International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 24 No. 4 Nov. 2018, pp. 1544-1553

http://www.ijias.issr-journals.org/

Temporal Analysis of Road Accidents by ARIMA Model : Case of Tunisia

Aymen GHÉDIRA[1-2], Karim KAMMOUN[3], and Chaker BEN SAAD[1]

[1]University of Sousse, Higher Institute of Transport and Logistics, Tunisia

[2]University of Tunis, School of Economics and Commerce, DEFI Laboratory, Tunisia

[3]University of Sfax, the Higher Institute of Industrial Management of Sfax, Tunisia

ABSTRACT: Tunisian policy for road safety is neither clear nor reliable. Each road organism acts on own its side without any coordination with other stakeholders and therefore without achieving results. This paper aims to facilitate the decision-making process on road safety in Tunisia through the time series analysis of road accidents. The analysis work will allow identifying better the respective weight of the factors associated with the road accident frequency. Methodologically, ARIMA (Auto-Regressive Integrated Moving Average) model is used to meet our above goal. Moreover, the methodology of Box Jenkins intervenes as a statistical solution to solve the problem of time series analysis. The results show that the time series of accidents are mainly characterized by two different periods in terms of trend. A low decrease in the number of accidents before the revolution (2011) (between 2007 and the end of 2010) and the irregular evolution in the rest of the series. Then, models are developed in accident cases and ARIMA (0, 1, 2) is identified as the best model. A three-year forecast is made using the best model and it shows that the number of road accidents would decrease due to several factors in Tunisia. Consequently, this study shows that the temporal analysis of the time series of road accidents can attract the attention of decision-makers to the importance of the application of key road safety measures in the short, medium and long term as well as the nature of the relationship between the different decision-makers horizons.

KEYWORDS: Road safety; Time series analysis; Road accidents; ARIMA.

1 INTRODUCTION

Road traffic accidents represent a global problem, which has attracted an increasing interest because of its social and economic consequences. Worldwide, the number of people killed each year in road accidents is estimated at nearly 1.25 million, while the number of injured could reach 50 million (World Health Organization, 2015) [1]. In Tunisia, citizens always live a dramatic situation. Indeed, with a daily average of more than 21 accidents causing more than 32 injured and over 4 dead (National Road Safety Observatory of Tunisia), and with socioeconomic costs, which represent nearly 190 million dollars annually (World Health Organization, 2013), it is not surprising that more attention and resources should be devoted by decision-makers to this major problem.

Nowadays, the National Road Safety Council meets once a year to make recommendations and to contribute to one way or another in the road safety improvement process. But nothing is done due to a lack of a national policy for road safety. Thereafter, each actor acts on his side without any achieving coordination and therefore without real results. The aim of this article is to analyze temporally a time series of road accidents and provides a future number of observations in order to form a judgment on the current and future situation of road safety and consequently make easier the process of developing a national policy.

In this article, we have applied an ARIMA model proposed by Box and David (1970) to reconstruct the pattern of processes subjected to random shocks over time: between two successive observations in a series of measurements concerning the

Corresponding Author: Karim KAMMOUN

activity of the process, a random event called perturbation affects the temporal pattern of this process and thus modifies the values of the observations. To do this, Box and Jenkins' methodology is applied using studio software to construct an ARIMA model that best describes the pattern of a time series using the following procedure: identification, estimation, diagnosis and prediction.

In this paper, our first goal is to examine related research studies; secondly, we will present a recent data to use and materials. Moreover, we will discuss some models and applied methodologies. Finally, the results are analyzing.

2 LITERATURE REVIEW

Progress in road safety is a primary objective for several countries, and as a result, a very high number of scientific works are being done just to analyze and improve the state of road safety. On the other hand, these road safety studies are different, in terms of the nature of the problem as well as the data used, and the approach followed. Indeed, concerning the temporal analysis of accidents, various works are involved; whose time series analysis is the most used approach.

2.1 RELATED WORKS

The temporal analysis of road accidents or fatalities is based mainly on time series and easily treated, measurement tools, to relate the number of observations, which are evaluated according to road traffic, demand for daily travel, climatic conditions and other factors. Indeed, the literature has many works such as (Christens, 2003; Yuan, and al., 2013) which examine the factors responsible for the variation of road accidents according to the years. The same for the effects of strategies and the economic evolution on the occurrence of accidents and the number of deaths per month, there are also various studies such as (Lassarre, 2001; Levine, and al., 1995). The most used type of time series in temporal analysis examines the annual or monthly number of accidents and victims in some countries. The objective of the analysis is to describe accident phenomenology qualitatively and quantitatively, that one may help us to understand and predict the variation in the road safety situation. Researchers are trying to identify a more appropriate technique for analyzing time series. Consequently, it becomes possible to have a vision for the future to predict the frequency of accidents (Basic, and al., 2010).

2.2 THE DYNAMIC UNIVARIATE MODELS

According to European Commission COST Action 329 (2004), there are two types of much answered dynamic univaried models: the structural models by Harvey and the ARIMA models by Box and Jenkins. Concerning the structural models, the applications of these models are very diverse in terms of problems and variables. Historically, the first test was made by Harvey and Durbins in 1986. It's showed the consequences of the introduction of the seat belt law in the United Kingdom. The second test of this model was used by Ernst and Bruning in 1990 dealing with the seat belt in West Germany. In addition, the diversity of scientific works related to the analysis of an explanatory factors influence on the level of road safety where structural models of the time series are used. Other applications are distinguished in road safety analysis based on this method by Lassarre (2001), Scuffham and Langley (2002), and was applied by Christens (2003), Gould (2005) and Van den Bossche (2006). For the second dynamic uni-varied model, ARIMA model shows a case of ARMA model. According to the literature, ARIMA is developed by Box and Tiao (1975) and is generally used extensively to model or predict time series data. Looking at applications in road safety, many scientific works can be referred to, for example Wagenaar (1984) and Van den Bossche, and others. Moreover, in terms of advantages, the approach of this model represents a flexible form for the identification and estimation of the parameters for a studied series, and the prediction of future observations.

2.3 THE FORECAST IN ACCIDENTOLOGY

About accident prediction, the literature has a variety of works whose used models usually depend on the availability of the data and the nature of the problem. Indeed, a separation has been translated in the choice of models that analyze the effects of some factors according to observations of accidents and forecasting models of future observations. The use of ARIMA model is an example to predict the number of road fatalities in Malaysia between 2015 and 2020 (Rohayu, and al., 2012), the same model was used but this time to predict the number of accidents in China (Yuan and al., 2013).

3 METHODOLOGY

3.1 DATA AND MATERIAL

This study is based mainly on a time series that aims to analyze the evolution of road accidents as a function of time. The data used is derived from the National Road Safety Observatory (NRSO) database, which enables to formulate time series of road accidents. NRSO is the only responsible source that has a set of time series based on the number of accidents, deaths and injuries, plus a set of potential explanatory variables. About data, we have used the monthly history of road accidents which extends over 9-years from 2007 to the end of 2015. The choice of this period is mainly related to the availability of data proposed by the NRSO. This analysis has been converted into computer language by RStudio software. It is an integrated development environment, functional, free and multiplatform, whose programming language is used for statistical analysis.

3.2 CONCEPTS AND STATIONARITY PROCESS

The basic notion that intervenes when modeling a time series is that of stationarity. Consequently, it becomes very important to study the set of stochastic characteristics of the time series before running statistical tests. So, it is necessary to examine the stability of variance and mean of the time series over time. A time series is stationary if there is no systematic change in mean (no trend) (1), variances and strictly periodic variations (2) and (3) have been removed. Then, a stationary temporal series has no tendency or seasonality the statistical parameters of the series do not change over time. Mathematically, a series is said to be stationary if:

$$E(Y_t) = \text{constante pour tous } t. \quad (1)$$

$$Var(Y_t) = \text{constante pour tous } t. \quad (2)$$

$$Cov(Y_t - Y_{t+k}) = \text{constante pour tous } t. \quad (3)$$

On the one hand, if the characteristics of a time series have a change over time, it is assumed to be non-stationary. So, we must achieve the stationary by the differentiation process before analyzing or modeling the time series. Consequently, the differenced data must be characterized by less variation around the systematic mean of the series compared to the original series. Generally, for non-seasonal data, first order differencing operation may be sufficient to attain the stationarity (4). On the other hand, other types of the series require order differencing operation (5) or more:

$$\nabla y_{t+1} = y_{t+1} - y_t \quad (4)$$

$$\nabla^2 y_{t+2} = \nabla y_{t+2} - \nabla y_{t+1} - 2y_{t+1} + y_t \quad (5)$$

3.3 ARIMA (P, D, Q) MODEL

In the analysis of a time series, mainly two categories of models can be distinguished, the first category assumes that the data used are a function of time, whereas the models of the second category estimate the values of data used depending on the data values that precede them. The second category then has an ARIMA model (AutoRegressive Integrated Moving Average). The ARIMA model has been formalized and popularized by Box and Jenkins in 1976. ARIMA (p, d, q) combines three types of temporal processes: The autoregressive process (AR), which considers that each value in a time series can be determined by the weighted sum of a set of previous values, with a random error term; The integrated process (I) which considers that each value exposes a constant difference with the previous value; The moving average (MA) which considers that each value can be determined as a function of the errors related to the previous value, with a more error term.

For model parameters, "p" is the order of the autoregressive process AR (p), "d" the order of differencing or degree of integration process I (d) and "q" the order of Moving average MA (q). In general, the ARIMA model (p, d, q) follows the formula below.

$$y_t = \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \mathcal{E}_{t-j} + \mu + \mathcal{E}_t \quad (6)$$

Where autoregression coefficient "ϕ" expresses the strength of the linear connection between two successive values (AR (p) refers to as "y_t" depends on the "p" previous values), the term "θ" is moving average coefficient multiplied by a past random error "$\mathcal{E}$".

3.4 BOX AND JENKINS METHODOLOGY

In practice, it is very difficult to use equation 6 in the case where the number of lag "p" or "q" is large. Then, Box and Jenkins methodology intervenes as a statistical solution to solve the problem of time series analysis. It presents a forecasting method which implements knowledge about autocorrelation analysis while testing a set of ARIMA models to identify the appropriate model for a time series and to make the forecast. This method has four main stages: model identification, parameter estimation, diagnostics checking and forecasting.

4 ANALYSIS AND RESULTS

4.1 DESCRIPTIVE ANALYSIS

According to Fig. 1, the time series plot shows a monthly road accident cases in Tunisia from 2007 to 2015. The time plot exhibits a variation in the mean of the series which gives evidence of trend in the time series and consequently the absence of stationarity. In general, accident cases are characterized by an incoherent decrease from 2007 to the end of 2010 (December 2010). In January 2011, a minimal peak of 384 accident cases appeared which was mainly caused by the Tunisian revolution (interruption of travel demand). But the number of accidents increases in February 2011 and it remained irregular until the end of the time series (The data in the series do not show a seasonal pattern) except in January 2015, a maximum peak of 1233 accident cases appeared which was mainly due to the exceptional climatic conditions (a cold wave that hit the country) that increased risk of accidents.

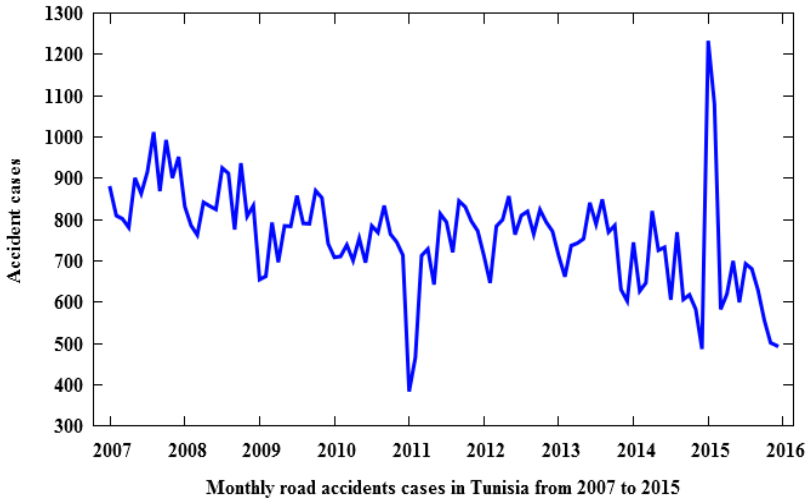

Fig. 1. Time plot of monthly road accident cases in Tunisia from 2007 to 2015

4.2 STUDY OF STATIONARITY

Studying stationarity is a fundamental phase before the analysis of the original time series. The existence of a trend requires the need to apply differencing to achieve stationarity to stabilize the time series. Fig. 2 shows a first differencing time series which was performed to remove trend component and replaced the original time series. Then, a stationarity test (the Dickey-Fuller test and the increased Dickey-Fuller test) was executed for the differenced time series to detect the existence of stationarity. The stationarity test can be exceeded, and a second differencing can be performed. So, an autocorrelation function for this last time series differencing treatment should be done allowing to detect any form of non-stationarity or over-differencing.

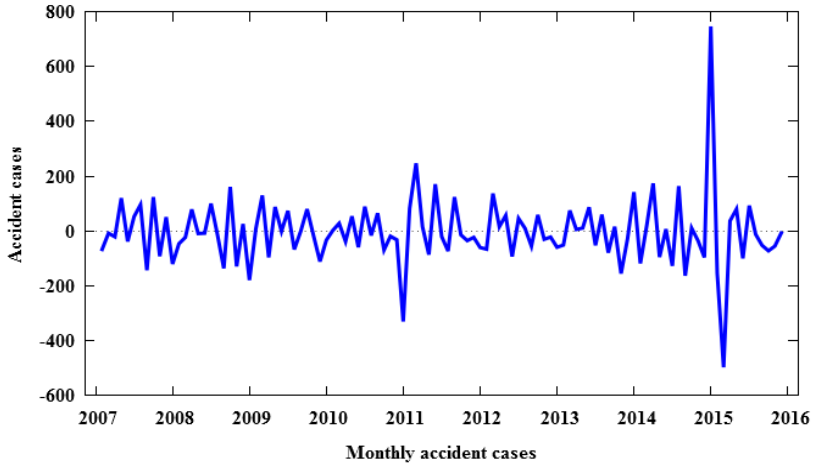

Fig. 2. First differencing of monthly accident cases in Tunisia

According to Fig. 3, the autocorrelation function has a negative peak of -0.5 at the first lag (it is a sign of over-differencing). Then, there is no need to make a second differencing and consequently the order of differencing "d" is 1.

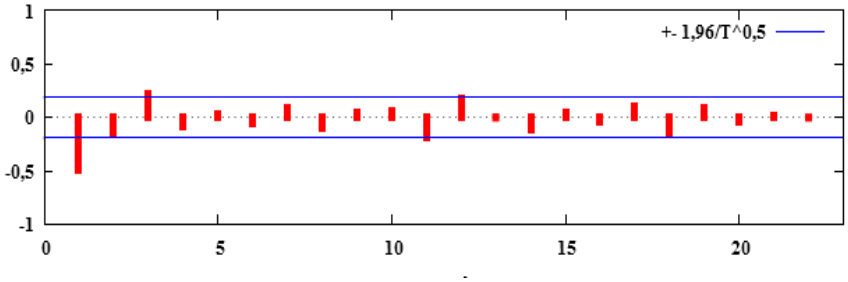

Fig. 3. ACF plots of the second differencing of monthly accident cases in Tunisia

4.3 MODEL IDENTIFICATION

After checking stationarity, in this step we have identified the parameters "p" of autoregressive process "AR" and "q" of moving average process "MA" necessary to correct the residual autocorrelations. The latter are the correlations of the time series according to a Lag (except for the lag autocorrelation 0 which is equal to 1). We used for this step the first differencing time series. ARIMA model (p, d, q) identification is based on the autocorrelation function (ACF) test for the determination of "p" and the partial autocorrelation function (PACF) for determining "q". In general, if ACF plot of the first differencing has only its first terms "q <= 3" and different from zero, we can consider an MA (q). Fig. 4 shows that the autocorrelation function (ACF) of the first differencing time series has only two significant negative peaks that exceed the error limits. Then, ARIMA model has a moving average process of order 2 or MA (2).

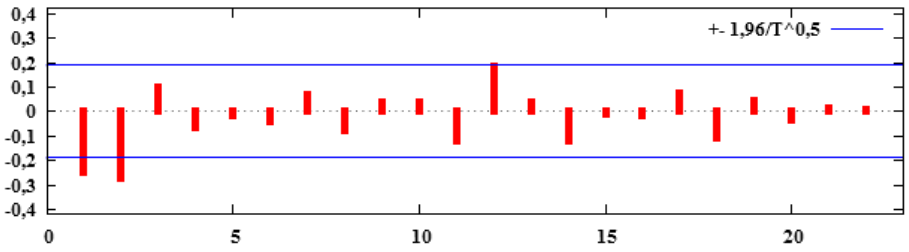

Fig. 4. ACF plots of the first differencing of monthly accident cases in Tunisia

For the determination of the autoregressive process order "p", if the PACF plot of the first differencing has only its first terms "p <= 3" and different from zero, we can consider an AR (p). Fig. 5 shows that the partial autocorrelation function (PACF) of the first differencing time series has only two significant negative peaks that exceed the error limits. Then, ARIMA model has an autoregressive process of order 2 or AR (2).

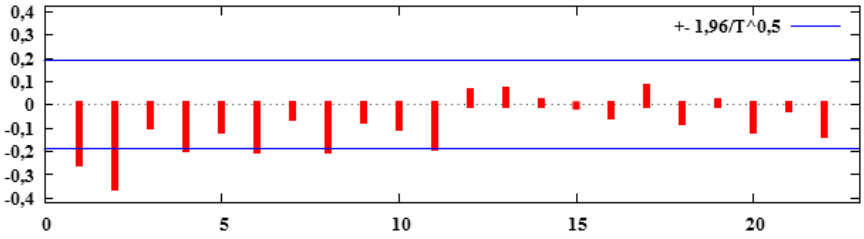

Fig. 5. PACF plots of the first differencing of monthly accident cases in Tunisia

As a result, a set of ARIMA models have been proposed: ARIMA (2, 1, 2), ARIMA (2, 1, 0), ARIMA (0, 1, 2).

4.4 ESTIMATING MODEL PARAMETERS

This part is intended to estimate the autoregressive parameter (p) and moving average parameter (q), which aims to obtain an appropriate model. Then modeling each model separately using RStudio to estimate all models previously proposed. The software allows exhibiting a result an estimated values (coefficients) and standard errors of the parameters. Indeed, we can provide the approximate values of "t" or "t-values" (the coefficient is divided by the standard error). In cases where the approximate value of "t" is not logical, the model that has this illogical parameter estimation should be eliminated without completing the other steps of Box and Jenkins methodology.

Table I shows that most values of "t" are statistically significant, while the value of "t" is insignificant for the coefficient MA (1) because it is less than 2.

Table 1. STATISTICAL INDICATORS OF ARIMA (2, 1, 2) MODEL

Coefficient	Estimation	standard deviation	Value of "t"
AR(1)	0,5259	0,0955	5,30
AR(2)	0,2670	0,0960	2,18
MA(1)	0,0000	0,0423	0,00
MA(2)	1,0000	0,0423	23,98
Akaike information criteria= 1305,34			

Table II shows that values of "t" are statistically significant, and they exceed 2.

Table 2. STATISTICAL INDICATORS OF ARIMA (2, 1, 0) MODEL

Coefficient	Estimation	standard deviation	Value of "t"
AR(1)	0,3305	0,0903	5,30
AR(2)	0,3485	0,0896	2,18
Akaike information criteria= 1323,33			

Table III shows that values of "t" are statistically significant for the ARIMA (0, 1, 2) model.

Table 3. STATISTICAL INDICATORS OF ARIMA (0, 1, 2) MODEL

Coefficient	Estimation	standard deviation	Value of "t"
MA(1)	0,5558	0,1008	5,51
MA(2)	0,4442	0,0964	4,61
Akaike information criteria= 1305,02			

4.5 DIAGNOSTICS CHECKING

Selecting the most appropriate model should not only provide relatively accurate forecasts. It must contain only statistical noise without any regular pattern (residue plot must not expose auto-correlations). Therefore, this step consists mainly of plotting the residuals over time to examine the existence of systematic trends (it should be noted the absence of trends), to represent the autocorrelation function of residuals to identify any autocorrelation (it should be noted the absence of autocorrelation between residuals) and run a normality test (diagram Q - Q).

The figures above show the diagnostics of the residuals from ARIMA (0, 2, 1). First, the standardized residuals plot (Fig. 6) shows no obvious trend and pattern and looks like an independent and identical distribution. Second, in terms of autocorrelation, the ACF plot of residuals (Fig. 7) shows no evidence of significant correlation in the residuals. Finally, the normality test plot (Fig. 8) shows that most of the residuals are located on the straight line except some few outliers deviating from the normality. In conclusion, the model is adequate and fits well. Like the diagnostics checking of ARIMA (0, 1, 2), it is also observed that ARIMA (2, 1, 2) and (2, 1, 0) models exposed similar diagnostics characteristics as ARIMA (0, 1, 2) model.

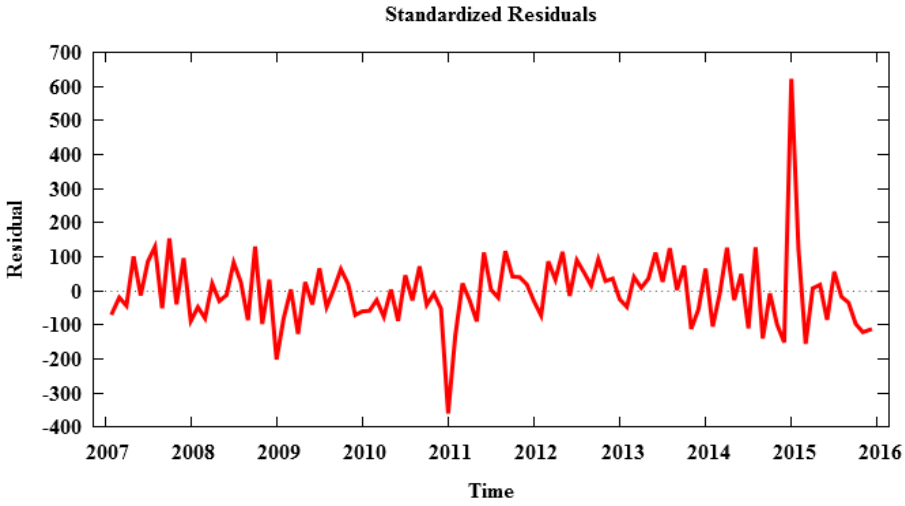

Fig. 6. The normality test plot of ARIMA (0, 1, 2)

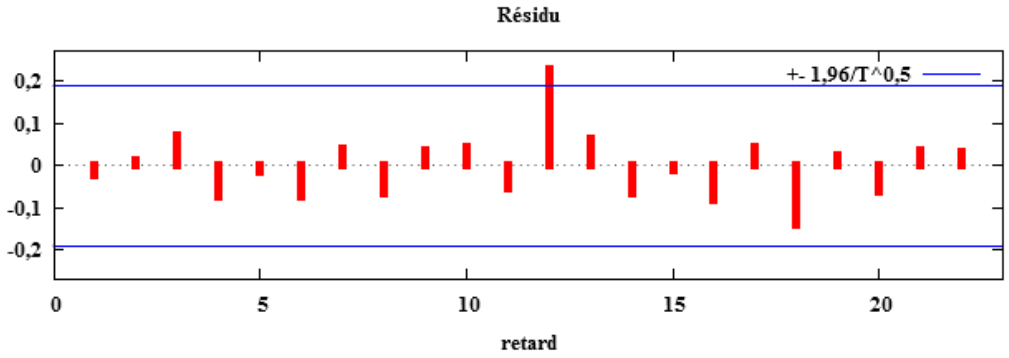

Fig. 7. The ACF plot of residuals of ARIMA (0, 1, 2)

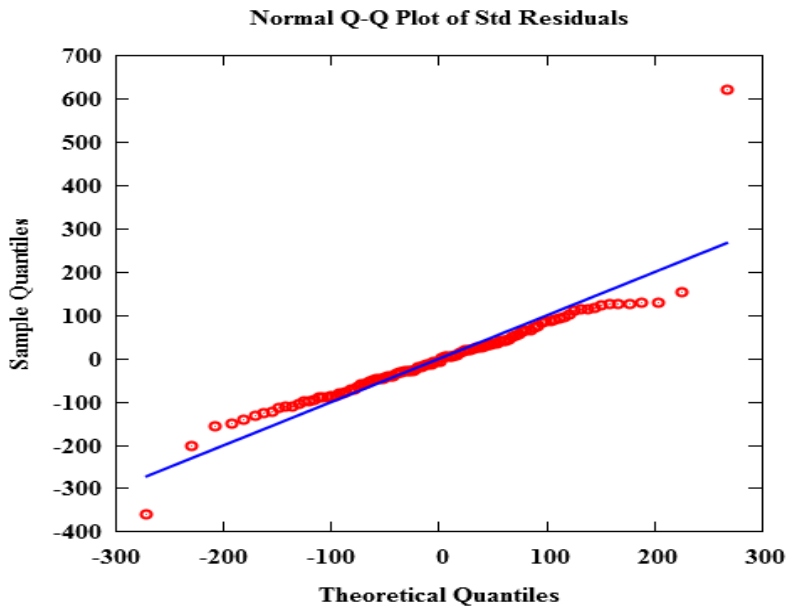

Fig. 8. The normality test plot of ARIMA (0, 1, 2)

4.6 SELECTION OF THE MOST APPROPRIATE MODEL FOR FORECASTING ACCIDENT CASES

Regarding diagnostics checking tests of the previously proposed ARIMA models, the standardized residuals plots (standard errors) of models were distributed in an independent and identical manner (irregular variations). In addition, the autocorrelation functions (ACF) plots of the residual of all models have shown no correlation. Furthermore, the normality test plots were significant for all models.

From TABLE I, II and III, the values of "t" demonstrate that coefficients of ARIMA (0, 1, 2) and ARIMA (2, 1, 0) models are significant except for ARIMA (2, 1, 2) is not significant, the parameters of ARIMA (0, 1, 2) and ARIMA (2, 1, 0) are compared. Comparing the Akaike information criteria (AIC) of the models, it is noticed that ARIMA (0, 1, 2) has the minimum AIC and residual variance. According to the above discussion, the ARIMA (0, 1, 2) model is the most favorable for forecasting accident cases in Tunisia.

4.7 FORECASTING ACCIDENT CASES IN TUNISIA

For this study, the forecast presents the final step of the Box and Jenkins methodology. Based on our time series, we could use the appropriate model ARIMA (0, 1, 2) to forecast accident cases. Consequently, we could formulate our model by the following equation:

$$y_t = y_{t-1} - (\theta_1 \mathcal{E}_{t-1} + \theta_2 \mathcal{E}_{t-2}) + \mu + \mathcal{E}_t \quad (7)$$

with, θ_1 = - 0,5558 and θ_2 = -0,4442

Based on equation (7), fig. 9 represents the visual representation of accidents cases from 2007 until the end of 2015 (red line). Moreover, based on the adjusted forecasts for the same period, the figure below shows a future road accident cases (blue line) with its confidence interval (green area). The aggressive and irresponsible behavior of Tunisian drivers constitutes the first major factor in road accidents. According to Fig.9, the road accident cases in Tunisia will represent a surprising decrease over the next three years.

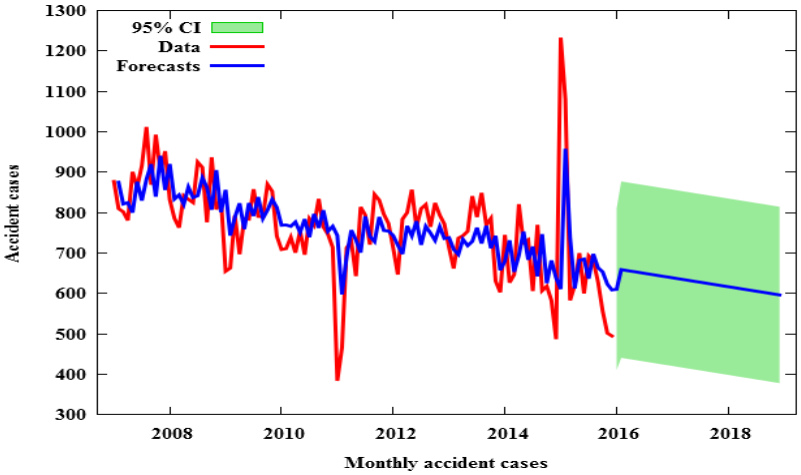

Fig. 9. Graph of the accident cases, its forecast and confidence intervals

By making the connection with the daily life of Tunisians, this decrease can be explained by several causes such as the political and security instability that marks the country nowadays and which makes it possible to slow or even stop the road traffic flows. Moreover, the degradation of climatic conditions (snow falls and floods) which affect the country every winter and which have a direct effect on the fluidity of traffic flow and the mobility of individuals; in addition to the vehicular intensification in urban areas which can slow traffic.

5 CONCLUSIONS

In this study, there are two main objectives. The first is the emphasis on the importance of temporal analysis (time series analysis) to make the best use of the data collected by the NRSO. The second is the implementation of effective road safety policy against accidents by predicting future road accidents. In scientific research, the time series of road accidents represent a very useful data in the road safety analysis. In terms of visualization, these series can be used to display and describe events related to road safety and the evolution of accidents or road victims over time. Therefore, the time series analysis related to the accidentology has a very important place in the decision-making process to improve the level of road safety. Concerning the results of our study, we can conclude from the descriptive analysis of the time series of road accident data in Tunisia over 9 years from 2007 to the end of 2015, that accidents show a small decrease before the revolution. But, in general it is characterized by an irregular evolution in the rest of the series. After applying the Box and Jenkins methodology, ARIMA (0, 1, 2) is identified as the most appropriate model to our data. Moreover, the forecast executed by this model shows a decrease in the number of accidents in Tunisia for the next three years. Based on our daily lives, the reduction in road accidents can be explained generally by an increase in number of strike actions repetitive cold waves, an increase in vehicle fleets which can slow the movement of vehicular flow in urban areas. This result can draw attention to the importance of applying key measures (time series analysis) to have a clear vision on the exact level of road safety and to improve the prevention strategy, based on scientific studies and research.

REFERENCES

[1] W. H. Organization, World report on road traffic injury prevention, Geneva: WHO, 2013.
[2] W. H. Organization, World report on road traffic injury prevention, Geneva: WHO, 2015.
[3] B. G. e. P. D.A, «Distribution of Residual Autocorrelations in Autoregressive Moving Average Time Series Models, » Journal of the American Statistical Association, vol. 65, 1970.
[4] P. F. Christens, Statistical modelling of traffic safety development, Ph. D. thesis, 2003.
[5] Y. X.-j. Z. Z.-b. T. M.-j. C. a. Y. G. Yuan, «Autoregressive Integrated Moving Average Model in predicting road traffic injury in China », Zhonghua Liu Xing Bing Xue Za Zhi l, vol. 34, n°17, pp. 736-9, 2013.
[6] S. Lassarre, «Analysis of progress in road safety in ten european countries», Accident Analysis and Prevention, pp. 33: 743-751, 2001.
[7] N. &. K. K. &. N. L. Levine, «Daily fluctuations in Honolulu motor vehicle accidents», Accident Analysis and Prevention, pp. Vol. 27. No. 6. P. 785-796., 1995.

[8] B. S. &. B. T. &. J. D., «Temporal and time series forecasting as a tool for traffic safety analysis», International Symposium "Road accidents prevention 2010", 2010.
[9] European Commission COST Action 329, «models for traffic and safety development and interventions», EUR20913, 2004.
[10] A. C. a. J. D. Harvey, «The effects of seat belt legislation on British road casualties: A case study in structural time series modelling », Journal of the Royal Statistical Society, pp. 149(3), 187–227, 1986.
[11] G. a. E. B. Ernst, «Funf Jahre danach: Wirksamkeit der 'Gurtanlegepflicht fur Pkw Insassen», Zeitschrift fur Verkehrssicherheit, pp. 36(1), 2–13, 1990.
[12] P. A. a. J. D. L. Scuffham, «A model of traffic crashes in New Zealand», Accident Analysis and Prevention, pp. 34(5), 673–687., 2002.
[13] P. G. Gould, «Econometric modelling of road crashes», Ph. D. thesis, 2005.
[14] F. A. M. Van den Bossche, «Road Safety, Risk and exposure in Belgium», PhD, 2006.
[15] G. E. a. G. C. T. Box, «Intervention analysis with applications to economic and environmental problems», Journal of the American Statistical Association, pp. 70-79, 1975.
[16] A. C. Wagenaar, «The effects of macroeconomic conditions on the incidence of motor vehicle accidents», Accident Analysis and Prevention, pp. 16 (3) 191-205, 1984.
[17] F. W. G. a. B. T. Van den Bossche, «A regression model with ARMA errors to investigate the frequency and severity of road traffic accidents», chez Proceedings of the 83rd Annual Meeting of the Transportation Research Board, Washington, 2004.
[18] S. S. A. S. J. M. a. W. S. Rohayu, «Predicting Malaysian road fatalities for year 2020», Kuala Lumpur: Malaysian Institute of Road Safety Research 2020, 2012.