Forecasting of Road Accident in Kerala: A Case Study

Christine Maria Sunny, Nithya S, Sinshi k S, Vidya Vinodini M D, Aiswaria Lakshmi K G, Anjana S, TK Manojkumar*
Center for Excellence in Data Engineering and Computational Modeling
Indian Institute of Information Technology and Management-Kerala (IIITM-K)

***Abstract*—Traffic accidents are the foremost reason for death and injuries around the world, fatalities are still on the ascent in many creating nations including India. Data Analysis of road accidents makes a strong impact for taking preventive measure to overcome the mishap. In this manuscript, we have addressed the prediction problem of road accidents using time series analysis across all districts of Kerala. Time series analysis is useful in discovering the trends in road accidents which enables the prediction of future patterns. In the present MS, we used the time series road accidents data in Kerala, India for the period January 1999 – December 2016 to understand the patterns in the data and to develop appropriate model to predict about future patterns which may enable authorities to take preventive steps. We subsetted the data till 2013 December as training data for the model selection and rest of the data is used for model valuation. Two models discussed here are the "Holt-Winters (HW) exponential smoothing" and "Seasonal ARIMA (SARIMA)". Both the models will provide the forecast values within the confidence interval of the test data.**

Keywords—Forecast;TimeSeries;ARIMA Model ; Holt-Winters

I. Introduction

Many states in India is having issues related to road traffic concerns such as slow moving traffic, higher accident rates for past several years etc. The situation is getting worsened year by year due to the increasing population [1]. In India along with the population growth, other factors such as improved financial status resulted in the increased number of vehicles. This leads to increased number of road accidents and their resulting fatalities as a growing social and economic problem. The concerned authorities in Kerala are finding it difficult to address the issues such as monsoon maintenance works of the roads and drains. The absence of street lights on most roads has added to the difficulties faced by road users. Apart from the street conditions, other issues such as human errors, over-speed, lack of knowledge about rules, violation of traffic rules, vehicle conditions/unauthorized extra fittings, etc. also contribute significantly to road accidents[2]. The World Health Organization (WHO) has reported that traffic fatalities will be the third leading cause of deaths thus world wide by 2020.

Recent reports about road accidents in India by National Transportation Planning and Research Centre (NATPAC) indicate that after Maharashtra, Kerala positioned second among different states in country with respect to the road accident rate. According to data in 2004, the total number of road accidents reported is 41219 in Kerala and in 2005 come across to near 43000. Later the state implemented strict laws in roads, installed speed control devices/ speed breaks in roads, and also extended the road safety campaigning activities. The reported number of accidents went down due to taking immediate actions. Still when we look at the data of past three years 2014 the figures were 36282 and, coming to 2016 it becomes 39420 so it moving towards to 40K again. The recent trend is also showing the upward trend in the road accidents in Kerala. It is in this direction that this paper is prepared to study the trends, patterns and forecast of road traffic accidents in Kerala.

II. Methods

The review of literature for this work illustrates the papers that discussed about the method carried out for Road accidents, their challenges; scope for development and after that the investigation of something beyond late, contemporary ways to deal with forecasting, particularly with reference to Time Series (TS) analysis. In this regard many traffic accidents models were developed by many researchers. Developing countries are much more affected from traffic accidents than developed countries.

Jha et al gave a comparative study for traffic forecasting using time series analysis. They applied a regression model for prediction of causality in road accidents the results were compared with the results obtained from time series analysis. The comparison indicates that the time series is more useful in prediction with lesser error rates [3]. Mutangi et al. studied the road accident in Zimbabwe to develop a prediction model for future occurrences. The authors studied the quality of the model using different methods and found that ARIMA(0,1,0) is highly useful in representing and predicting the annual traffic accident rate of Zimbabwe[2][4]. Brajesh et al gave a model to find forecasted value of accident death. They used Damped Trend Exponential smoothing (DTES), ARIMA. From the results they concluded that ARIMA (1,1,1) is good for forecasting the accident mortality in India [5].

Traffic accident data of Britain was analyzed by M. A. Quaddus et al using ARIMA model. The authors adopted integer-valued autoregressive (INAR) and Poisson and Negative Binomial (NB) methods for the analysis. The study indicate that in terms of model goodness fit, both models

978-1-5386-4855-1/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: University of Thessaly. Downloaded on March 18,2025 at 08:32:53 UTC from IEEE Xplore. Restrictions apply.

ARIMA and INAR Poisson are comparable when we are using aggregated time series traffic accident data. The authors concluded that ARIMA model is inferior in performance to INAR Poisson model if the data is represented in disaggregated time series [6]. There are many studies which are addressing serial dependency issues of ARIMA using statistical methods. Different groups have introduced models to further improve the results such as DRAG model which is a special case of the ARIMA model, Auto Regressive (AR) model for road safety analysis etc. [7]. Dr.Bhuvana Vijaya(2014) analyzed mathematical formulation of total fatalities per vehicle with respect to vehicle ownership and total fatalities with respect to both vehicles and population. From the results the authors concluded that that a higher degree positive correlation is observed in Kerala compare to neighboring state of Tamil Nadu, even the population and area is lower than neighboring state[8]. We used data obtained from Kerala State Crime Records Bureau for the period of 18 years from 1999 Jan to 2016 for this study. The aim of the study is to identify the trends in the historical data and predict the future trends of the road accidents in Kerala. We used the R software for the analysis and prediction.

III. Theoretical Framework

A. Exponential Smoothing

This technique is used for steadily re-examining a forecast based on more recent experience is exponential smoothing. So that latest observations are given more weightage than previous observations. [9][10].

For limited forecasting, just for one month into the future the best method to use is single exponential smoothing. The single exponential smoothening assumes that that the data varies in a constant mean region, therefore the method is useful for data without any regular trend/pattern.

The equation corresponds to simple exponential smoothing is given by the relation:

$$"S_{t+1} = \alpha X_t + (1-\alpha) S_t \text{ "} \qquad (1)$$

The new forecast value is the sum of old value and the term corresponding to the error occurred for previous forecast. The weight factor exponentially diminishes with the estimation parameter "estimation of the parameter" α. It value lies between the 1 and 0, if it occurs to be 1 then past perceptions are totally ignored. Suppose the value is 0, the present perception completely avoided and smoothed value has only terms related to the past smoothed results.

The initial value of S_t shows a major role in figuring all the resulting values. Setting it to X_1 is one technique of initialization. When the value becomes smaller, it will be more important for the selection of initializing the S_t value

Double Exponential Smoothing and tripple exponential also useful in prediction and forecasting. If the data found to have a trend, then double smoothening is useful. In this method level and trend has to be taken care at each period. As discussed in the earlier case, level is represented as smoothed estimate of data of a period. Trend is also represented as estimate of the mavarage growth during the period [9][10].

If the data shows both trend and seasonality, the triple smoothening is useful. Compared to double smoothening, there will be another term to handle the seasonality. The set of equations which include these terms are known as "Holt-Winters" (HW) method. Literature indicates there are two different two main HW models proposed which are [9][10].

Multiplicative Seasonal model
Additive Seasonal model

Additive Seasonal Model is useful when the data having the property that seasonality can be additive. In this case the time series exhibits steady fluctuations for seasons in spite of the trends in entire data set.

According to earlier studies, the time series can be represented as:

$$Y_t = (b_1 + b_2 t) S_t + \epsilon_t \qquad (1)$$

Where

b_1 - permanent component (base signal).

b_2- represent linear trend

S_t- represent seasonal factor which is additive

ϵ_t- factor for representing error

The next period can be forecasted using following equation where

$$Y_t = R_{t-1} + G_{t-1} + S_{t-L} \qquad (2)$$

R_t is the overall smoothening factor, where as G_t and S_t represents trend and seasonal factors.

B. "Auto Regressive Integrated Moving Average" (ARIMA)

The forecasting method using ARIMA of a time series involves mainly two steps

- First step is analyzing the series
- Develop a suitable model which can forecast the data provided in the data set.

If the data is stationary, then only one can apply the ARIMA model. Since most of the time series we know are non-stationary. In this condition, the model is now called Auto Regressive Integrated Moving Average(ARIMA). The model can be represented as ARIMA (p,d,q) x (P,D,Q) model[11]. Here P represents the number of autoregressive terms, the number of seasonal differences is represented by the term"D" and number of moving averages for season is given by the term "Q".

The Auto Regressive (AR) model represent the current value of the series *Xt* can be represented in terms of previous p values

$$X_{t-1}, X_{t-2}, \dots\dots\dots\dots . X_{t-p} \qquad (3)$$

Here p is the parameter which defines the number of previous values required for forecasting the current value. "d" denotes the levels of differencing so that the initial series

Authorized licensed use limited to: University of Thessaly. Downloaded on March 18,2025 at 08:32:53 UTC from IEEE Xplore. Restrictions apply.

becomes a stationary one. The term “q” represents errors due to lagged forecast. The equation for prediction is given by

$$Y_t = b_0 + b_1Y_{t-1} + b_2Y_{t-2} + \cdots + b_pY_{t-p} + e_t \quad (4)$$

The seasonal part of an ARIMA model has the same structure as the non- seasonal part.

Seasonal ARIMA (SARIMA) can be used for modeling any time series which is homogeneous and non- stationary. The real practical problem is to choose appropriate values for p,d,q,P,D and Q. This problem was addressed by checking the auto correlation function (ACF) and also the partial auto correlation function (PACF) for the series. Plotting ACF and PACF provides the diagnosis check for the time series. We used the information from “Akaike Information Criteria” (AIC) and “Bayesian Information Criteria” (BIC) for selecting the best model for our study. Finally using this model we forecasted for the period 2017-2020.

IV. Evaluation of Result

We have used R software for the analysis and forecasting. The data is loaded and Fig 1 shows the time series plot of total reported accident cases. By looking at the trends we can identify that data is having additive seasonality. The trend is given in Fig. 2

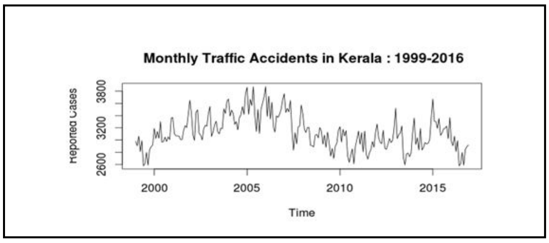

Fig. 1. Time Series Plot for Monthly Traffic Accident Data

Further the plot indicate that the non-stationary trends in the data [12]. The mean is not constant and increasing with time ADF (Augmented Dickey-Fuller) test on the initialization data (1999-2013) showed p-value of 0.01 suggesting further that the MTA data is non-stationary. First differencing the p value for ADF test is 0.01 which indicate the data became stationary.

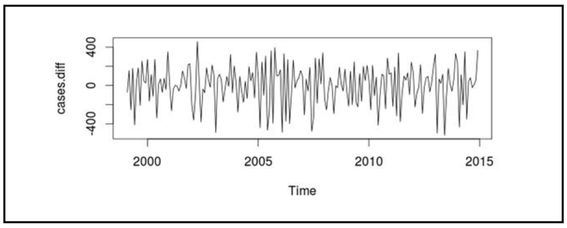

Fig. 2. Train Data after first difference

Fig 3 and Fig 4 shows the plot of ACF and PACF with data after differencing.

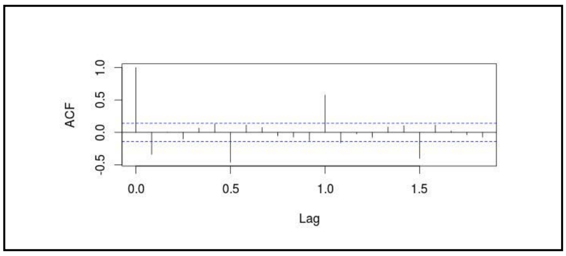

Fig. 3. ACF Plot for monthly traffic accidents after differencing

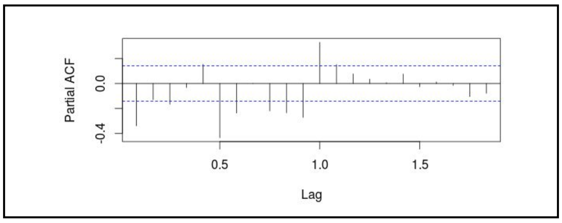

Fig. 4. PACF Plot for monthly traffic accidents after differencing

A. Model Selection

The seasonality of the differenced series can be investigated using above plots Fig. 3 and Fig. 4. The computation indicate that SARIMA (0,1,1) (2,0,0) reported lowest AIC and BIC[13]. The values corresponding to AIC and BIC are 2417.97 and 2434.23 respectively for the Monthly Traffic Accident data. Therefore this particular model was chosen for further analysis.

The residuals of SARIMA (0,1,1) (2,0,0) were further analyzed by Ljung Box Test. The results does not show any evidence for getting non zero auto correlation (p=0.92). ADF test proves that for p=0.01, the residuals are almost stationary. These results confirm that SARIMA (0,1,1) (2,0,0) is the final model.

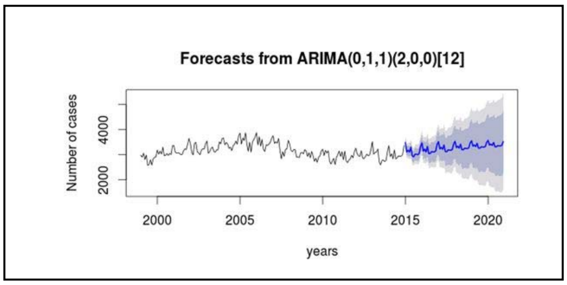

Fig. 5. Forecast Plot for Monthly Traffic Accident Data

Comparison with Test Data: The Fig.6 shows the clotted line mimics the data used for testing. The blue one shows the forecasted values.

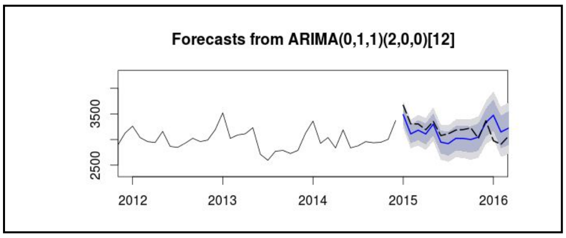

Fig. 6. Comparison Plot for Test data and Forecasted data

Authorized licensed use limited to: University of Thessaly. Downloaded on March 18,2025 at 08:32:53 UTC from IEEE Xplore. Restrictions apply.

ACF plots of the residuals are shown in Fig.7.

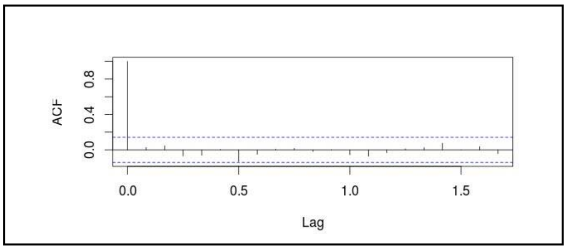

Fig. 7. ACF plots of the residuals

B. Analysis of deaths in Monthly Traffic Accidents.

Next we analyzed the number of death cases of MTA in Kerala. The Fig. 8 represents the data. The non stationary behavior of the data is clear from the plot.

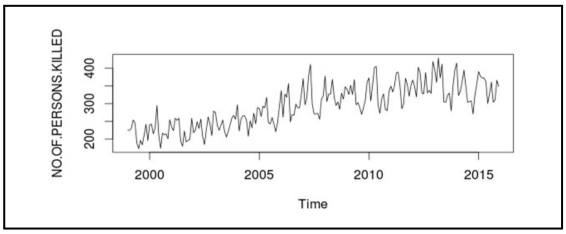

Fig. 8. Time Series plots of Killed persons in accidents

We further analyzed data using ADF to check whether the data is stationary. At p=0.01, the test rejected the null hypothesis further confirms the non stationary behavior of the death data. The ACF and PACF plot of differenced data once given in Fig.9 . It confirms the stationary of the data.

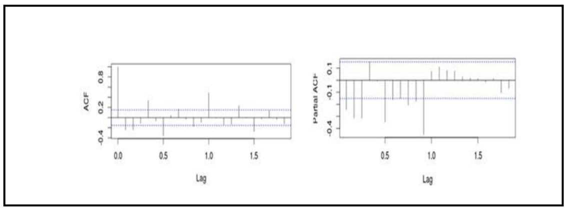

Fig. 9. ACF and PACF plots of Killed persons in accidents

According to the plot of ACF and PACF based model for forecasting is found to be SARIMA (0, 1, 1) (0, 0, 1).The analysis of residuals using LB statistics confirms the non-zero autocorrelation .

The stationary behaviors of residuals were confirmed by ADF test.

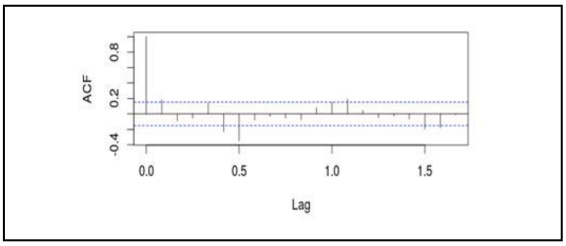

Fig. 10. ACF Plot of residuals

Thus we can say that the model selected is appropriate; therefore the model can be used for forecasting of the above cases. The time series plot of persons killed in road accidents shows the trend with additive seasonal component. Here we use HoltWinters function, a predictive model for forecasting. The estimated values of α, β and γ are 0.06, 0 and 0.3 respectively. The values of α, β and γ is used for filtering. The Fig.11 shows the Holt-Winters filtering plot for observed vs filtered value of no: of persons killed in road accidents.

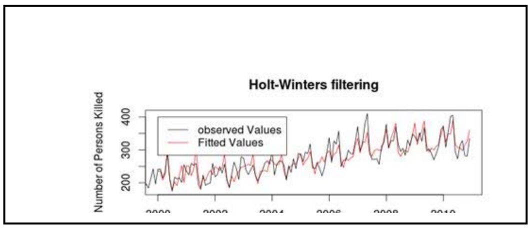

Fig. 11. Observed value vs Filtered value

From the plot further confirms success the Holt-Winters exponential method in predicting the seasonality. While calculating residuals Box-Ljung static gives DF=20, p-value=0.08, hence it indicate that residuals are stationary[14]. Fig.12 represents the forecasting using HoltWinter function.

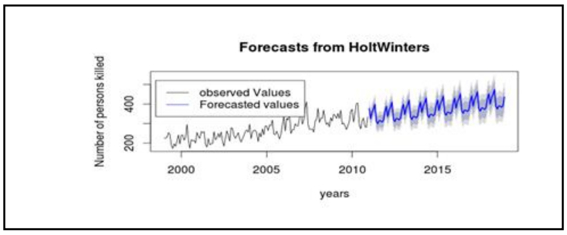

Fig. 12. Forecasting using HoltWinters Method

We further performed analysis on the number of persons injured in MTA data. The plot of the time series data given in the Fig.13, confirms that the data is non stationary and the mean of the data is changing with respect to time.

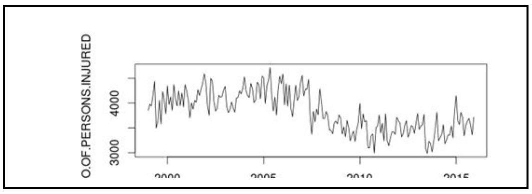

Fig. 13. Time Series Plot : No: of persons injured in MTA

The ACF and PACF plot of the differenced data is given in Fig. 14.

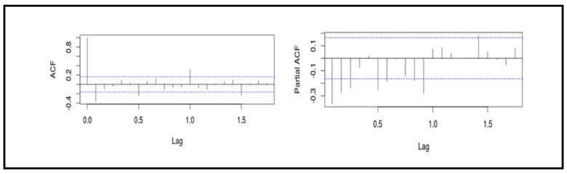

Fig. 14. ACF and PACF of the differenced series

The residuals were analyses using Ljung Box statistics. The Fig.15 shows the residual plot of ACF

Authorized licensed use limited to: University of Thessaly. Downloaded on March 18,2025 at 08:32:53 UTC from IEEE Xplore. Restrictions apply.

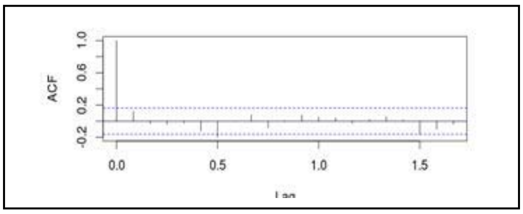

Fig. 15. ACF plot for residual

The time series plot of persons injured in road accidents shows the trend with additive seasonal component. Holt Winters function is used as a predictive model for forecasting which gives more accurate result. The Fig.16 shows the Holt-Winters filtering plot for forecasting for the years up to 2018 no: of persons injured in road accidents.

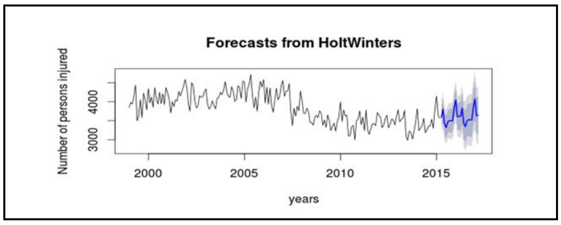

Fig. 16. Holt-Winters filtering plot for forecasting

V. Segmented Data

According to MTA in 2012 the number of accidents =36174, number of injuries=41915 and the number of death=4286, and which is higher than 2011. In the years 2013 and 2014 shows a downward trend in the number of death cases. This can be attributed to strict means in implementing law during that time. In 2016, reported maximum number of death cases and major death cases are due to two wheelers [15] by looking althrough the predicted value vs. original.

In 2003 the model overestimated by 11%.But this value reduced to 4% in 2007 and 0.3% in 2011. Interestingly the value got underestimated by 5% in 2015 and 8% in 2016.In 2017 predicted value comes closer to original value. A similar study was reported in UK. The country was able to reduce the number of casualties by 50%.The result indicate that the risk of an accident in monsoon season is greater than in dry weather.

TABLE I. Original Vs Predicted

YEAR	*Original*	*Predicted*
2003	39496	43960
2007	39917	41754
2011	35216	35333
2012	36174	35254
2013	35215	35469
2014	36282	35700
2015	39014	36867
2016	39420	36190
2017	38470	38932

VI. Conclusion

We analyzed the road accident data for total number of accidents in Kerala, number of death cases due to motor accidents and the number of injuries. All these subsets of data shows non stationary behavior which was confirmed by ADF test. Then these subset of data became stationary after differencing once. We employed Box- Jenkins and Holt Winter exponential smoothing on the data we got after differencing once. Holt-Winters exponential method gives accurate result while forecasting. The implications of these findings show that road accidents and person killed is increasing. Our conclusion is that strict measures of implementing laws will definitely reduce the road accidents.

References

[1] Bollapragada, R., Poduval, S., Bingi S, C. and Brahmbhatt, B.,"Solving Traffic Problems in the State of Kerala, India: Forecasting, Regression and Simulation Models", Vikalpa, 41(4), pp.325-343, December 2016

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

[3] Jha, K., Sinha, N., Arkatkar, S.S. and Sarkar, A.K.," A comparative study on application of time series analysis for traffic forecasting in India: prospects and limitations", Current Science (00113891), 110, no.(3), Feb 2016.

[4] Hurvich, C.M. and Tsai, C.L.,"Regression and time series model selection in small samples",Biometrika, 76(2), pp.297-307, June 1989 .

[5] Brajesh and Dr. Chander Shekhar. "Accidental mortality in India: Statistical models for forecasting",International Journal of Humanities and Social Science Invention, pages 35-45, 2015.

[6] Quddus, M.A.," Time series count data models: an empirical application to traffic accidents",Accident Analysis & Prevention, 40(5), pp.1732-1741,Sep 2008.

[7] Commandeur, J.J., Bijleveld, F.D., Bergel-Hayat, R., Antoniou, C., Yannis, G. and Papadimitriou, E., "On statistical inference in time series analysis of the evolution of road safety". Accident Analysis & Prevention, 60, pp.424-434.,Nov 2013.

[8] Vijaya, R.B. "Analysis of road accidents of southern states in India using smeed's model." International Journal of Research in Mathematics Computation, 1:13-19,2014.

[9] Kalekar, P.S.,"Time series forecasting using holt-winters exponential smoothing",Kanwal Rekhi School of Information Technology, 4329008, pp.1-13, December 2004.

[10] Balas, V.E., Jain, L.C.," Soft computing applications", In Proceedings *of* the 5th international workshop soft computing applications (SOFA) (Vol. 195, pp. 01-04),2013.

[11] Smith, M. and Agrawal, R.,"A Comparison of Time Series Model Forecasting Methods on Patent Groups". In MAICS (pp. 167-173),2015

[12] Zheng, X. and Liu, M.,"An overview of accident forecasting methodologies". Journal of Loss Prevention in the process Industries, *22*(4), pp.484-491,2009.

[13] Zhang, X., Pang, Y., Cui, M., Stallones, L. and Xiang, H.,"Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model". Annals of epidemiology, 25(2), pp.101-106,2015.

[14] Ahmed, M.S. and Cook, A.R.," Analysis of freeway traffic time-series data by using Box-Jenkins techniques" (No. 722), 1979.

[15] Official webportal of kerala police. http://www.keralapolice.org/public-information/crime-statistics/road-accident.

Authorized licensed use limited to: University of Thessaly. Downloaded on March 18,2025 at 08:32:53 UTC from IEEE Xplore. Restrictions apply.