J. King Saud Univ., Vol. 7, *Eng. Sci.* (2), pp. 199–218 (A.H. 1415/1995)

Time Series Forecasts for Traffic Accidents, Injuries, and Fatalities in Saudi Arabia

Ali S. Al-Ghamdi

Civil Engineering Department, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

(Received 3/3/1993; Accepted for publication 15/6/1994)

Abstract. This paper develops three forecasting models for the number of traffic accidents, injuries and fatalities in Saudi Arabia. Like other developing countries, this country suffers from traffic accident problems. For example, in 1991 there were 3,232 deaths and 25,516 injuries in 37,127 reported accidents. Although there have been some efforts in studying traffic accidents in this country, more research is needed. This paper uses a time series technique in order to predict the numbers of accidents, injuries and fatalities in Saudi Arabia. Thus, three forecasting models are developed. According to these models, the trends for the number of accidents, injuries and fatalities show no decrease in the future. Consequently, the need for improving existing safety programs is vital, and more research to investigate the causes of the increasing trends is required.

Introduction

Forecasting traffic accidents, injuries, and fatalities is an important task for traffic safety planners. These forecasts are usually beneficial in providing a better understanding of accident trends and the effectiveness of existing safety countermeasures. That is, it is of interest to safety planners to assess the current policies and safety measures by looking at future accident trends and taking corrective actions.

In this study three forecasting models for traffic accidents, injuries and fatalities in Saudi Arabia were developed based on traffic data (1980-1991) obtained from the General Traffic Department [1]. Like other developing countries, this country is suffering both human losses and economic losses in a large number of accidents [2]. Looking carefully at fatality rates (deaths per 1000 accidents), shown in Fig. 1, one may conclude that there is an obvious increase in the trend of these rates since 1989, which may question the efficiency of the existing safety plans. Statistics, in 1991,

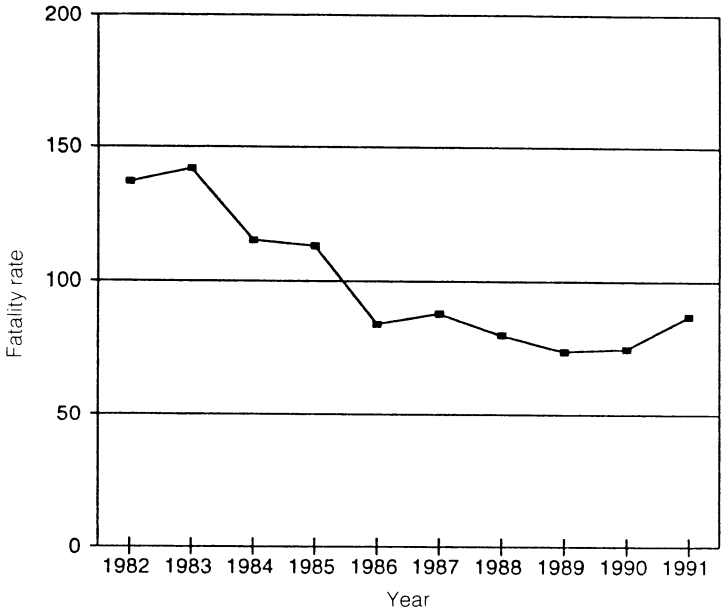

Fig. 1. Fatality rates over ten years (1982-1991)

showed among 37,127 traffic accidents there were 3,232 deaths and 25,516 injuries [1]. In a country with a small size of population (12 million), and given that 50% of this population (women) is not allowed to drive, and alcohol consumption is prohibited based on Islamic rules, these figures are relatively large, as discussed by Al-Ghamdi [2]. The need for planned methods to decrease these numbers and improve road saftety is vital. Forecasting the accident figures is one of these methods, which could help in the assessment of current safety improvement programs and where they stand with regard to the future. The time series modeling technique is employed in the present study to develop the three models.

Statistical theory concerns random samples in independent observations. However, in time-series analysis, successive observations are usually not independent. Thus, this analysis must take into account the time order of the observations. As a result of this dependence, future values may be predicted from past observations.

On the whole, a series may consist of a trend, seasonal variation, and other fluctuations. The method used in this study-Box and Jenkins-can model such variations. Before explaining the use of this method in modeling traffic accident data in Saudi Arabia, a brief description of each of these variations will be given. A trend can be

generally defined as a change in the mean over time and can be referred to the upward or downward movement that characterizes a time series over a certain period of time. Seasonal variations can be defined as periodic patterns that complete themselves within a calendar year and are then repeated on a yearly basis. Along with trend and seasonal variations, a time series could have other variations such as cyclical fluctuations.

To develop the three models in this study, Box-Jenkins methodology is used. The model-building strategy consisting of three stages, presented in Fig. 2, are discussed below. Since the author follows the same stages to develop the three models, for illustration purposes a detailed description for developing one of these models (*i.e.* fatal accident model) is given below. Then, the forms for the three models will be stated at the end of this paper.

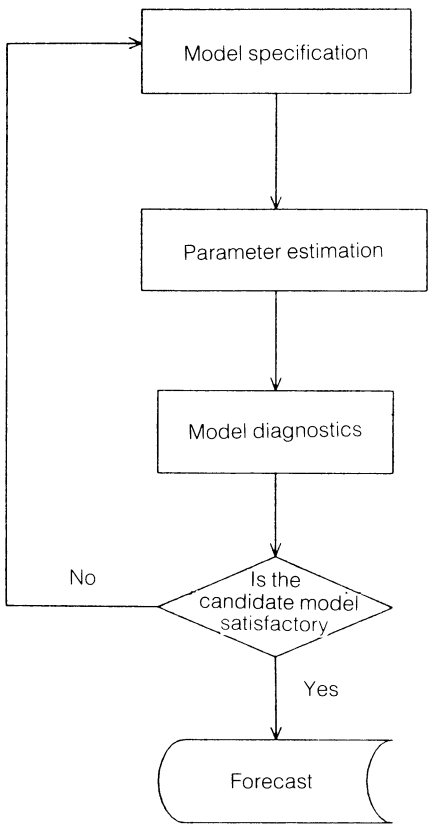

Fig. 2. The model-building strategy

Time series analysis in traffic accident studies

It is not new to apply time series modeling technique to traffic data. Traffic phenomena, as Wiorkowski and Heckard [3] indicated, show very high period to period correlations which decay gently as points on the series are more widely departed in time. Thus, this class of models discussed by Box and Jenkins can be representative for the nature of traffic data in time. Chang and Paniati [4] used Box-Jenkins modeling techniques to assess the effects of raising speed limit on rural interstate highways to 65 mph in the United States in 1987. The long-term patterns (January 1975-September 1988) of rural interstate fatalities were first examined through a systematic trend analysis using the autocorrelation function and the U-statistic. Box-Jenkins modeling techniques were then used to forecast the number of fatalities that would have occurred if the speed limit remained at 55 mph. In their study, Vaziri, Kermanshah and Hutchinson [5] focused on regression and intervention modeling of Lexinton/Fayette County specialized transportation monthly riderships. They found that intervention models from time-series analysis properly replicated these monthly riderships. The superiority of intervention modeling when compared with regression analysis was found to be in capturing ridership seasonality, proper incorporation of the time-lag structure, and the intervening events of fare and services. Khasnabis and Lyoo [6, pp.30-36] tested the feasibility of using Box-Jenkins method of time series analysis for forecasting truck accidents. They showed excellent correspondence between the observed data and the model output. However, they recommended further studies to test the feasibility of using time series analysis as an accident prediction tool.

The time series literature, as briefly discussed above, has demonstrated the applicability of time-series analysis techniques for forecasting traffic accidents. Researchers should be aware of any intervening events (*e.g.*, seat-belt legislation or speed-limit change) during the time period spanned by the underlying series. In case of observing such events, methods of intervention analysis developed by Box and Tiao [7] should be considered in order to model the effects, if any, of these intervening events. Methods of intervention analysis has been applied on traffic data, such as in [3;5;8].

Basic Concepts

In the time series modeling Box-Jenkins [9;10] methodology is commonly used. The models generated by this technique are often referred to as ARIMA models (acronym stands for Auto-Regressive Integrated Moving Average). ARIMA models are characterized by the parameters p,d,q and written as ARIMA (p,d,q). In the

case where a seasonal effect is present this notation is modified to include this effect: ARIMA (p,d,q) (P,D,Q) where P,D and Q are the parameters of seasonality. In ARIMA analysis the time-sequenced data in a series $(\ldots, Z_{t-1}, Z_t, Z_{t+1}, \ldots)$ are supposed to be statistically dependent, and this dependency is modeled.

Box-Jenkins methodology applies only to stationary data series. A time series is said to be stationary when its mean, variance, and autocorrelation are constant through time. An Auto Correlation Function (ACF) is one way of measuring how the observations within a data series are related to each other. The autocorrelation coefficient can be established by using the following formula [10]:

$$r_k = \frac{\sum_{t=1}^{n-k} (Z_t - \bar{Z})\ (Z_{t+k} - \bar{Z})}{\sum_{t=1}^{n-k} (Z_t - \bar{Z})^2}$$

If z_{t+k} and z_t are not correlated at all in the available data, the value of r_k is equal to zero. Along with the acf, the PACF (Partial Auto Correlation Function) is used as a guide in selecting one or more ARIMA models that might fit the available series.

A very important concept in ARIMA modeling is differencing, which is a simple operation that involves calculating successive changes in the values of a data series. To difference a data series, consider a new variable (w_t) which is the change in z_t, that is,

$$w_t = Z_t - Z_{t-1} \ ; \ t = 2,3, \ldots, n$$

In practice, nonstationary data are more frequent. Differencing is a simple tool to transform such data into stationary ones.

Model-building Strategy for Fatal Accidents

As just mentioned and shown in Fig. 2, this strategy consists of three stages:

1. Model specification;
2. Parameter estimation;
3. Model diagnostics.

with a few interactions of this model-building strategy, it is hoped to arrive at the best possible model for the given series. Now a discussion of each stage in the course of developing the fatal-accident model is presented.

Model specification

In the model specification stage the goal is to select a model which is tentatively believed to fit the underlying series of data. In other words, the arguments (parameters) of the selected ARIMA should be defined. This identification can be reached through several steps, including the investigation of time, ACF, and PACF plots.

By looking carefully at the time plot of fatality data, as shown in Fig. 3, one can detect the nonstationary behavior. That is, the mean is not constant, and the variability around the mean is not stable. This conforms with the ACF and PACF plots, where correlations do not come down to zero (*i.e.* some values of ACF and PACF over lags do not decrease fast to zero and they are significant at the 5% level), as shown in Figs. 4 and 5. To remove the stationary trend, the data must be differenced. After differencing the data (d = 1), the time plot and the ACF and PACF plots show the stationary trend. Consequently, the stage of identifying other parameters in the ARIMA model starts.

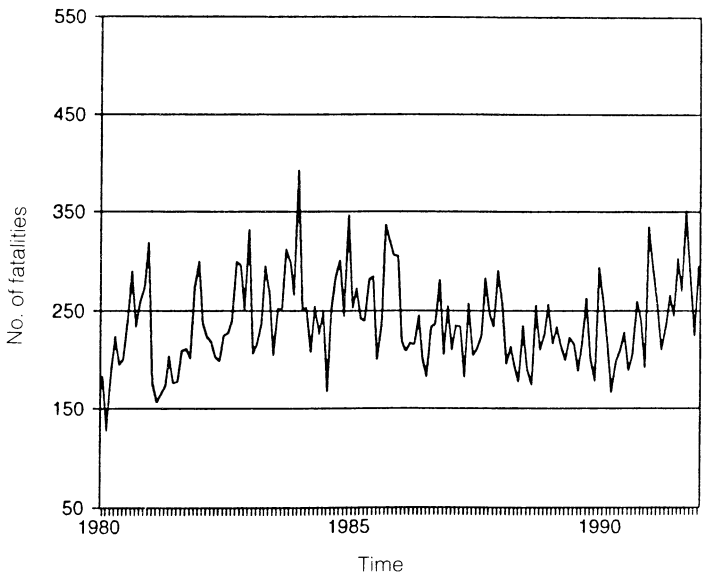

Fig. 3. Time plot of the fatality series

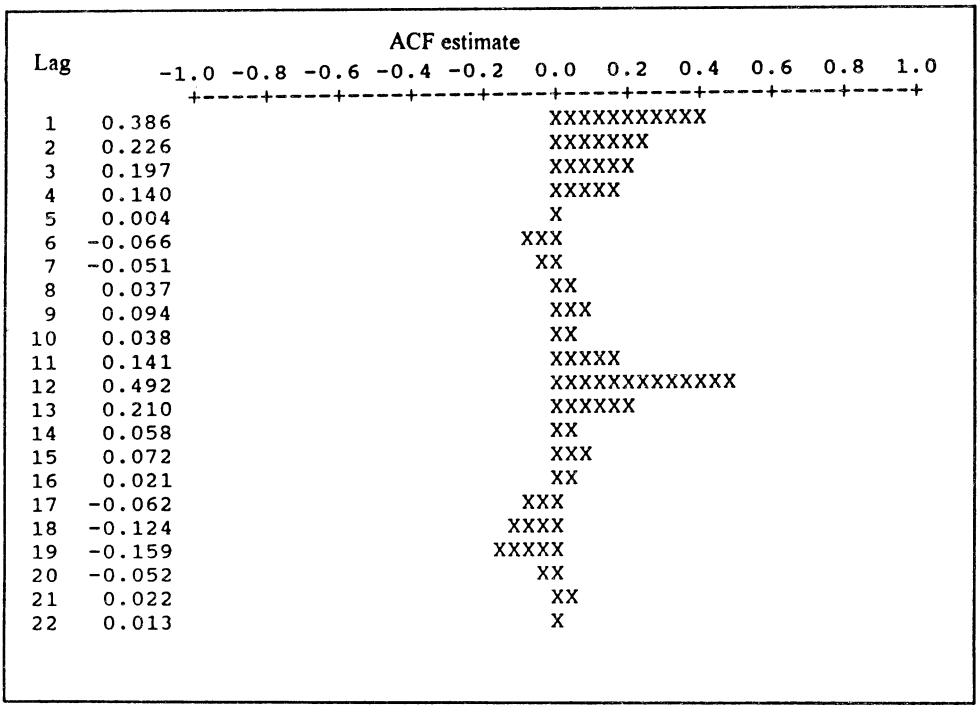

Fig. 4. The ACF plot of the original fatality data

The ACF and PACF plots for the first differenced data suggest p = 0 and q = 1 for removing the trend effect and P = 1, D = 0, and Q = 1 for removing the seasonal effect. For example, the ACF plot, in Fig. 6, shows a strong lag 1 correlation and a cut-off after that (the autocorrelation values after lag 1 are not significant at the 5% level). Notice also the strong lag 12 correlation which supports the presence of seasonality effect. The PACF plot, in Fig. 7, also shows a sort of damped exponential pattern. These findings from both plots strongly suggest the use of the ARIMA mixed model.

As a matter of fact, the process presented in Fig. 2 was repeated severl times by using the Box-Jenkins technique through the MINITAB [11] statistics package to reach the appropriate model. More than eight model specifications were tried before the selected ARIMA (0,1,1) (1,0,1) was reached as the most suitable and reasonable model to fit the accident fatality data. Although a few ARIMA models, such as

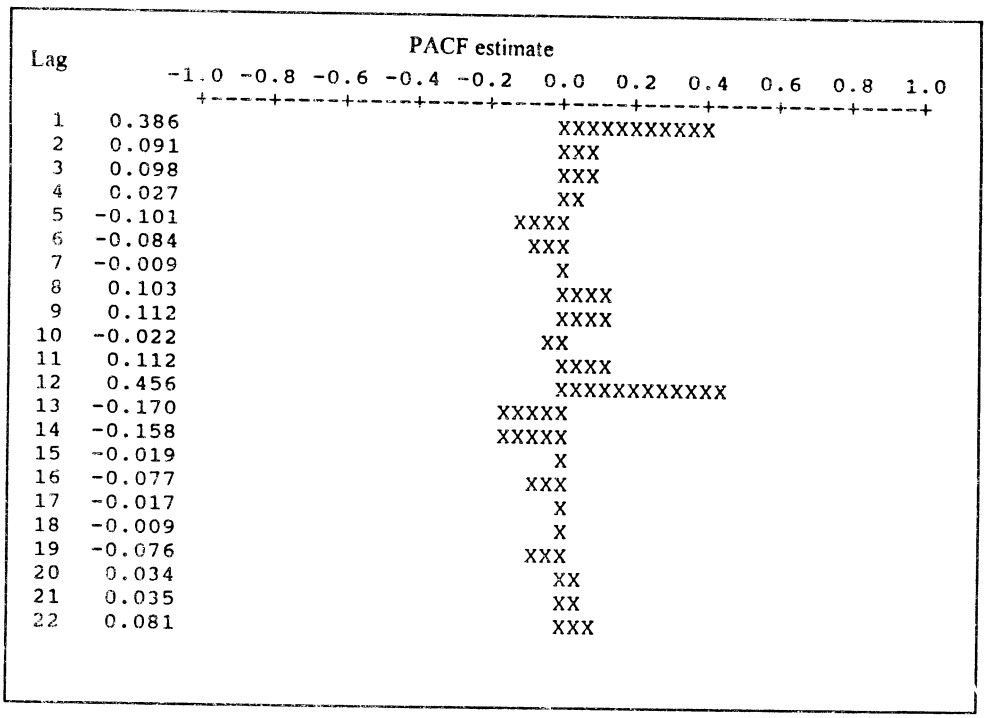

Fig. 5. The PACF plot of the original fatality data

ARIMA (0,1,2) (1,0,1), have shown a slightly better fit, the selected model was preferred for simplification purposes (*i.e.* as a few parameters as possible). The selected model passed the diagnostic tests, as will be shown shortly.

Parameter estimation

The model identified, ARIMA (0,1,1) (1,0,1), for fatalities can be written in the following backshift form:

$$(1 - \Phi_{12} B^{12})(1 - B) Z_t = (1 - \theta_1 B)(1 - \Theta_{12} B^{12}) \alpha_t \qquad (1)$$

where:

Φ_{12} = The autoregressive seasonal parameter (SAR 12),
Θ_{12} = The moving average seasonal parameter (SMA 12),

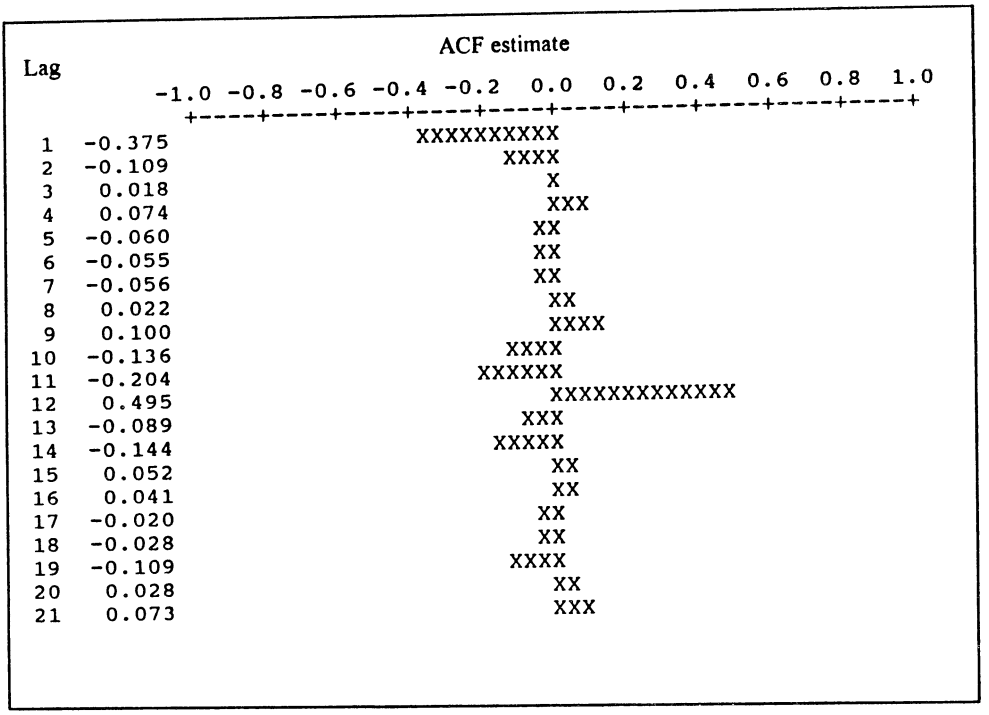

Fig. 6. The ACF of the first-differenced fatality data

θ_1 = The moving average parameter (MA),
B = The backshift operator,
α_t = Random error (in the language of time series called white noise), and the notations in the parentheses are MINITAB notations shown in Fig. 8.

Expanding (1) and solving for Z_t, the following difference formula is obtained:

$$Z_t = Z_{t-1} + \Phi_{12}(Z_{t-13} - Z_{t-12}) - \theta_1 \alpha_{t-1} + \Theta_{12}(\theta_1 \alpha_{t-13} - \alpha_{t-12}) + \alpha_t \quad (2)$$

Including the parameter estimates, listed in the MINITAB printout presented in Fig. 8, the model shows as:

$$Z_t = Z_{t-1} + 0.9874(Z_{t-13} - Z_{t-12}) - 0.7395\,\alpha_{t-1} + 0.7389 (0.7395\,\alpha_{t-13} - \alpha_{t-12}) + \alpha_t \quad (3)$$

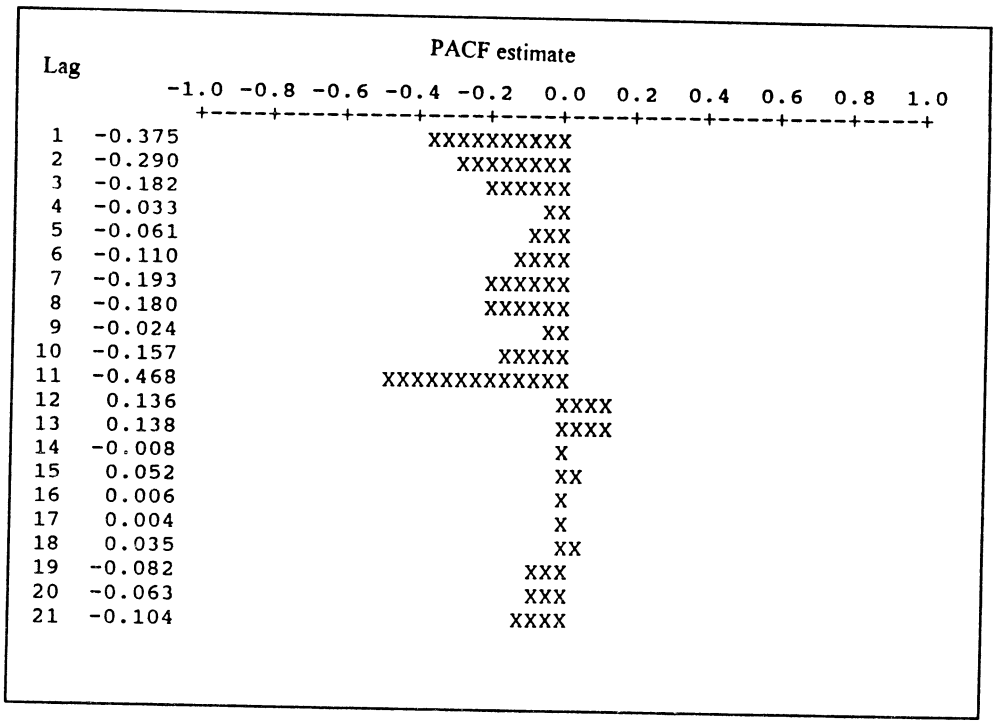

Fig. 7. The PACF of the first-differenced fatality data

This model states that the estimated fatality at time t dependes upon the number of fatalities in several previous time periods along white noise terms (error terms).

Model diagnostics

The goal of this stage is to test the goodnees-of-fit of the specified model. If the fit is poor, appropriate modifications will be conducted. Two complementary approaches were used in this diagnostic process to assess the model. First, certain characteristics of residuals, including normality and randomness, were analyzed. Second, overparameter-models (*i.e.*, models with more parameters than the specified model) were analyzed to judge the significance of adding more parameters to the selected model.

The residual randomness can be tested through the time plot of residuals, as shown in Fig. 9. In this plot no systematic trend can be detected, and the plot suggests

Final Estimates of Parameters

Type		Estimate	St. Dev.	t-ratio
SAR	12	0.9874	0.0223	44.28
MA	1	0.7395	0.0575	12.87
SMA	12	0.7389	0.0753	9.82

Differencing: 1 regular difference

No. of obs.: Original series 144, after differencing 143

Residuals: SS = 138735 (backforecasts excluded)

MS = 991 DF = 140

Modified Box-Pierce chisquare statistic

Lag	12	24	36	48
Chisquare	9.5 (DF = 9)	17.7 (DF = 21)	22.2 (DF = 33)	35.2 (DF = 45)

Fig. 8. The parameter estimates from MINITAB output

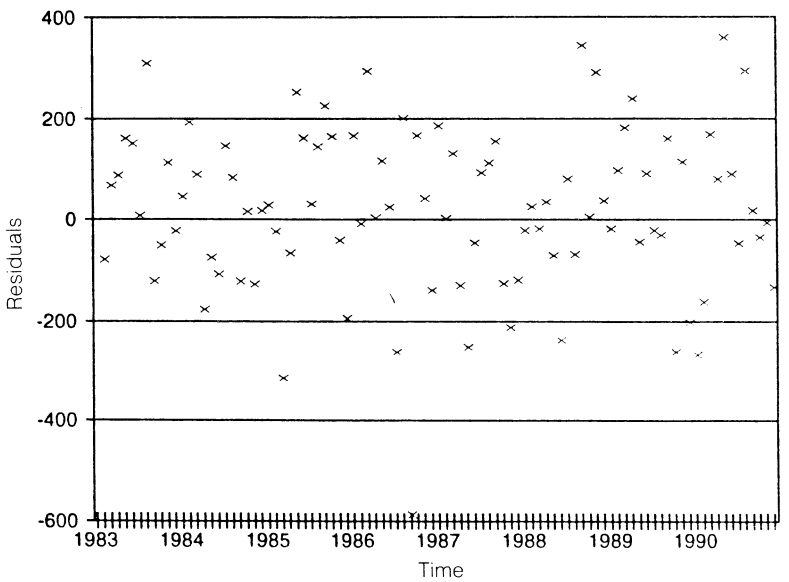

Fig. 9. Time plot of residuals

a rectangular scatter around a zero horizontal level. This finding supports the randomness of the residuals of the specified model.

The normality of the residuals can be assessed by computing a histogram of residuals, by plotting normal scores versus residuals, by conducting the normal scores correlation test, and by plotting residuals versus fitted values. The bell shape of the histogram plot in Fig. 10, the linearity of the normal-scores plot in Fig. 11, and the absence of any pattern in the residual-fitted value plot in Fig. 12 indicate that residuals tend to be normally distributed. In addition, the normal-score correlation with a value of 0.993 supports this result.

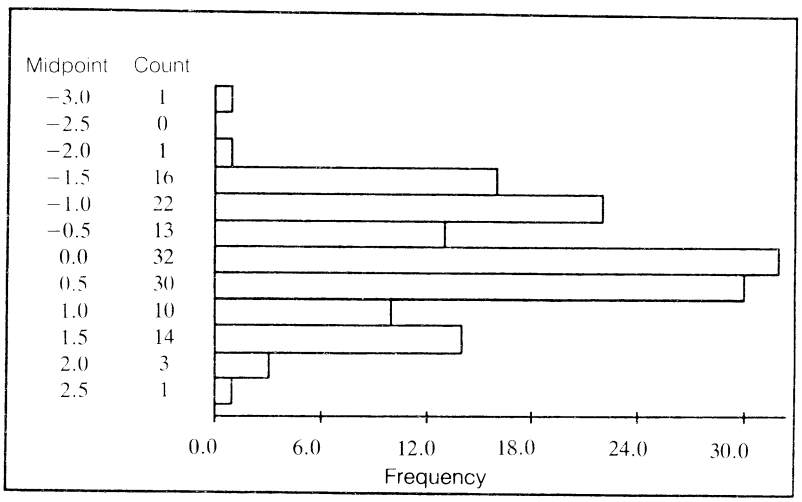

Fig. 10. Histogram of residuals

Investigating the residuals' autocorrelation plot helps in checking the independence of the noise terms in the selected model. This plot, presented in Fig. 13, shows that the residuals' autocorrelations are all within plus or minus two standard deviations (0.17), which gives no reason to question the independence of the noise terms.

Besides looking at the residuals' autocorrelations at individual lags, it is meaningful to test their magnitudes as a group. Box and Pierce [12, p. 153] proposed a statistic (Q) for this purpose. For a large sample size, if the correct model is estimated, Q has a chisquare distribution. That is, fitting an erroneous model would result in inflating Q. The model can be rejected if the observed Q exceeds an appropriate critical values in a chisquared distribution. The value of Q of the specified model, shown in the MINITAB output in Fig. 8, is not significant, indicating the appropriateness of the model.

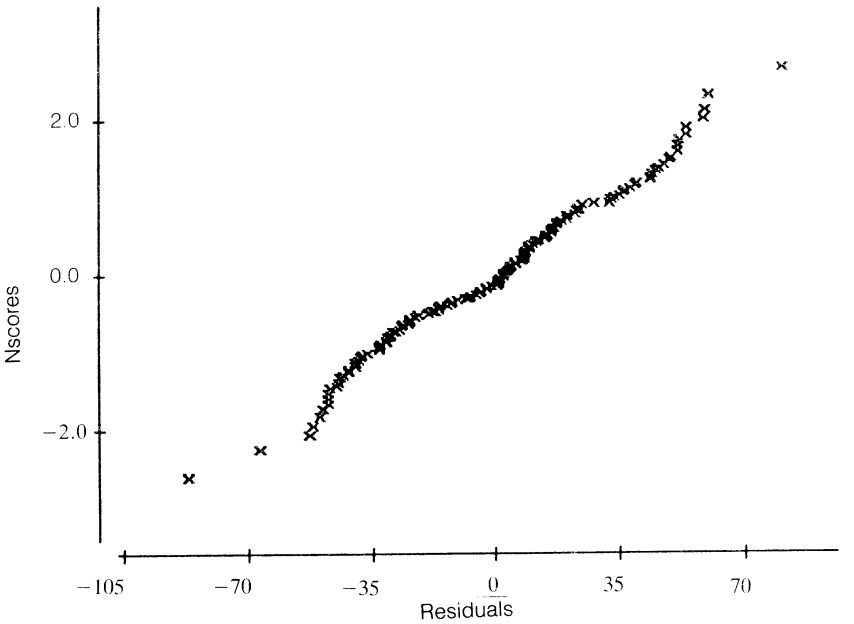

Fig. 11. Normal-score values vs. residuals

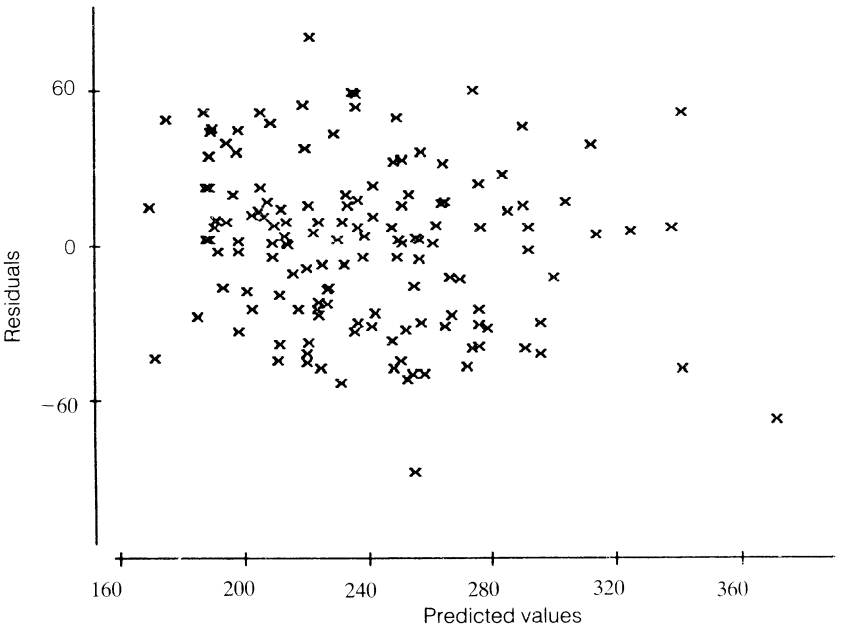

Fig. 12. Plot of residuals vs. fitted values

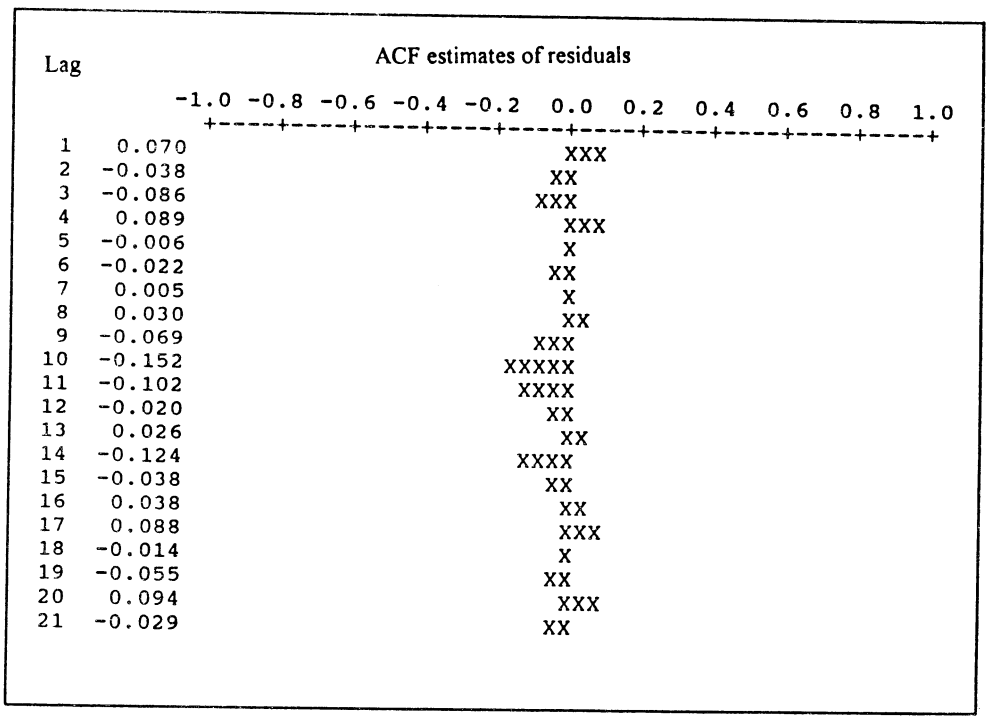

Fig. 13. The ACF plot of the residuals

In the overfitting analysis (overparametrized analysis), a more general model, namely ARIMA (0,1,1) (1,0,2), which contains the specified model as a special case, was fitted. Looking at the parameters of this general model (Table 1), one can notice two things. First, the value of the added parameter (−0.04) is close to zero and, based on its t-ratio (−0.41), insignificant. Second, the estimates for other parameters, excluding the added parameter, have not changed significantly from their original estimates in the specified model (Fig. 8). Thus, the added term is redundant and will not add more information to the specified model.

Above all, there is a very encouraging agreement between the actual numbers of fatalities and their corresponding fitted values for the 12-month period (1991), as illustrated in Fig. 14. Thus, the specified ARIMA (0,1,1) (1,0,1) for fatalities seems very realistic.

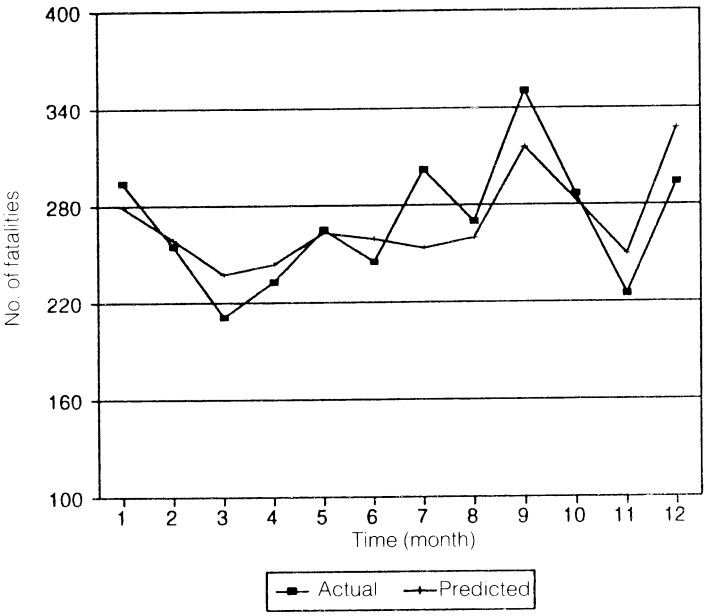

Fig. 14. Actual data and their corresponding fitted values for 12 months of 1991

Table 1. The overparameterized estimates

Parameter	Estimate	Standard deviation	t-ratio
SAR 12*	0.9831	0.0258	38.15
MA 1**	0.7425	0.0575	12.91
SMA 12***	0.7358	0.0930	7.91
SMA 24***	−0.0400	0.0979	−0.41

* seasonal autoregressive
** moving average
*** seasonal moving average

Forecasting Models

The Forecasting model for ARIMA (0,1,1) (1,0,1) can be derived through several steps, illustrated in detail in Cryer [12, p. 161]. The forecasting concept takes the past values as given data to predict future values at different points in time. The esti-

mate of future noise terms in this process is assumed to be zero, but the estimate of present and past noise terms is the residual:

$$\hat{a}_t = Z_t - \hat{Z}_t$$

After going through these steps, the three forecasting models are derived: for fatalities, for injuries, and for total accidents. These three models are illustrated below:

For fatalities: ARIMA (0,1,1) (1,0,1)

$$Z_t = Z_{t+l-1} + 0.9874\,(Z_{t+l-13} - Z_{t+l-12}) - 0.7395\,a_{t+l-1} + 0.7389$$

$$(0.7395\,a_{t+l-13} - a_{t+l-12}) + a_{t+l}$$

The noise terms $a_{t-13}, a_{t-12},\ldots,a_t$ will enter into the forecasts for $l = 1,2,\ldots13$, but for $l > 13$ the autoregressive part of the model takes over and the above model becomes:

$$\hat{Z}_t = \hat{Z}_{t+l-1} + 0.9874\,(\hat{Z}_{t+l-13} - \hat{Z}_{t+l-12}) \qquad \text{for } l > 13$$

For injuries: ARIMA (0,1,1) (1,0,0)

$$\hat{Z}_t = \hat{Z}_{t+l-1} - 0.5985\,(\hat{Z}_{t+l-13} - \hat{Z}_{t+l-12}) - 0.6245\,a_{t+l-1} - a_{t+l}$$

$$\hat{Z}_t = \hat{Z}_{t+l-1} - 0.5985\,(\hat{Z}_{t+l-13} - \hat{Z}_{t+l-12}) \qquad \text{for } l > 13$$

For total accidents: ARIMA (2,1,2) (1,1,0)

$$Z_{t+l} = Z_{t+l-1} + 0.7905\,(Z_{t+l-12} - Z_{t+l-13}) - 0.2487\,(Z_{t+l-24} - Z_{t+l-25}) -$$

$$0.6829\,a_{t+l-1} - 0.7386\,(a_{t+l-12} + 0.6829\,a_{t+l-13}) - 0.7371\,(a_{t+l-24} + 0.6829$$

$$a_{t+l-25}) + a_{t+l}$$

$$\hat{Z}_t = \hat{Z}_{t+l-1} + 0.7905\,(\hat{Z}_{t+l-12} - \hat{Z}_{t+l-13}) - 0.2487\,(\hat{Z}_{t+l-24} - \hat{Z}_{t+l-25}) \qquad \text{for } l > 13$$

The expression Z_{t+l} means that based on the available history of the series up to time t, $Z_t, Z_{t-1},\ldots, Z_1$, we would like to forecast the value Z_{t+l} l time periods in the future. The time t is called the origin of the forecast and l represents the lead time for the forecast.

Using the above forecasting models, Figs 15 and 16 show the forecasts over three-year period in the future. Although fatal forecasts show the same level, accident and injury forecasts tend to increase.

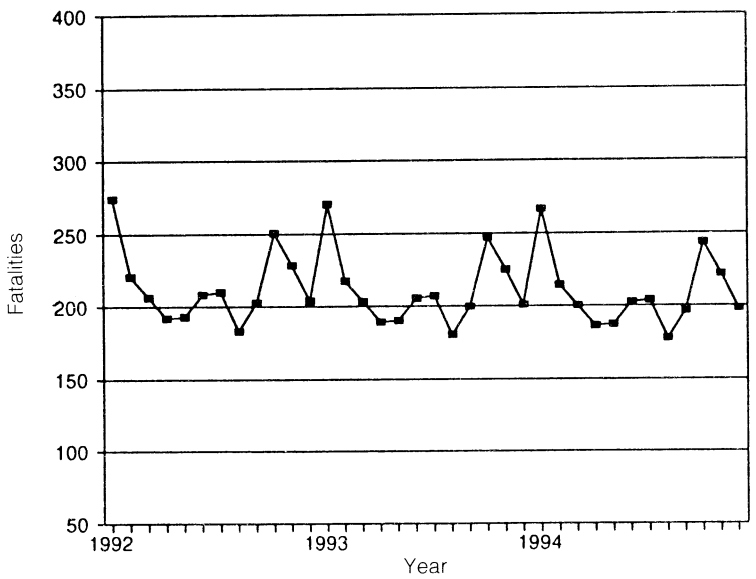

Fig. 15. Fatality forecast over three-year period

Conclusion

This study shows the applicability of time series modeling to forecast traffic accidents, injuries and fatalities in Saudi Arabia. Three models for forecasting purposes are derived based on Box-Jenkins methodology. The ARIMA (2,1,2) (1,1,0) is specified for predicting the number of traffic accidents at any point in time. For predicting injuries and fatalities, ARIMA (0,1,1,) (1,0,0) and ARIMA (0,1,1) (1,0,1) are identified, respectively. The diagnostic process for each of these models accompanied with 95% confidence intervals for their predicted values indicate the good fit of these models.

Broadly speaking, one of the purposes of time series modeling is to predict the future values of the series. This is an important task in many studies, including traffic accidents. Based on this technique, traffic analysts can predict that a traffic accident

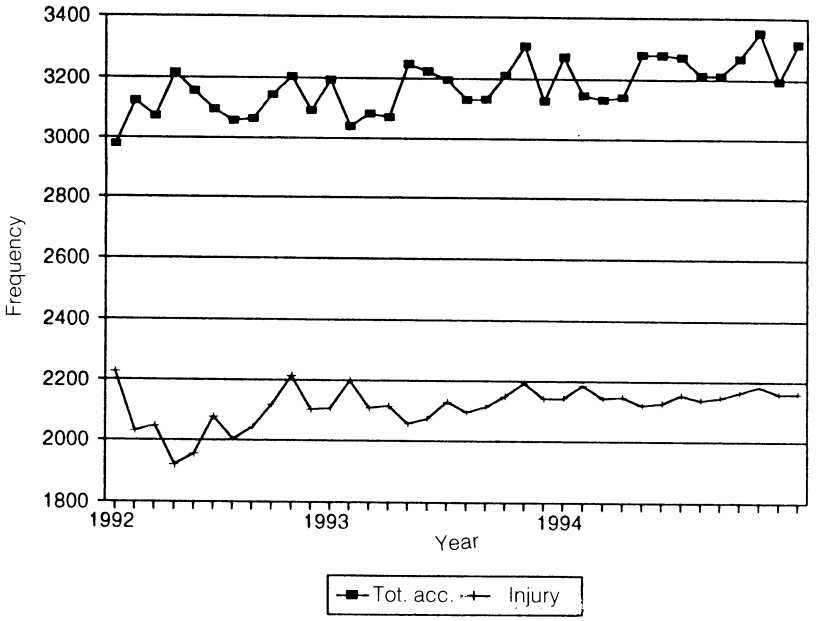

Fig. 16. Total-accident and injury forecast over three-year period

trend is raising; then appropriate corrective modification for existing road safety policy can be taken. Accordingly, the three models developed in this study could be of great help to traffic planners in Saudi Arabia to evaluate their safety programs. It should be mentioned that time-series modeling gives reasonable forecasts in short time in the future (*i.e.* 2 to 3 years ahead), therefore, when time goes by new data become available and updating the developed models can be made to improve forecasts. This updating can be carried out through a certain procedure that will be the subject of another study for the author. The three models developed in this study have shown the need for a quick revision of the current traffic safety programs in Saudi Arabia since no decrease takes place in the future predictions of the numbers of accidents, injuries, and fatalities. An improvement in traffic safety plans is urgently needed to reduce these numbers in the future.

Acknowledgement: The author would like to thank the officials in the General Traffic Department, Riyadh, Saudi Arabia, for making the data used in this study available.

References

[1] Ministry of Interior. *Traffic Statistics for the Years from 1983 to 1991.* Riyadh: Saudi Arabian Ministry of Interior, General Department of Traffic, 1991.

[2] Al-Ghamdi, A. "Forecasting Traffic Accidents in Saudi Arabia by Using a Time Series Model." Transportation Research Board (TRB). Presented in the *72nd Annual Meeting*, Washington, D.C., January (1993).

[3] Wiorkwski, J.J. and Heckard, R.F. "The Use of Time Series Analysis and Intervention Analysis to Assess the Effects of External Factors on Traffic Indices: A Case Study of the Effects of the Speed Limit Reduction and Energy Crisis in the State of Texas." *Accident Analysis and Prevention*, 9, (1977), 229-247.

[4] Chang, G. and Paniati, J.F. "Effects of 65-mph Speed Limit on Traffic Safety." *Journal of Transportation Engineering*, (ASCE), 116, No. 2 (1990), 213-226.

[5] Vaziri, M.; Kermanshah, M. and Hutchinson, J. "Short-Term Demand for Specialized Transportation: Time-Series Model." *Journal of Transportation Engineering*. (ASCE), 116, No. 1 (1990), 105-121.

[6] Khasnabis and Lyoo. "Use of Time Series Analysis to Forecast Truck Accidents." In: *Transportation Research Record 1249, TRB*. Washington, D.C.: National Research Council, 1989.

[7] Box, G.E.P. and Tiao, G.C. "Intervention Analysis with Applications to Economic and Environmental Problems." *Journal of the American Statistical Association*, 70, No. 349 (1975), 70-79.

[8] Bhattacharyya, M.N. and Layton, A.P. "Effectiveness of Seat Belt Legislation on the Queensland Road Toll. An Australian Case Study in Intervention Analysis." *Journal of the American Statistical Association*, 74, No. 367 (1979), 596-603.

[9] Box, G. and Jenkins, G. *Time Series Analysis: Forecasting and Control*. Oakland, California: Holden Day, 1976.

[10] Pankratz, A. *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. New York: John Wiley & Sons, 1983.

[11] MINITAB Inc. *MINITAB: Reference Manual, release 7*. State College, PA: Minitab Inc., 1989.

[12] Cryer, J. *Time Series Analysis*. Boston: PWS-KENT Publishing Company, 1986.

تنبؤات التسلسل الزمني لحوادث المرور ومصابينها وقتلاها في المملكة العربية السعودية

علي سـعـيد الغامـدي

قسم الهندسة المدنية، كلية الهندسة، جامعة الملك سعود، ص.ب. ٨٠٠، الرياض ١١٤٢١، المملكة العربية السعودية

(استلم في ١٩٩٣/٣/٣م ؛ قبل للنشر في ١٩٩٤/٦/١٥م)

ملخص البحث. في هذه الدراسة تّم تطوير ثلاثة نماذج تنبؤيّة لعدد الحوادث المروريّة والمصابين والوفيات الناجمة عن هذه الحوادث في المملكة العربية السعودية. مثل الدول النامية الأخرى، تعاني المملكة من مشكلات حوادث المرور. على سبيل المثال كان هناك (٣,٢٣٢) وفاة و (٢٥,٥١٦) مصاب في (٣٧,١٢٧) حادث مروري مسجل وقعت خلال عام ١٩٩١م. وبالرغم من أنّ هناك جهودًا لدراسة حوادث المرور في المملكة إلا أنّ هناك حاجة لأبحاث أكثر. استخدم في هذه الدراسة أسلوب السلاسل الزمنية للتنبؤ بأعداد الحوادث والمصابين والوفيّات في المملكة حيث تمّ بناء ثلاثة نماذج تنبؤيّة لذلك. وقد اتضح من هذه النماذج أنّ أنماط الحوادث، المصابين والوفيات لم تظهر تناقصًا في المستقبل. من ذلك تبدو الحاجة الملحّة إلى تحسين برامج السلامة المروريّة المعمول بها حاليًا وأبحاث أكثر لفحص أسباب هذه الأنماط المتزايدة.