Time series modelling strategies for road traffic accident and injury data: a case study

Ghanim Al-Hasani, Md Asaduzzaman, and Abdel-Hamid Soliman

Staffordshire University, Stoke on Trent ST4 2DE, UK,
md.asaduzzaman@staffs.ac.uk

Abstract. The paper aims to provide insights of choosing suitable time series models and analysing road traffic accidents and injuries taking road traffic accident (RTA) and injuries (RTI) data in Oman as a case study as the country faces one of the highest numbers of road accidents per year. Data from January 2000 to June 2019 from several secondary sources were gathered. Time series decomposition, stationarity and seasonality checking were performed to identify the appropriate models for RTA and RTI. SARIMA (3, 1, 1)(2, 0, 0)(12) and SARIMA (0, 1, 1)(1, 0, 2)(12) models were found to be the best for the road traffic accident and injury data, respectively, comparing many different models. AIC, BIC and other error values were used to choose the best model. Model diagnostics were also performed to confirm the statistical assumptions and two-year forecasting was performed. The analyses in this paper would help many Government Departments, academic researchers and decision-makers to generate policies to reduce accidents and injuries.

Keywords: time series modelling, model diagnostics, SARIMA, road traffic accidents (RTA), road traffic injuries (RTI)

1 Introduction

Road traffic accident (RTA) is one of the prime reasons for fatalities and disabilities globally. It has been considered as one of the significant health problems in term of death and disability [1]. Over fifty million injuries occur and more than 1.2 million people die in roadway-related accidents yearly. The RTA is going to be the fifth main cause of death in the world by 2030 [5]. Since over 50% of young adults, aged 15 to 44 years, die due to RTA a significant economic impact is discernible through the loss of earning upon their families [8]. Moreover, road crashes including deaths and injuries cost from 1-2% ($100bn) of the gross national product in low and middle-income countries in addition to the total development aid received by these countries [7].

Although time series models for continuous variables have been well-studied, autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models by Box and Jenkins (1970) have also been used to model count data recently [10]. However, finding an appropriate model for RTA and RTI time series data through suitable criteria and diagnostic checking are yet

nontrivial tasks [3]. The major essential steps involved in time series modelling process are- model specification, fitting and diagnostics of the model [2].

Due to the high concern of the government departments in major developed and mid-developed countries and the change of road safety policies, there has been a significant shift of trend of the number of accidents and injuries in recent years. Although there is a decline in the trend of the number of accidents in many Gulf countries including Oman the severity and number of injuries have been found yet to be significant. Recently a substantial decrease in the number of injuries observed for Oman. The number of accidents and injuries are high volatility, there are also shift of trends in accidents and injuries over time and they possess high seasonality, which makes the analysis, model selection and forecasting complex tasks. In this paper, we study the times series model identification from a set of models, performed suitable diagnostic checks for the case study datasets on RTA and RTI in Oman and forecasted accidents and injuries for next two years. The rest of the paper is organised as methodology in Section 2, data description in Section 3, results and discussion are given in Section 4 and some concluding remarks in Section 5.

2 Methodology

Time series is a sequence of values of a variable recorded over time, most often at a regular time interval. Time series are usually decomposed into four components: trend (T_t), cyclical pattern (C_t), seasonal variation (S_t) and random error (I_t). A time series can be expressed by an additive model defined as

$$X_t = T_t + C_t + S_t + I_t, \tag{1}$$

which can be used when the variation around the trend does not vary with the series. The multiplicative model defined as

$$X_t = T_t \cdot C_t \cdot S_t \cdot I_t, \tag{2}$$

is appropriate when the trend is proportional to the series. Often graphical approach (plot of a series) is used to identify whether a time series is additive or multiplicative.

There are three parts in the ARIMA model: AR is the autoregressive part, I is the differencing part and MA is the moving average part.

An ARIMA (p, d, q) model for a time series sequence $\{X_t,\, t = 1, 2, \ldots, n\}$ can be written as

$$\phi(B)(1 - B)^d X_t = \theta(B) A_t, \tag{3}$$

where p is the order of the AR process, d is an order of differences, q is the order of the MA process, A_t is the white noise sequence, ϕ is a polynomial of degree p, B is a backshift operator and θ is a polynomial of degree q.

However, an ARIMA model could not analyse time series with seasonal characteristics; therefore, seasonal autoregressive integrated moving average (SARIMA)

models have been developed [12]. SARIMA models perform better than the historical average, linear regression, and simple ARIMA models for data with seasonal variations. In fact, SARIMA models are capable of taking into account the trend and seasonality. A SARIMA$(p,d,q)(P,D,Q)_s$ model can be written by the following equation

$$\phi(B)\Phi(B^s)(1-B^s)^d X_t = \theta(B)\Theta(B^s)A_t, \tag{4}$$

where Φ, Θ, P, D and Q are seasonal counterparts of ϕ, θ, p, d and q, respectively, and s is the seasonality.

There are several approaches to fit time series models such as the least square method, method of moment and maximum likelihood method. However, choosing the most suitable model can be cumbersome. Several criteria such as Akaike information criterion (AIC) and Bayesian Information Criteria (BIC) have been used as an essential tool to choose the best model from a set of models [11].

Once a suitable model is fitted, diagnostic checking of the model is performed, which concerns evaluating the quality of the model. This study assesses time series models through three residuals: root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute scaled error (MASE), which are frequently used in time series analysis. RMSE is the standard deviation of the residuals, defined as

$$\text{RMSE} = \left[\frac{\sum_{i=1}^{n}(x_{f_i} - x_{o_i})^2}{n}\right]^{1/2}, \tag{5}$$

where n is a simple size, x_{f_i} are the forecasted values, x_{o_i} are the observed values. The mean absolute percent error (MAPE) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error, as defined in the equation below

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_{f_i} - x_{o_i}|}{|x_{f_i}|} \times 100. \tag{6}$$

The mean absolute scale error (MASE) is used to compare models of a time series through scale-free for assessing forecast accuracy across series [4]. MASE is defined as

$$\text{MASE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|x_{o_i} - x_{o_{i-1}}|}\right|\right), \tag{7}$$

where $e_t = x_{o_i} - x_{f_i}$ and the outcome values are independent of the data scale. However, if the outcome value less than one indicates to better forecasting. Alternatively, when the MASE value is greater than one, that means, the forecast is worse for the data.

3 Data

The data of road traffic accident and injuries (RTA and RTI) in Oman are maintained by the Royal Omani Police (ROP) and 'Statistical Summary Bulletins'

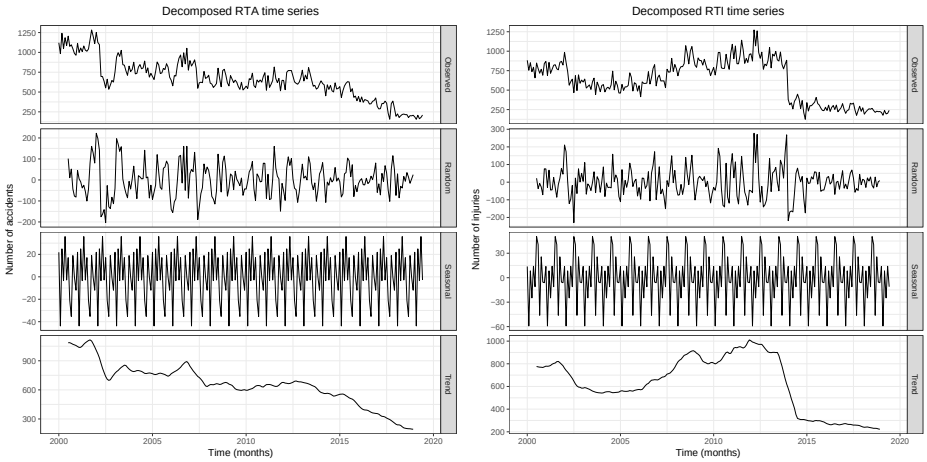
are published annually by the Directorate of Road Traffic as part of the ROP. Summary data are published by the National Centre for Statistics & Information (NCSI) in monthly reports, called, 'Monthly Statistical Bulletin' in Oman.

The RTA and RTI data for this study have been collected from two sources: 'Statistical Summary Bulletins' from the Directorate of Road Traffic and 'Monthly Statistical Bulletin' from the National Centre for Statistics & Information (NCSI). The data cover all road accidents and injuries for the period of 2000 to 2016 published in [9]. Additional data from 2016-19 have been collected from the 'Monthly Statistical Bulletin' of the National Centre for Statistics and Information (NCSI) [6]. As a result, we consider monthly time series data of road traffic accidents and injuries from January 2000 to June 2019.

4 Results and discussion

The time series data in this study represent the number of monthly road traffic accidents (RTA) and injuries (RTI) in Oman from January 2000 to June 2019. The resulting data consist of a total of 234 observations for both RTAs and RTIs.

Over the past two decades, the incidence of RTAs in Oman has fallen from a high of $1,283$ RTAs in October 2001 to a low 156 in February 2019. The mean is 660 for RTAs with a standard deviation of 247.5. Similarly, RTIs varied from a high of $1,273$ in March 2012 to a low of 125 in December 2014 with a mean of 620 RTI and standard deviation 265. The decomposition of trend, seasonality and random error components are shown in Figure 2.



Time series decomposition of RTA data Time series decomposition of RTI data

Fig. 2: Time series decomposition of RTA and RTI data

Table 1: Assessment of different models for RTA data in Oman

Model	AIC	BIC	RMSE	MAPE	MASE
$(4,1,1)(2,0,0)_{12}$	2746.56	2774.17	84.47	10.93	0.89
$(4,1,1)(1,0,0)_{12}$	2748.10	2772.25	85.19	11.12	0.91
$(3,1,1)(2,0,0)_{12}$	2744.69	2768.84	84.49	10.95	0.89
$(5,1,1)(2,0,0)_{12}$	2747.42	2778.21	84.17	10.91	0.89
$(4,1,0)(2,0,0)_{12}$	2754.42	2778.58	86.35	10.85	0.91

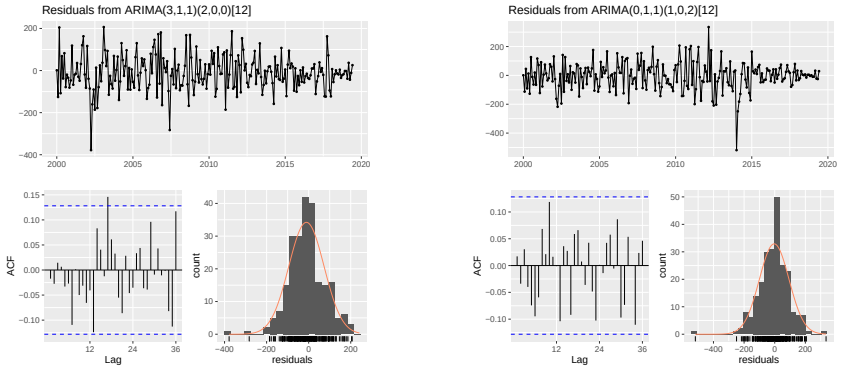
Table 2: Assessment of different models for RTI data in Oman

Model	AIC	BIC	RMSE	MAPE	MASE
$(1,1,2)(0,0,2)_{12}$	2810.48	2831.19	97.75	13.66	0.85
$(0,1,1)(0,0,2)_{12}$	2807.22	2821.03	97.92	13.65	0.85
$(0,1,1)(1,0,2)_{12}$	2804.33	2821.58	96.73	13.16	0.83
$(0,1,1)(0,0,1)_{12}$	2808.89	2819.24	98.76	13.53	0.85
$(1,1,1)(0,0,2)_{12}$	2808.72	2825.97	97.80	13.68	0.85
$(0,1,2)(0,0,2)_{12}$	2808.74	2825.99	97.80	13.68	0.85

Different SARIMA models have been fitted in R and compared using the values of AIC, BIC, RMSE, MAPE and MASE. The values of AIC, BIC, RMSE, MAPE and MASE for different SARIMA models for RTA given in Table 1. While developing different models for RTA data, models with first-order difference ($d = 1$) only considered as the data found to be stationary at lag 1. The analyses indicate that the best model for the RTA data in Oman is SARIMA $(3,1,1)(2,0,0)_{12}$ as the model has the lowest AIC (2744.69) and BIC (2768.84) values. For the RTI data, a number of models have been compared (Table 2) and the model SARIMA$(0,1,1)(1,0,2)_{12}$ is found to be the best. Although BIC value (2821.58) is not the lowest for the SARIMA$(0,1,1)(1,0,2)_{12}$ model due to more parameters than other models but the AIC value (2804.33) is the lowest. Additionally, this model has the lowest value in RMSE (96.73), MAPE (13.16) and MASE (0.83), which suggest that the model is better than the other models for the RTI time series in Oman.

Although the model $(3,1,1)(2,0,0)_{12}$ for RTAs have slightly higher RMSE and MAPE errors is yet adequate and better than the other models considering AIC and BIC values. Moreover, the residual diagnostic, more specifically, the autocorrelation of the residuals were checked by the Ljung-Box test ($Q = 22.5$, p-value$= 0.21$), which shows that the test is insignificant. Figure 4 shows that residuals for both models are white noise and ACF residuals fall near to the zero. It can be deduced from further goodness of fit analysis that the SARIMA$(3,1,1)(2,0,0)_{12}$ model fitted the data reasonably well.

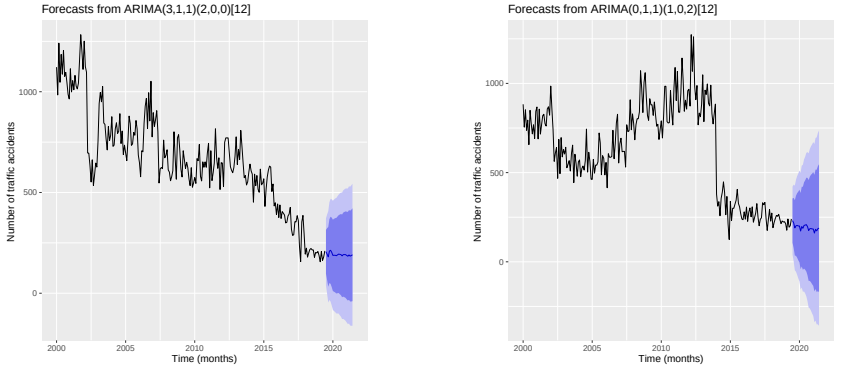
Diagnostic checking and model validation were also performed as the procedures as mentioned earlier for the RTI data. The model SARIMA $(0,1,1)(1,0,2)_{12}$ has shown the highest adequacy than other models considering RMSE, MAPE and MASE as shown in Table 2. Results of the Ljung-Box test ($Q = 21.6$,



Residuals of the fitted model in RTA Residuals of the fitted model in RTI

Fig. 4: Residuals of the fitted models

p-value= 0.4136) suggest that autocorrelation coefficients are not significantly different from zero. The ACF residuals indicate that autocorrelation is near to zero and do not deviate from a zero mean, i.e. the residuals follow a white noise process. These suggest that the SARIMA$(0,1,1)(1,0,2)_{12}$ model fitted injuries data in Oman well. Based on the final models for RTA and RTI, we have forecasted the number of accidents and injuries for the next 24 months shown in Figure 6.



Observed (black) and forecasted values (blue) of traffic accidents in Oman Observed (black) and forecasted values (blue) of traffic injuries in Oman

Fig. 6: Observed (black) and forecasted values (blue) of RTA and RTI in Oman

5 Conclusion

This study has aimed to develop a time series model to forecast road traffic accidents and road traffic injuries in the Sultanate of Oman. A peak occurred with 1,283 RTAs which then declined to the lowest point of 156 RTAs in February 2019. Based on the Box and Jenkins approach, SARIMA$(3,1,1)(2,0,0)_{12}$ model was selected to forecast RTAs for the next 24 months in Oman. This model forecasted the high occurrence of RTA in June, July and August in the following years. On the other hand, SARIMA$(0,1,1)(1,0,2)_{12}$ model was selected for predicting the number of traffic injuries, which shows a downward trend. The policymakers in Oman should keep under their consideration the results of this study. Both models in RTA and RTI shows that there is a higher chance of accidents and injuries during the summer season. For future work, we would like to study the fatality-related crashes in Oman. Furthermore, investigation of the spatial factors and the socio-economic impact of the RTAs in Oman would be considered.

References

1. Boulieri, A., Liverani, S., de Hoogh, K., Blangiardo, M.: A space–time multivariate Bayesian model to analyse road traffic accidents by severity. Journal of the Royal Statistical Society: Series A (Statistics in Society) **180**(1) (2017) 119–139
2. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
3. Cryer, J.D., Chan, K.S.: Time series analysis with application in R. Springer (2008)
4. Hyndman, R.J., et al.: Another look at forecast-accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting **4**(4) (2006) 43–46
5. Mannering, F., Bhat, C.: Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research **1** (2014) 1–22
6. NCSI: Monthly Statistical Bulletin. National Centre for Statistics & Information, Sultanate of Oman (2000-19)
7. Peden, M., Hyder, A.: Road traffic injuries are a global public health problem. BMJ **324**(7346) (2002) 1153
8. Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A.A., Jarawan, E., Mathers, C.D., et al.: World report on road traffic injury prevention (2004)
9. Police, R.O.: Traffic Statistic. Director General of Traffic (2013-19)
10. Quddus, M.A.: Time series count data models: an empirical application to traffic accidents. Accident Analysis & Prevention **40**(5) (2008) 1732–1741
11. Raeside, R., White, D.: Predicting casualty numbers in Great Britain. Transportation Research Record: Journal of the Transportation Research Board (1897) (2004) 142–147
12. Zhang, X., Pang, Y., Cui, M., Stallones, L., Xiang, H.: Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model. Annals of Epidemiology **25**(2) (2015) 101–106