Diagnostic Time Series Models for Road Traffic Accidents Data

Ghanim Al-Hasani[1], Aamir M. Khan[2], Hamed Al-Reesi[3], Abdullah Al-Maniri[4]
[1] Staffordshire University, Stoke on Trent, UK, *ghanem528@hotmail.com*
[2] University of Buraimi, Al Buraimi, Oman, *jadoon.engr@gmail.com*
[3] Ministry of Health, Sohar, Oman, *abuessa06@gmail.com*
[4] Oman Medical Specialty Board, Muscat, Oman, *abdullah.a@omsb.org*

Abstract

Statistical modelling of road traffic accidents forms the original insight and implementation for road safety policies. Recently, there is an emerging trend to more progressively analysing road traffic accidents data using time series techniques. However, with sophisticated tools utilized for identification and optimally fitting time series models, it is important to bear in mind the possible bias in the resulting responses. The purpose of this study is to evaluate optimal time series models for determination orders of parameters in the road traffic accidents data compared to the manual process. Time series traffic accidents data were gathered for eighteen years from secondary sources, and statistical time-series analyses were performed. Time series decomposition, stationarity and seasonality were checked to identify the appropriate models for road traffic accidents. Meanwhile, optimally analysis of data was conducted with a comparison of both the results. AIC, BIC and other error values were used to choose the best models and model diagnostics tools were applied to confirm the statistical assumptions. SARIMA(0,1,2)(1,0,2)12 and SARIMA(0,1,2)(0,0,2)12 models resulted the best models manually and automatically, respectively. The diagnostic process showed that SARIMA (0,1,2)(1,0,2)12 performed better than the optimal model. Therefore, the modellers who prefer to use the optimal function as a tool for time series model selection should consider the model's accuracy. It would be better for assessments to compare a variety of models and select the one having the best goodness of fit.

Keywords: *Road Traffic Accidents, Time Series, ARIMA, SARIMA, Diagnostic Time Series*
JEL classification: *Q17, Q18, F14*

Introduction

Road traffic accidents (RTA) cause severe problems for the societies in developed as well as developing countries and result in loss of lives and high cost. RTAs are one of the prime reasons for fatalities and disabilities globally that resulted in high economic burden. Indeed, by 2030, road traffic accidents are expected to be the fifth main cause of death globally (Mannering & Bhat, 2014).

For decades, statistical analysis and modelling have been playing a significant role in getting insights from road traffic accidents data. Practitioners in this field encourage to use statistical modelling, analysis and forecasting to identify the root causes of problems and establish a foundation for evolving policies and economics based interventions (Zhang, Pang, Cui, Stallones, & Xiang, 2015)

Therefore, time series analysis has encountered immense applicability on a massive scale. Prior to that, statisticians used descriptive methods and informal arguments for analysis. Time-series models for continuous variables like ARMA (autoregressive moving average) and ARIMA (autoregressive integrated moving average) models, popularized in the landmark work by George Box (Box & Jenkins, 1970), have been explored well over the years. However, still, these time series models find novel applications like the one used to model count data (Quddus, 2008). In addition, a growing number of studies in road safety employed time series models for forecasting the number of road traffic accidents, injuries or deaths.

There are several tools and software that utilize state-of-the-art research in analysing RTA.

Statisticians and practitioners in a road traffic accident (RTA) research have majorly been using R based tools for their models like time series analysis. However, with such sophisticated tools utilizing identification and optimally fitting time series models, it is still important to bear in mind the possible bias in their responses. The purpose of this study is to evaluate optimal (automatic) time series models, obtained using auto.arima ( ) function in open source R tools. The order values of parameters determined in this manner for the road traffic death data are compared to the manual process, and the outcome diagnostics are performed.

In the current age where detailed real-time data collected over time for RTAs can be gained at frequently low cost, time-series models can be well-suited to get novel perspective on RTA research and classic safety issues (Lavrenz, Vlahogianni, Gkritza, & Ke, 2018). However, Cryer et al. (Cryer & Chan, 2008) pointed out the potential problems when practitioners are looking to find an appropriate model for time series data through diagnostics of suitable criteria. Along with the development of ARIMA in their milestone work (Box & Jenkins, 1970), the authors Box and Jenkins also suggested a process for identifying, estimating, and checking models for a specific time-series dataset. Referred to as the Box-Jenkins Method in the updated edition of the book (Box, Reinsel, & Ljung, 2015), this is the process of stochastic model building with an iterative approach that consists of the following three essential steps:

- **Model Specification.** Using the data and all related information to help select a sub-class of the model that may best summarize the data.
- **Model Fitting/Estimation.** Use the data to train and estimate the parameters of the model (i.e. the coefficients).
- **Model Diagnostics/Checking.** Evaluate the selected fitted model in the context of the available data and check for areas where the model may be improved.

It is an iterative process so that we continually loop through this cycle as new information is gained during diagnostics and incorporate that information into new model classes. The approach starts with the assumption that the process that generated the time series can be approximated using an ARMA model if it is stationary or an ARIMA model if it is non-stationary. Once a suitable model is fitted, diagnostic checking of the model is performed, which concerns evaluating the quality of the model. This ensures that the fitted model has reasonably well-satisfied the underlying assumptions. However, if there are no inadequacies found, the model fitting is assumed to be complete, and the model can then possibly be used to forecast future values. Otherwise, in the case of inadequacies, another model is searched, and thus, we return to the model specification step again (Cryer & Chan, 2008).

Diagnostic time series model is crucial to examine the goodness of fit for the tentative model. Evaluation criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC) have been used as a vital tool for selecting the best time series model from the group of models (Ham, et al., 2017; Raeside & White, 2004; Ozaki, 1977). To ensure that the selected model reasonably satisfies the underlying assumptions, it is necessary to apply diagnostics checking for evaluating the quality performance of the model. Consequently, the model can be used for forecasting if it scores fairly accurate, or otherwise, we return to the identification step again (Cryer & Chan, 2008). For checking the adequacy of the fitted time series model, the residual diagnostics information is also quite useful. When the residuals behave like white noise, the model is adequate (Lavrenz, Vlahogianni, Gkritza, & Ke, 2018; Cryer & Chan, 2008). There are several types of forecast-residuals to assess the accuracy of the time series (Hyndman & others, 2006). Therefore, our study evaluates time series models with the following very commonly used data analysis error metrics: root means square error (RMSE), mean absolute percentage error (MAPE) and mean absolute scaled error (MASE). At a more detailed level, using ACF and PCAF plots of the time series residuals may suggest whether the residuals are uncorrelated.

Furthermore, there are several tests that are extremely useful as diagnostic to the residuals' correlation such as; the Ljung-Box test (Zhang, Pang, Cui, Stallones, & Xiang, 2015), Box-Pierce test (Szeto, Ghosh, Basu, & O'Mahony, 2009), Portmanteau lack of fit test (Manikandan, Prasad, Mishra, Konduru, & Newtonraj, 2018; Parvareh, et al., 2018). In this era, the sophisticated tools and developed software encourage modellers to apply complicated evaluation and tests of statistical models

Rest of the paper is structured as follows. The second section describes the literature review, narrating the state-of-the-art of the field by describing the notable work done by the researchers. Section three provides a brief introduction to the theory of ARIMA, followed by the description of evaluation metrics. Section four provides the crux of our research contribution with a discussion about the data and the results achieved. Finally, the last section concludes the presented work summarising the key takeaway points.

Methodology

Time series model

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

$$\phi(B)(1 - B)^d X_t = \theta(B)A_t$$

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 (Zhang, Pang, Cui, Stallones, & Xiang, 2015). SARIMA models perform better than the traditional average, linear regression, and simple ARIMA models for data with seasonal variations. In fact, SARIMA models are capable of considering the trend and seasonality. A SARIMA (p, d, q)(P, D, Q)s the following equation can write model

$$\phi(B)\Phi(B^s)(1 - B^s)^d X_t = \theta(B)A_t$$

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

Time series diagnostic tools

Statisticians established some tools to employ for diagnosing time series models. One of those, Akaike Information Criterion (AIC) is defined as

$$AIC = -2\log(L) + 2K \tag{1}$$

where L is the maximized likelihood, and K is the number of parameters. AIC is used to obtain the order of the times series models (p, d, q, P, Q, D) which are the coefficients for ARIMA (p, d, q) (P, D, Q)s model. Another diagnostic metric, Bayesian Information Criteria (BIC) is defined as

$$BIC = -2\log(L) + K\log(n) \tag{2}$$

where L is the maximized likelihood, K is the number of parameters, and n is the number of data points in the time series. The root means square error (RMSE) metric is the standard deviation of the residuals, represented by the equation,

$$RMSE = \left[\frac{\sum_{i=1}^{n}\left(x_{f_i} - x_{O_i}\right)^2}{n}\right]^{1/2} \tag{3}$$

where n is the sample size, x_{f_i} are the forecast values, and 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 errors, as identified by the equation,

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|x_{f_i} - x_{O_i}\right|}{\left|x_{f_i}\right|} \times 100. \tag{4}$$

The mean absolute scale error (MASE) is used to compare models of a time series through scale-free for assessing forecast accuracy across series (Hyndman & others, 2006). MASE is identified as the equation,

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}\left|x_{O_i} - x_{O_{i-1}}\right|}\right|\right), \tag{5}$$

where $e_t = x_{O_i} - x_{f_i}$ and the output values are independent of the scale of the data. Here if the output MASE value is less than one, it points to better forecasting. Alternatively, when the MASE value is greater than one, that indicates worse forecast for the time series data.

This study is employing the Ljung-Box test [Eq(6)] for testing and diagnosing the residuals of time series, which is defined as,

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\rho_k^2}{n-k} \tag{6}$$

where n is the sample size, ρ is the autocorrelation, K is the lags, and h is the lags to be tested.

Data

The fatal accidents data for this study has been collected from secondary sources in the Sultanate of Oman. Monthly fatal accidents were used from January 2000 to December 2018. R version 3.5.2 is used for time series data analysis.

Results and discussion

A total of 228 observations of time series data were analysed in this study. This time series represents the number of road traffic deaths every month in Oman from January 2000 to December 2018. Code implementation details and further analysis of presented research work are available online in the GitHub repository[1].

Over the past 18 years, the number of people killed on the road has risen from the lowest of 23 in September 2000 to a peak of 141 deaths in August 2012, as shown in Figure 1. From the figure (Figure 1), we conclude that this time series is non-stationary. Moreover, Augmented Dickey-Fuller test results of Dickey-Fuller = -1.7893, Lag order = 6 and p-value = 0.6644 indicate that the time series model needs to be stationrised, which in turn suggests using difference. Therefore, the primary focus is to develop a suitable ARIMA model for this time series. Ham et al. (Ham, et al., 2017) suggested to use *auto.arima( )* functions to conduct a search for all possible models and to determine the order of parameters for the ARIMA model.

[1] Code Repository: https://github.com/jadoonengr/Oman-RTA-Time-Series-Analysis

Applying this function to RTD data in Oman, SARIMA (0,1,2)(0,0,2)$_{12}$ resulted as the best model. Hence, the optimal time series model was SARIMA (0,1,2)(0,0,2)$_{12}$, which is seasonal for the current time series data.

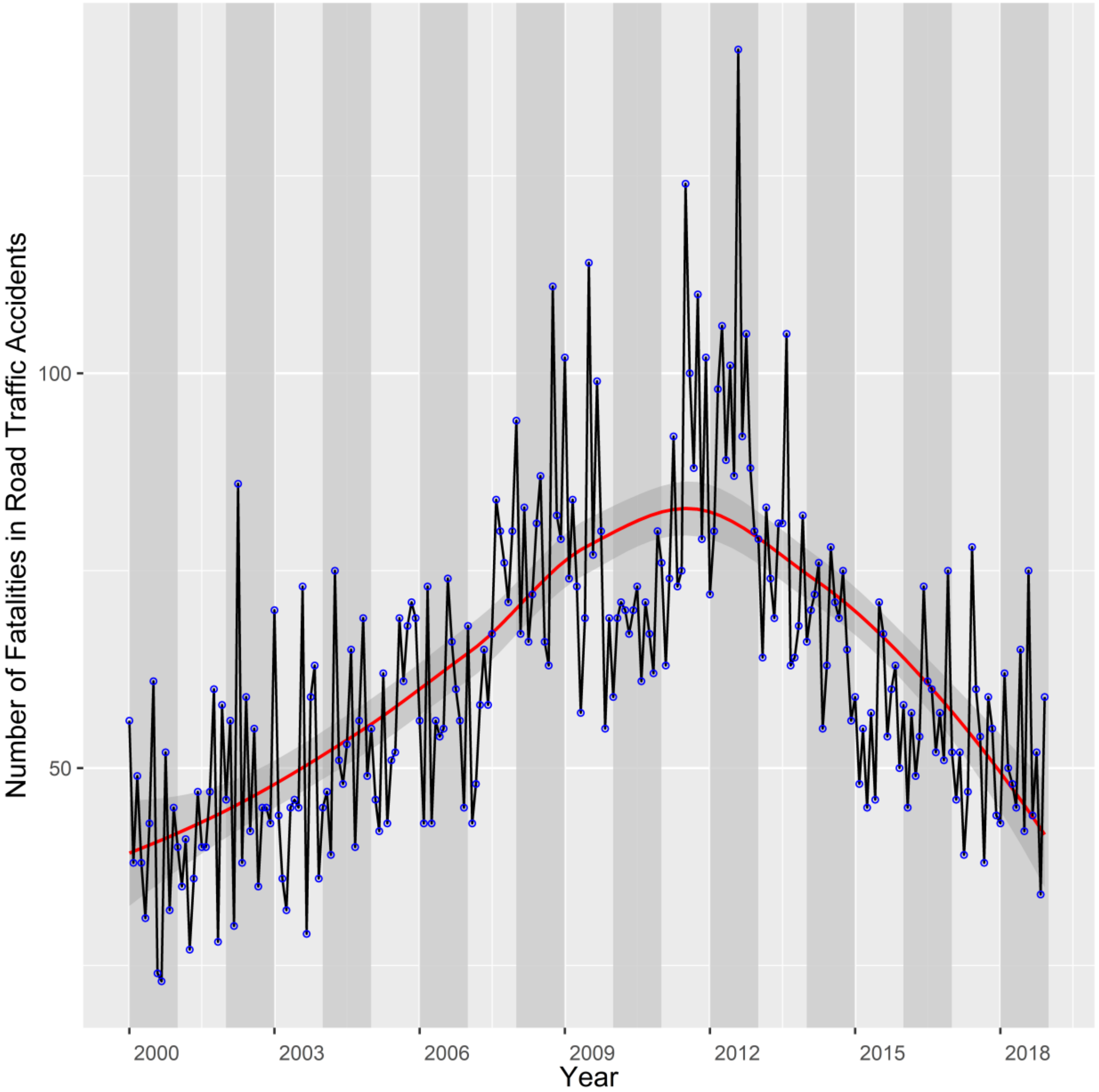

Figure 1: Road Traffic Deaths in Oman from 2000 to 2018

A different perspective on the *auto.arima( )* R function is provided by following the essential process to analyse time-series data as mentioned earlier in the literature review (Ham, et al., 2017). Also, this study used the same software R version 3.5.2 to conduct RTDs time series analysis by following the three main steps. Because of non-stationary time series curve, significant seasonality was found (as shown in figure Figure 1), it is necessary to take one difference (d=1) to be the stationary time series at lag 1. Therefore, the primary focus in the next step was to fit suitable seasonal ARIMA models in this time series. Supported by AIC and BIC values, several SARIMA models were fitted, and orders were estimated as shown in Table 1. As a result, the best time series model found is (0; 1; 2)(1; 0; 1)$_{12}$ due to lowest AIC and BIC values, 1812 and 1829 respectively. As shown in Table 1, even though the (0; 1; 2)(1; 0; 2)$_{12}$ model was the same model generated optimally by *auto.arima( )* function in the *forecast* R package; it has not got the lowest AIC (1815) and BIC (1834) values.

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

Model	AIC	BIC	RMSE	MAPE	MASE
(0,1,2) (1,0,1)$_{12}$	1812	1829	12.69	16.57	0.69
(0,1,2) (1,1,2)$_{12}$	1814	1834	12.66	16.6	0.69
(0,1,1) (1,0,2)$_{12}$	1816	1833	12.73	116.82	0.7
(0,1,3) (1,0,2)$_{12}$	1815	1839	12.66	16.6	0.69
(0,1,2) (0,0,2)$_{12}$	1815	1833	12.83	16.84	0.7
(0,1,2) (1,0,2)$_{12}$	1815	1834	12.66	16.6	0.69

Table 2: Comparison optimal and manual models for RTD data in Oman

Type	Model	AIC	BIC	RMSE	MAPE	MASE
Manual	(0,1,2) (1,0,1)$_{12}$	1812	1829	12.69	16.57	0.69
Optimal	(0,1,2) (0,0,2)$_{12}$	1815	1834	12.83	16.84	0.7

In the final step, it is important to highlight the performance of both model's support of time series diagnostic outcome. The comparison between both models is done in terms of root mean squared error (RMSE), mean absolute percentage error (MAPE) and mean absolute scale error (MASE) as depicted in Table 2. Apparently, the basic diagnostic tools obtained through manual model outperformed the optimal models due to the lowest obtained values of RMSE (12.69), MAPE (16.57) and MASE (0.69). These suggest that the manual model has higher goodness values and higher accuracy than the optimal time series model. Moreover, the autocorrelations of the residual's diagnostics were checked by applying the Ljung-Box test for both models. Ljung-Box test obtained insignificant residual's correlation with both models. Manual model's residual results (Q=14.21, df=20, and p-value=0.82) were obtained by Ljung-Box test and the values (Q=14.66, df=20, and p-value=0.80) belong to optimal model's residuals.

In comparison, the estimation of the parameters of manual and optimal time series models, both of them have significant parameters, as shown in Tables (Table 3 and Table 4). However, the manual model's parameters were found to be more significant (Table 3) than the optimal model's parameters (Table 4). Consequently, parameters generated by the manual time series model were affected more significant than the parameters generated by the optimal model. This indicates that the manual model had slightly better performance with time-series data. The study has unconfirmed the findings of (Ham, et al., 2017) that was employed *auto.arima( )* function to determine the time series model.

Table 3: Parameters estimates and their testing results of the Manual model

	Estimate	Std. Error	z value	Pr(> |z|)	
ma1	-0.909957	0.064616	-14.0824	< 2.2e-16	***
ma2	0.145944	0.063576	2.2956	0.0217	*
sar1	0.926239	0.102212	9.0620	< 2.2e-16	***
sma1	-0.818673	0.162108	-5.0502	4.414e-07	***

- - -

Signif. codes: 0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Table 4: Parameters estimates and their testing results of the Optimal model

	Estimate	Std. Error	z value	Pr(> |z|)	
ma1	-0.884333	0.064171	-13.7810	< 2e-16	***
ma2	0.127946	0.064044	1.9978	0.04574	*
sma1	0.135269	0.067033	2.0179	0.04360	*
sma2	0.182493	0.073128	2.4955	0.01258	*

- - -

Signif. codes: 0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Conclusion

In this investigation, the aim was to assess optimal (automatic) time series models, generated by *auto.arima( )* function in R. This paper has analysed the diagnostics of time series models and compared the diagnostic checking of optimal time series models to the manual ones. Time series data considered in the study represented monthly road traffic deaths in Oman from January 2000 to December 2018 with a total of 228 observations. These data were found to represent a non-stationary time series curve, significant seasonality, confirmed by Augmented Dickey-Fuller test.

Moreover, it was also shown that in an operational scenario, SARIMA $(0,1,2)(0,0,2)_{12}$ model resulted as the optimal model in this time series data. One of the more significant findings emerged from this study was that AIC, BIC values obtained for the model $(0; 1; 2)(1; 0; 1)_{12}$ had the best goodness of fit which was better than other models (including the optimal one). The primary diagnostic tools obtained through manual model outperformed the optimal model with the lowest values of RMSE (12.69), MAPE (16.57) and MASE (0.69). Furthermore, by checking residual diagnostics for both models, residuals of an optimal model are found less accurate than the manual one, even though there is insignificant autocorrelation of residuals for both models. In general, these findings suggest that the manual model has higher accuracy than the optimal time series model. The results of this research support the idea that practitioners may operate the *auto.arima( )* function from the *forecast* R package with high consideration of different models. It is highly recommended to assess and compare several models and employ diagnostic tools. Further research in this field, especially in *forecast* R packages, would be of great help to the computing discipline.

References

Box, G. E., & Jenkins, G. M. (1970). *Time series analysis forecasting and control.* San Francisco Holden-Day. Retrieved from http://openlibrary.org/books/OL4564956M

Box, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). John Wiley & Sons.

Cryer, J. D., & Chan, K.-S. (2008). *Time series analysis with application in R .* Springer.

Ham, S., Kim, S., Lee, N., Kim, P., Eom, I., Lee, B., . . . Yoon, C. (2017). Comparison of data analysis procedures for real-time nanoparticle sampling data using classical regression and ARIMA models. *Journal of Applied Statistics, 44*, 685-699.

Hyndman, R. J., & others. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting, 4*, 43-46.

Lavrenz, S. M., Vlahogianni, E. I., Gkritza, K., & Ke, Y. (2018). Time series modeling in traffic safety research. *Accident Analysis & Prevention, 117*, 368-380.

Manikandan, M., Prasad, V., Mishra, A. K., Konduru, R. K., & Newtonraj, A. (2018). Forecasting road traffic accident deaths in India using seasonal autoregressive integrated moving average model. *International Journal of Community Medicine and Public Health, 5*, 3962-3968.

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research, 1*, 1-22.

Ozaki, T. (1977). On the order determination of ARIMA models. *Applied Statistics*, 290-301.

Parvareh, M., Karimi, A., Rezaei, S., Woldemichael, A., Nili, S., Nouri, B., & Nasab, N. E. (2018). Assessment and prediction of road accident injuries trend using time-series models in Kurdistan. *Burns & trauma, 6*, 9.

Quddus, M. A. (2008). Time series count data models: an empirical application to traffic accidents. *Accident Analysis & Prevention, 40*, 1732-1741.

Raeside, R., & White, D. (2004). Predicting casualty numbers in Great Britain. *Transportation Research Record: Journal of the Transportation Research Board, 1897*, 142-147.

Szeto, W. Y., Ghosh, B., Basu, B., & O'Mahony, M. (2009). Multivariate traffic forecasting technique using cell transmission model and SARIMA model. *Journal of Transportation Engineering, 135*, 658-667.

Zhang, X., Pang, Y., Cui, M., Stallones, L., & Xiang, H. (2015). Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model. *Annals of Epidemiology, 25*, 101-106.