for annual data.

Mutangi (2015) analyzed the data of traffic accidents in Zimbabwe by three ARIMA models which were suggested based on the ACF and PACF plots of the differenced series. These were ARIMA(0,1,0), ARIMA(1,1,0) and ARIMA(1,1,1) and he decided that ARIMA (0,1,0) was the best model for the Zimbabwe annual Traffic Accidents data.

II. THE TIME SERIES ANALYSIS

A time series is a sequential set of data points, measured typically over successive times. It is mathematically defined as a set of vectors y_t, t=0,1,2, … where, the subscript t is the time point at which y is observed (Pankratz (1983)). According to Chatfield (1987) time series is a collection of observation segmental in time at regular intervals. There are four factors affecting time series observations: the trend effect, the seasonal effect, cyclical effect and random variation. The majority of the time series contains a trend effect either increasing or decreasing, therefore it's the most important effect that must be studied when analyzing the time series. This analysis can be done by several methods such as the classic approaches, least square method (OLS) , matrices, semi average, quadratic trend model and moving average method (MA) .

The Box-Jenkins methodology which is known by ARIMA models will be introduced, ARIMA model as a non-stationary time series model is made stationary by applying finite differencing of the data points. The mathematical formulation of the ARIMA has (*p*,*d*,*q*) form where *p*, *d* and *q* are integers greater than or equal to zero and refer to the order of the autoregressive (*p*), the order of difference (*d*), and the order of moving average (*q*) parts of the model .The integer *d* controls the level of differencing where *d equal 1* is generally enough in most cases. When *d* is equal to 0, then it reduces to an ARMA(*p*,*q*) models. The linear regression is estimated as it follows.

$$y_t = \beta_o + \beta_1 t + \varepsilon_t \qquad (2.1)$$

Where $y_t, \beta_o, \beta_1, and\, \varepsilon_t$ are the current observation, constant of regression line, the regression coefficient and the

Abeer S. Mohamed, Department of Statistics, Faculty of Commerce (Girls' Branch), Al-Azhar University

random errors which satisfy independent identical distribution (i.i.d) (normal distribution with mean equal zero and constant variance) respectively. Then the estimate of equation (2.1) can be written as shown in equation (2.2).

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 t \qquad (2.2)$$

III. The Data Set

The time series data set (shown in Table 1 and figure 1) presents the number of the traffic accidents according to the Central Agency for Public Mobilization and Statistics (CAPMAS) in Egypt.

Table 1: The Traffic Accidents from 2005 to 2015

Year	Accident Numbers
2005	21352
2006	18061
2007	22900
2008	20938
2009	22793
2010	24371
2011	16830
2012	15516
2013	15578
2014	14403
2015	14548

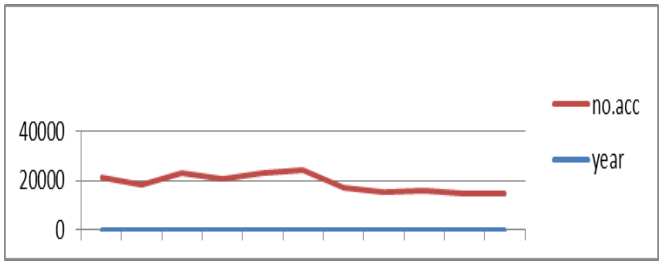

Figure 1: The observed values

IV. The Classical Methods

This section presnts the classical method to analyze the data set (in table 1) and presents the method accuracy by using some accuracy measures such as the mean absolute deviation MAD,MAPE and mean square error MSE.The classical time series analysis method decomposes the time series function $y_t = f(t)$ into up to four componentsMcClave and Synch(2001).

1.Trend: a long-term monotonic change of the average level of the time series.

2.The Trade Cycle: a long wave in the time series.

3.The Seasonal Component: fluctuations in time series that recur during specific time periods.

4.The Residual component: the influences on the time series that are not explained by the other three components.

4.1 the least square method OLS

The equation (2.2) is solved using the OLS method as it follows.

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} t\,y_t - \sum_{i=1}^{n} t \sum_{i=1}^{n} y_t}{n\sum_{i=1}^{n} t^2 - (\sum_{i=1}^{n} t)^2} \qquad (4.1)$$

$$\hat{\beta}_o = \bar{y}_t - \hat{\beta}_1 \bar{t},$$

$$where\, \bar{y}_t = \frac{\sum_{i=1}^{n} y_t}{n} \; and \; \bar{t} = \frac{\sum_{i=1}^{n} t}{n} \qquad (4.2)$$

By solving (4.1) and (4.2), the regression line can be written as it follows.

$$\hat{y}_t = 236131 - 794.77t \qquad (4.3)$$

The predictive values are shown in Table (2), the graph of predictive and observed data is shown in Figure (2), and the statistical analysis is presented in tables 3,4 and 5.

Table 2: The Predictive Values using OLS

Year(t)	y_t	$\hat{y}_t$
1(2005)	21352	22818.66
2(2006)	18061	22024.22
3(2007)	22900	21229.78
4(2008)	20938	20435.34
5(2009)	22793	19640.9
6(2010)	24371	18846.46
7(2011)	16830	18052.02
8(2012)	15516	17257.58
9(2013)	15578	16463.14
10(2014)	14403	15668.7
11(2015)	14548	14874.26
12(2016)		14075.86

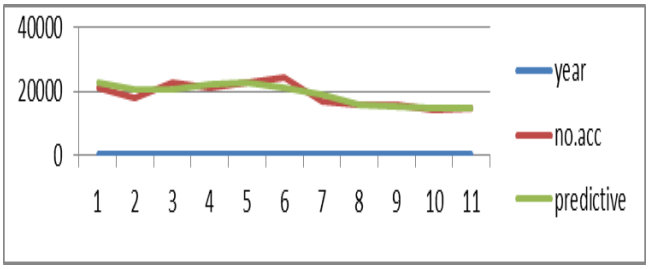

Figure 2: The observation and predictive values using OLS method

Table 3: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.710[a]	.504	.449	2756.30749

a. Predictors: (Constant), year

Table 4: ANOVA[a]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	69483005.682	1	69483005.682	9.146	.014[b]
	Residual	68375079.045	9	7597231.005		
	Total	137858084.727	10			

a. Dependent Variable: number of accident

b. Predictors: (Constant), year

Table 5: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23613.182	1782.421		13.248	.000
	year	-794.773	262.804	-.710	-3.024	.014

a. Dependent Variable: number of accidents

$$MAD = \frac{1}{n-2}\left|\sum_{i=1}^{n} e_t\right| = 2413.34 \quad (4.4)$$

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n} e_t^2 = 7597236 \quad (4.5)$$

$$MAPE = \frac{1}{n-2}\sum_{i=1}^{n}\left|\frac{e_t}{y_i}\right| = 11 \quad (4.6)$$

Where $e_t = y_t - \hat{y}_t$

4.2 The Matrices Method

To solve the equation (2.2) using matrices, we can rewrite (2.2) as it follows.

$$\underline{Y_t} = \underline{X}_t \underline{\hat{\beta}_t} \quad (4.7)$$

where

$$\underline{\hat{\beta}_t} = \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix} = \left[X^T X\right]^{-1} X^T y =$$

$$= \frac{1}{n\sum_{t=1}^{n} t^2 - (\sum_{t=1}^{n} t)^2} \begin{bmatrix} \sum_{t=1}^{n} t^2 & -\sum_{t=1}^{n} t \\ -\sum_{t=1}^{n} t & n \end{bmatrix} \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} t y_t \end{bmatrix}$$

(4.8)

$$\underline{\hat{\beta}_t} = \begin{bmatrix} 23613.2 \\ -794.76 \end{bmatrix},$$

Then

$$\hat{y}_t = 23613.1 - 794.77t \quad (4.9)$$

This method gives approximately the same resultsas that of OLS method.

4.3 The Moving Average (MA) Method

Moving average(MA) method is one of widely known technical indicators used to predict the future data in time series analysis. The moving average is extremely useful for forecasting long-term trendswhere the average represents the "middling" value of a set of numbers. First of all, the period of the moving average has to be decided and for this data set the period of 3 values is estimated for the observations shown in Tables5 and 6 and Figure 3.

Table 6: The Predictive Values using MA method

y_t	Moving summation (MS)	$\hat{y}_t = \frac{MS}{3(MA\ period)}$
21352		
18061	62313	20771
22900	61899	20633
20938	66631	22210.33
22793	68102	22700.67
24371	63994	21331.67
16830	56717	18905.67
15516	47924	15974.67
15578	45497	15165.67
14403	44529	14843
14548		

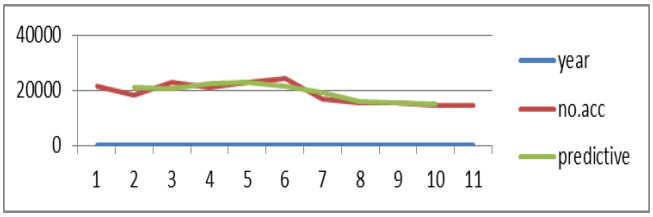

Figure 3: The observation and predictive values using MA method

$$MAD = \frac{1}{n-2}\left|\sum_{i=1}^{n} e_t\right| = 1418.62 \quad (4.10)$$

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n} e_t^2 = 3136740.44 \quad (4.11)$$

$$MAPE = \frac{1}{n-2}\sum_{i=1}^{n}\left|\frac{e_t}{y_i}\right| = 16 \quad (4.12)$$

4.4 The Semi Average Method

To solve the equation (2.2) using the semi average method, time series data is divided into two equations to be solvedto find the parameters of (2.2). Because the data have odd observations, then we must delete the observation number six as it follows.

$$\bar{y}_1 = \beta_o - \beta_1 \bar{t}_1 \quad (4.13)$$

$$\bar{y}_1 = 21208.8, \quad \bar{t}_1 = 3$$

$$\bar{y}_2 = \beta_o - \beta_2 \bar{t}_2 \quad (4.14)$$

$$\bar{y}_1 = 15375, \quad \bar{t}_1 = 9$$

Then

$$\hat{y}_t = 24125.7 - 972.3t \quad (4.15)$$

The predictive values and real data are illustrated in table (7) and Figure (4)

Table 7:The Predictive Values using semi average

year	y_t	$\hat{y}_t$
1	21352	23153.4
2	18061	22181.1
3	22900	21208.8
4	20938	20236.5
5	22793	19264.2
6	24371	18291.9
7	16830	17319.6
8	15516	16347.3
9	15578	15375
10	14403	14402.7
11	14548	13430.4

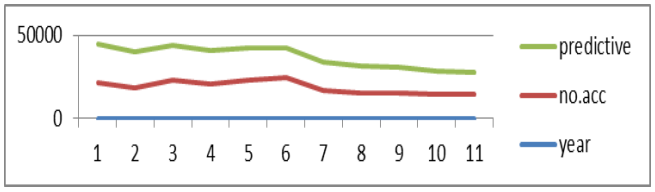

Figure (4): The observation and predictive values using semi average method

$$MAD = \frac{1}{n-2}\left|\sum_{i=1}^{n} e_t\right| = 6754.4$$

(4.16)

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n} e_t^2 = 835571322$$

(4.17)

$$MAPE = \frac{1}{n-2}\sum_{i=1}^{n}\left|\frac{e_t}{y_i}\right| = 24.4$$

(4.18)

4.5 The quadratic trend model

In some cases, a linear trend is inadequate to capture the trend of a time series. A natural generalization of the linear trend model is the polynomial trend model as in equation (4.19).

$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_p t^p$ (4.19)

where p is a positive integer.

The quadratic linear trend model is a special case of the polynomial trend model (p=1), (for economic time series we almost never require p > 2). Then the equation (4.19) can be written as it follows.

$y_t = \beta_0 + \beta_1 t + \beta_2 t^2$(4.20)

By using Minitap 17, the equationis estimated as it follows.

y_t= 20129+774 t+126.8t^2

(4.21)

The equation (4.21), the predictive and actual values are illustrated in Figure (5).

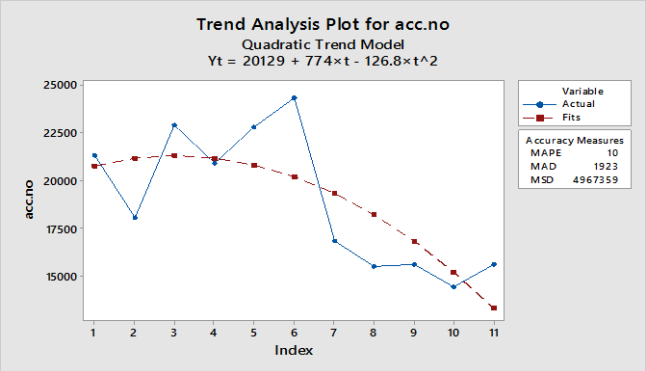

Figure (5): The observation and predictive values using quadratic trend model

$$MAD = \frac{1}{n-2}\left|\sum_{i=1}^{n} e_t\right| = 1923$$

(4.22)

$$MAPE = \frac{1}{n-2}\sum_{i=1}^{n}\left|\frac{e_t}{y_i}\right| = 10 \qquad (4.23)$$

V. The Box and Jenkins Methodology

Box - Jenkins analysis refers to a systematic method of identifying, fitting, checking, and forecasting. Identification determines the appropriate values of p, d, & q using the ACF, PACF, and unit root tests (p is the AR order, d is the integration order, q is the MA order).Estimation estimates an ARIMA model using values of p, d, & q. Diagnostic checking checks residuals of estimated ARIMA model(s) to check if they are white noise.Forecasting produces sample forecasts or set aside last few data points for in-sample forecasting.The Box-Jenkins model assumes that the time series is stationary and it recommends differencing non-stationary series one or more times to achieve stationarity (Box et al (1994)).

Using integrated autoregressive, moving average (ARIMA) time series model is appropriate for time series of medium to long length.

$$y_t = \mu + \phi_1 y_{t-1} + \ldots \phi_p y_{t-p} - \theta_1 y_{t-1} - \ldots \theta_q y_{t-q} - \varepsilon_t$$

ϕ_p *parameterof AR*, θ_q *parameterof MA, and*

the rondomerror $\varepsilon_t \sim (0, \sigma^2)$

In this section the four steps (identification, estimation, checking and forecasting)
are discussed.

5.1 Autocorrelation Function (ACF)

Autocorrelation Function (ACF) computes the correlation between different lags of a series. The ACF (ρ_k) represents the degree of persistence over respective lags of a variable. The autocorrelation function is presented in Figure (6)

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\sum_{i=1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{i=1}^{n}(y_t - \bar{y})^2}$$

(5.2)

ACF (0) = 1, ACF (k) = ACF (-k)

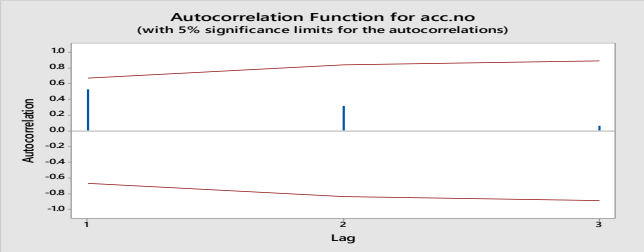

Figure (6): The Autocorrelation Function

Figure (6) proved that the ACF decreases quickly after one lag which means thatthe model AR becomes stationary after one difference.

5.2The Partial Autocorrelation Function (PACF)

The Partial Autocorrelation Function (PACF) expresses information useful in determining the order of an ARIMA model (Box et al(1994)). PACF coefficient θ_{kk} is the correlation between $y_t$and y_{t-k} after omitting the effect of $y_{t-1},\ldots,y_{t-k-1}$. The lag k partial autocorrelation is the partial regression coefficient, in the k^{th}order autoregression.

$$y_t = \theta_{k1}y_{t-1} + \theta_{k2}y_{t-2} + \ldots + \theta_{kk}y_{t-k} + \varepsilon_t \quad (5.3)$$

PACF is represented in Figure (7).

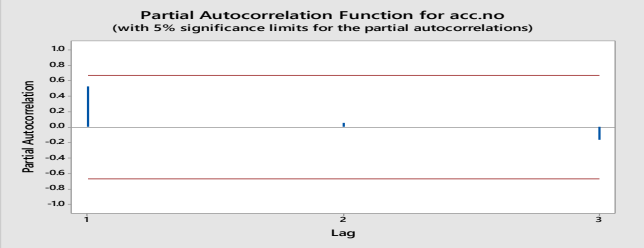

Figure (7): The Partial Autocorrelation Function

Figure (7) shows that the PACF decreases quickly after one lag. This means that the model MA is of order one. From ACF and PACF, we can judge that the number of traffic accidents appropriate model is ARIMA(1,1,1) and this concludes **identification step**.

$$y_t = \mu + \phi_1 y_{t-1} - \theta_1 y_{t-1} - \varepsilon_t \quad (5.4)$$

Then the parameters will be estimated and the estimation using spss 21is presented in Table 8 and 9. The residual statistic is also presented in Table 9.

Table 8 : Parameter estimation

Type	Coef	SE Coef	T	P
AR1	0.2315	0.4662	0.50	0.635
MA1	0.9088	03957	2.30	0.055
constant	-576.7	166.6	-3.46	0.011

$$y_t = -576.7 + 0.2315y_{t-1} - 0.9088_1 y_{t-1} \quad (5.5)$$

Table 9:The Model Fit

Fit Statistic	Mean
Stationary R-squared	0.265
R-squared	0.408
RMSE	3491.332
MAPE	12.111
MSE	10282406
MAE	2317.429
MaxAE	4044.052
Normalized BIC	17.237

Thediagnostic step depends on ACF and PACF of residuals of the data which are illustrated in Figure 9 and 10.

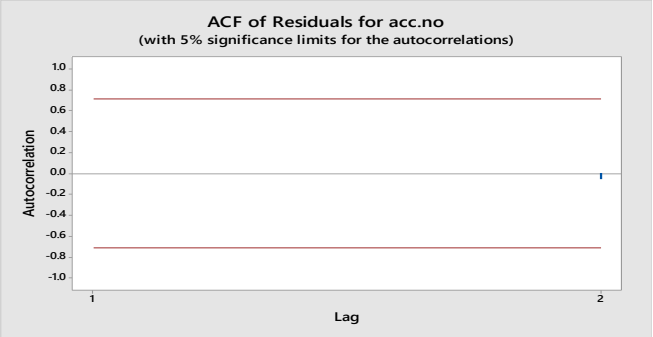

Figure (8)autocorrelation for the residuals of model ARIMA(1,1,1)

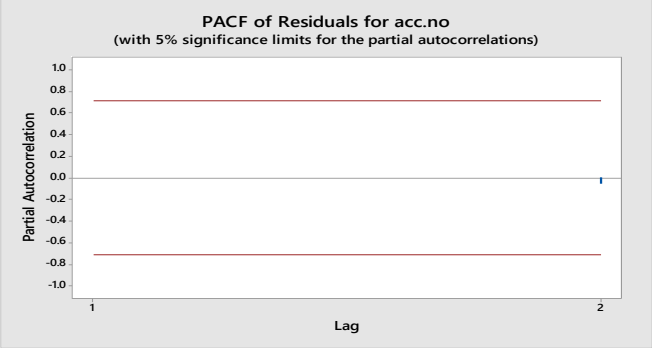

Figure9: partial autocorrelation for the residuals of model ARIMA(1,1,1)

Table10: Model statistics

Model	Model Fit statistics		Ljung-Box Q(18)
acc.no-Model_1 1	Stationary R-squared	MAE	DF . 0
	0.275	2340.675	

Figure (8), (9) and Table (10) illustrate the residuals of the model following random pattern. There is no correlation between the random errors (from Ljung-Box) which means that the model represents the data. This concludes the Forecastingfinal step where the current data will be introduced in Table 11 and Figure 10.

Table 11:The Predictive Values using ARIMA(1,1,1) ; 95% limits.

Period	Forecast	Lower	Upper	Actual
2008	20165.1	13878.9	26451.4	20938.0
2009	18955.3	12350.0	25560.6	22793.0
2010	18098.5	11411.4	24785.6	24371.0

2011	17323.4	10586.	24059.9	16830.0
2012	16567.3	9787.8	23346.7	15516.0
2013	15815.5	8994.8	22636.2	15578.0
2014	15064.7	8203.3	21926.2	14403.0
2015	14314.2	7412.3	21216.1	15578.0
2016	13563.7	6621.7	20505.8	
2017	12813.3	5831.3	19795.3	
2018	12062.8	5041.1	19084.5	

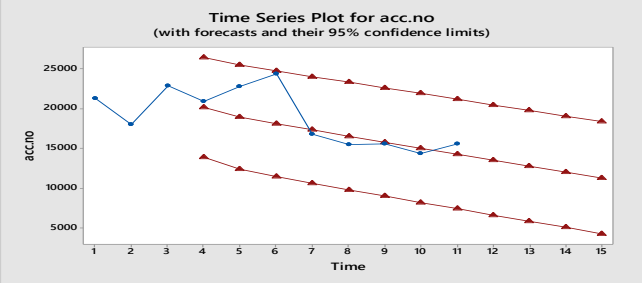

Figure (10):The observation and predictive values using ARIMA(1,1,1)

VI. The Conclusions

Ranging global rate of road death toll per 10 thousand vehicles, between 10 and 12 dead, but he arrives in Egypt to 25 people, twice the world average, and also has a death toll of road accidents per 100 km in Egypt, 131 people were killed, while the global average ranges between 4 and 20 people, which means that the rate in Egypt more than 30 times the global average, and also the cruelty of the incident tells us that Egypt is happening with 22 people per 100 wounded, while the global average 3 deaths per 100 injured.

Therefore this paper has presented the results of the statistical analysis of traffic accidents data in Egypt and also many models. The traffic accidents data was statistically analyzed by classical method and Box and Jenkins method. Among the classical methods quadratic trend linear method was the best because of its least values of accuracy measures (MAPE, MAD, and MSE). Box and Jenkins was the best in representing the time series data. Because the classical method treats the regression relationship as deterministic, it is very sensitive to any data update. Time series have many stochastic trends and the Box and Jenkins model can be modified to accommodate any data update it can be checked.

Reference

[1] Ali, S. A (2009). Traffic accident in Egypt -2 factorsaffect on traffic accident in Egypt. Journal of Engineering Sciences, Assiut University, Vol. 37, No. 2, pp.483-505, March 2009.

[2] Box,G. E. P G. Jenkins,M. and Reinsel.G. C. (1994).*Time Series Analysis, Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Clifs, NJ.

[3] Balogun ,O.S. Awoeyo, O.O and Dawodu, O.O.(2014).Use of time series analysis of road data in Lagos state. International Journal of Advanced Research (2014), Volume 2, Issue 5, 1045-1059.

[4] Chatfield .C. (1987). *The Analysis of Time Series. An Introduction*. London Chapman and Hall (Third Edition).

[5] McClave,J.T. and Synch, T.(2001). Statistics for Business and Economics. Prentice Hall..

[6] Mutangi, K.(2015). Time Series Analysis of Road Traffic Accidents in Zimbabwe. International Journal of Statistics and Applications 2015, 5(4): 141-149

[7] Momani, M and Nill,P.N (2009). Time Series Analysis Model for Rainfall Data in Jordan:Case Study for Using Time Series Analysis. American Journal of Environmental Sciences 5 (5): 599-604

[8] Pankratz, A (1983).*forecasting with univariat Box-Jenkins models concepts and cases*. Jon Willy&sons.New York.

[9] Makridakis,S. G, Wheelwright,S. C. and Hyndman ,R J. (1998). *Forecasting: Methods And Applications*. Jon Willy & Sons, New York.

[10] Tularam, G.A. and Mahbub, I.(2010). Time Series Analysis of Rainfall and Temperature Interactions in Coastal Catchments. Journal of Mathematics and Statistics 6 (3): 372-380

Abeer Sayed is a assistant professor in the faculty of Commerce, Statistics Department, AlAzhar University. She got her Doctor of philosophy (Ph.D.), Master of Science (M.Sc.), and B.Sc. from the same department. Her area of interest includes statistics, data analytics, Big data, and data mining.