International Journal of Traffic and Transportation Engineering 2022, 11(1): 1-7
DOI: 10.5923/j.ijtte.20221101.01

Prediction of Road Accidents Trend in Tanzania Using ARIMA Model: The Road Safety Implication by 2021-2030

Vitalis Agati Ndume[1,*], Edwin C. Rutalebwa[1], Angela-Aida K. Runyoro[2]

[1]Dar Es Salaam Institute of Technology, Tanzania
[2]National Institute of Transport, Tanzania

Abstract The purpose of this study is to predict the road accidents and justify whether Tanzania can reach the target of the global Second Decade of Action for Road Safety 2021-2030 calls. The study applied time-series modeling to determine and predict road traffic accidents patterns in the selected regions in Tanzania. Regions selection was based on those with the high rate of accidents. The secondary data obtained from Tanzania Road Safety Squad were used in analysis. Data were then loaded on R- packages for analysis. A time-series analysis using ARIMA Model was conducted to characterize and predict the frequency of road traffic accidents that lead to injury. The traffic accidents were categorized into four separate groups; these are accidents related to the car driver's behavior, motorcyclists, bicyclists and pedestrian. ARIMA model was used to model time series in each group from 2013 to 2021 and to predict the accidents up to 10 years later (2030). The analysis was carried out using R-4.1.1 statistical software package. The main contribution of our study in the field of road safety is estimation of number of death that can occur due to road accidents by 2030 which is estimated to decrease by 97%.

Keywords ARIMA Model, Road accidents, Time series, Driving error, Overtaking, Pedestrian accident

1. Introduction

The road traffic accident system can be considered to consist of essentially two predominant engineering sections, these are the roads and its environment and the vehicle including its contents but excluding driver; and two predominant human systems including driver on one side and pedestrian or form animal on the other side (Abdulhafedh, A. 2017). A road traffic accident therefore is a type of failure of one object when interacting with another object on the transportation protocol.

Traffic accidents' modeling is an aspect that requires deeper analysis and innovation to predict and if possible prevent the roads accident. A wide range of statistical approaches to traffic accident modeling are available. Abdulhafedh, A. (2017) provide a general overview of these models. The statistical modeling approach are meaningful for identifying most critical factors and association among the parameters for road accident but lack forecasting results of the factors contributing to the roads accidents (Kumar & Toshniwal, 2016; Meißner & Rieck, 2021). Another widely used approach is the data mining approach. This study focus on using mathematical structure of ARIMA model in predicting the road accidents in Tanzania for the next 10 years from 2021 and it uses more than 8-years' time series data in prediction. The ARIMA models have been used extensively in modeling traffic accidents data based on time series framework (Avuglah, R. K. etal., 2014)).

Time Series Modeling (TSM) is a method which deals with time dependent data. In this method data depends on the series of times where time refers to year, month, quarters, days, hours, minutes or seconds. The interest of TSM in this study is to predict the future of road accident in Tanzania by year 2030. This is in relation to the recent UN announcement of the 2nd Decade of action for road safety which call for each country member to half the road accidents and injuries by 2030 (Peden & Sminkey, 2004). According to (ERIC, 2019) the time series data is collected at different points in time. This is opposed to cross-sectional data which observes individuals, companies, etc. at a single point in time. Since such data points in time series are collected at adjacent time periods there is a potential for correlation between observations.

2. Study Methods and Material

2.1. Study Design

The data was obtained from accident data repository of

* Corresponding author:
ndumev@gmail.com (Vitalis Agati Ndume)
Received: Dec. 29, 2021; Accepted: Jan. 24, 2022; Published: Jan. 27, 2022
Published online at http://journal.sapub.org/ijtte

Tanzania and was proof checked by the division of Tanzania Road Safety Squad. A consent letter was approved and submitted to the Tanzania road safety squad with the title "research on road safety program in Tanzania the framework of IRAP".

The accidents data are collected as the event of road accident occurs and they are accumulated for some months before submission to the regional office for compilation. Each region is responsible for the data quality, and finally the data are submitted to the central head quarter for national reporting. A purposeful data gathering approach was meaning for this study. Either vast data was required for modeling.

2.2. Structure of the Region for Reporting Road Accidents

The region report for road accidents in Tanzania does not follow the government physical division of the regions. The ministry of Home Affairs divided Dar es Salaam region in three (3) divisions i.e. Ilala, Temeke and Kinondoni therefore the road accident data are compiled for the divisions and not as a single region of Dar es Salaam.

2.3. Data Analysis

The time-series analysis models were applied to model the observed frequency of accidents in the study regions and to predict the future accidents. According to (Abdulhafedh, 2017; Harris, 2013) the Autoregressive Integrated Moving Average (ARIMA) model formulation is appropriate in our study data since it's the best model used to model the time series. For this study the yearly total frequencies time series were used. As explained by (Avuglah & Harris, 2014) the ARIMA model was expressed by ARIMA (p, d, q), where the p, d, and q represented the number of autoregressive, differences, and moving average parameters, respectively

The ARIMA (p,d,q) equation is given by

$$\varphi_p(B)(1-B)^d Y_t = \theta_q(B)\epsilon_t \quad (1)$$

where $\varphi_p(B) = (1 - \sum_{i=1}^{p} \varphi_i B^i)$, $\theta_q(B) = (1 - \sum_{j=1}^{q} \theta_j B^j)$ and B is the Box-Jenkins operator given by $B^\tau X_t = X_{t-\tau}$ for an integer τ. The constants $\varphi_1, \varphi_2, \dots, \varphi_p$ are autoregressive parameters to be estimated and $\theta_1, \theta_2, \dots, \theta_q$ are the moving average parameters to be estimated. The variables $Y_1, Y_2, \dots, Y_t$ are a series of observed time series and $\epsilon_1, \epsilon_2, \dots, \epsilon_t$ are a series of unknown random errors (or residuals) that are assumed to follow a normal distribution. This model is applicable for a non-seasonal time series. For a seasonal time series, a modified model called SARIMA is appropriate.

Implementation of the Box-Jenkins methodology (in this case ARIMA) in practice, which most statistical software packages will perform, includes:

1. *Model Identification*: Using plots of the data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA models is selected. This amounts to estimating (or guesstimating) an appropriate value for d followed by estimates for p and q.
2. *Model Estimation*: The autoregressive and moving average parameters are found via an optimization method like maximum likelihood.
3. *Diagnostic Checking*: The fitted model is checked for inadequacies by studying the autocorrelations of the residual series (i.e., the time-ordered residuals).

Note that purpose of the differencing in equation (1) is to remove the trend and seasonal components in the series so that the resulting series is stationary. It does not model the trend in the series. The aim of this study is also to describe the trend and seasonal patterns of the road accidents. In order to model the trend and seasonal patterns, the time series is decomposed in to three components (assuming that the series has additive property): Trend, seasonal and stationary components:

$$Y_t = T_t + S_t + \epsilon_t \quad (2)$$

The modeling approach in this study is to first model the trend and seasonal variation, before modeling the short-term correlation in the stationary residual series. This involves a two- stage process:

1) The first stage is to estimate the trend and seasonal variation $\widehat{T_t} + \widehat{S_t}$, and
2) The second stage is to calculate the residual series $\widehat{\epsilon_t} = Y_t - (\widehat{T_t} + \widehat{S_t})$ which should be stationary, and model its short-term correlation using a time series model (in our case ARIMA model).

There are many methods for modeling trend and seasonal variation in a time series, and three of the most common ones are regression, moving average and differencing. In this study a regression method is used. The idea is to represent the trend and seasonal variation as

$$T_t + S_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_p z_{tp} \quad (3)$$

Where $(z_{t1}, z_{t2}, \dots, z_{tp})$ are known covariates and $(\beta_1, \beta_2, \dots, \beta_p)$ are the unknown regression parameters. For a time, series with linear trend and with no seasonal component, equation (3) becomes

$$T_t = \beta_0 + \beta_1 t + e_t \quad (4)$$

Where e_t are residuals and $\boldsymbol{t} = (t_1, t_2, \dots, t_n)$ is a vector of index of data points. For instance, if the time series involves 100 years of annual records, $\boldsymbol{t} = (1,2,3, \dots, 100)$.

3. Results and Discussion

Over the 9 years' observation (from 2013 to 2021), a total of 70039 road traffic accidents were reported. The time plot is given in Figure 1. The time plot exhibits a systematic change, therefore giving evidence of trend in the data. Moreover, the trend shows an exponential decay. This is reviled on Figure 2 when the accident cases are plotted on a logarithmic scale. Therefore we will be working with transformed data as $T_t^* = \ln(T_t)$ so that equation (4) becomes

$$T_t^* = \beta_0 + \beta_1 t + e_t \quad (5)$$

The fitted regression line is shown in equation (6). All coefficients are significant at (p-value<0.001). This shows that the number of road traffic accidents in Tanzania is decreasing at exponential rate of **0.360** per year. With this assumption Tanzania may go beyond the expectation of the second decade of action of road safety 2021-2030.

$$T_t^* = 10.371 - 0.360t \quad (6)$$

The results in Table 1 show number of accidents per year for each category. The result indicates that over a total of 71216 road traffic accidents that were reported, about 36.72% were due to drivers' negligence, 20.98% were due to motorcyclists, and 8.57% were due to speeding respectively. These three categories constitute more than (66.27%) of all road traffic accidents. This is also depicted in Figure 3. This calls for more intervention in human factor in road safety. At the same time road engineering design such as Road condition (3.70%), Road blocks (3.59%) and Rumination (1.56%) remain with moderate contribution. Even though pulling carts (0.26%), livestock (0.23%) and passenger (0.17%) have low contribution but they cannot be neglected in the intervention. It is noted that drugs such as alcohol (0.79%) remain in lower side of accident contribution. The frequency of driver error was very high in year 2013 and was very much reduced in year 2021 (data of September 2021). Such reduction might have been brought forward due to imposing serous rules by Land Transport Regulatory Authority.

Table 1 list 15 category of factors that associate with road accident. It is surprising that accident due to driver error (26149) and motocyclist (14943) account for more than 50% of the total road accidents.

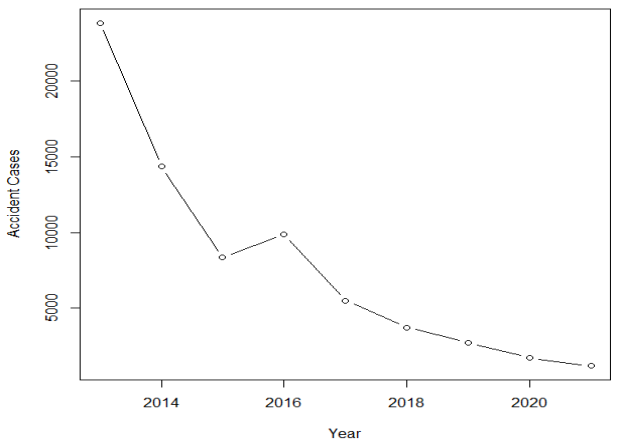

Figure 1. Time plot of accident cases in Tanzania from 2013 to 2021 with accident cases on a linear scale

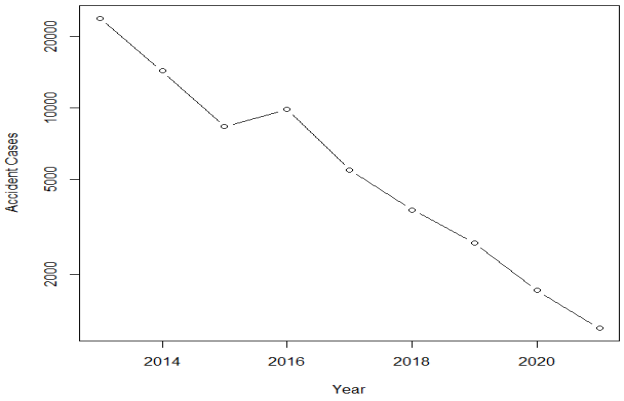

Figure 2. Time plot of accident cases in Tanzania from 2013 to 2021 with accident cases on a logarithmic scale

Table 1. Number of accidents per year for each category

Category/Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total	%	Cum. %
Driving error	8476	5053	2868	3624	2169	1542	1160	753	504	26149	36.72%	36.72%
Motor cycle	5118	3163	2009	1808	1135	725	492	317	176	14943	20.98%	57.70%
Speed	2039	950	691	730	489	364	288	272	282	6105	8.57%	66.27%
Car default	1623	990	502	755	375	197	105	80	70	4697	6.60%	72.87%
Pedestrian	1313	979	470	609	385	259	162	93	20	4290	6.02%	78.89%
Overtake	1475	800	474	638	262	217	236	91	45	4238	5.95%	84.84%
Cyclist	979	689	410	447	205	103	72	33	32	2970	4.17%	89.01%
Road condition	951	612	298	404	184	80	65	23	20	2637	3.70%	92.72%
Road blocks	907	489	341	533	117	120	32	9	9	2557	3.59%	96.31%
Rumination	484	285	104	123	47	31	19	14	4	1111	1.56%	97.87%
Alcohol	91	91	66	97	84	61	46	13	11	560	0.79%	98.65%
Crosing Railway	102	99	40	22	7	10	8	3	3	294	0.41%	99.07%
Fire	52	52	11	25	18	14	10	4	5	191	0.27%	99.33%
Pulling Cart	96	41	22	12	7	7	0	1	1	187	0.26%	99.60%
Livestock	94	18	20	9	13	1	3	4	4	166	0.23%	99.83%
Passenger	42	33	11	20	6	1	6	1	1	121	0.17%	100.00%
Total	**23842**	**14344**	**8337**	**9856**	**5503**	**3732**	**2704**	**1711**	**1187**	**71216**	**100.00%**	

The result in Figure 3 indicate that in each road factor the year 2013 was characterized with high road accident while year 2020 and year 2021 remain constant in almost all categories. Motorcyclist and driver error are among the most factors causing road accidents in the country while cyclists and passengers remain with low causal of road accidents.

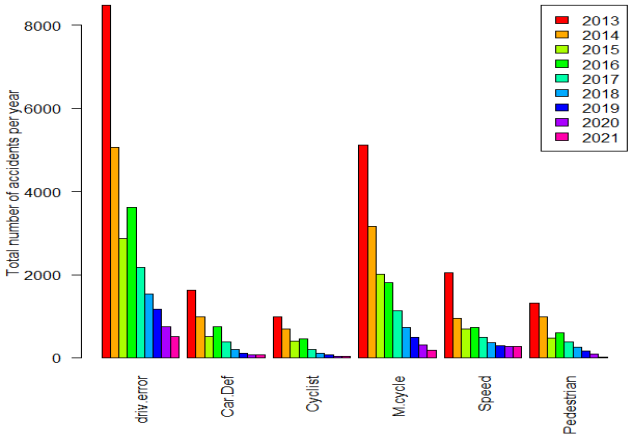

Figure 3. Total number of accidents per year for each category for the year 2013 to 2021

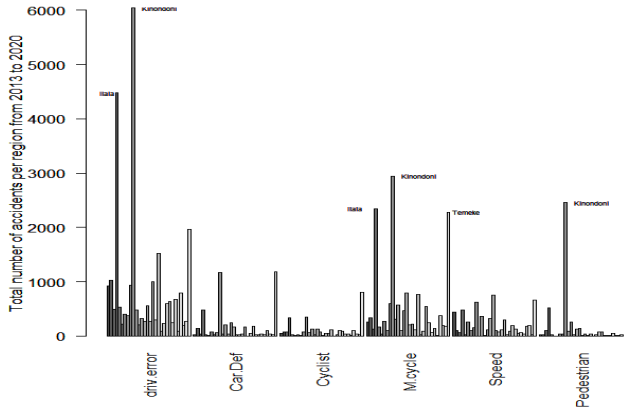

Figure 4. Total number of accidents per region from year 2013 to 2021

Three regions; Kinondoni, Ilala and Temeke; have the highest number of road traffic accidents as depicted in Figure 4 with Kinondoni region leading in the driving error, motorcycle and pedestrian. It is springing that road accident causative by higher speed remain almost constants in three years since 2019-2021.

From Figure 3 it can be observed and concluded that for the study period, there is generally a clear decreasing trend in the number of accidents per year for each category. The trends in each category can be described using equation (5) with $\boldsymbol{t} = (1,2,3,\ldots,8)$. Figure 5 depict the raw data for driving negligence variable for the case of Kindononi region plotted on the linear scale and Figure 6 corresponds to same case plotted on the logarithmic scale. The empirical regression equation for this case is given by equation (7).

$$T_t^* = 7.729 - 0.363t \qquad (7)$$

The empirical regression equation (7) summarized in table 2 tells us that the number of accidents due to driving negligence at Kinondoni region is approximately decreasing exponentially at the rate of about 0.363 (Coefficient (β_1)) accidents per year. Those due to motorcycle are reducing at rate 0.508, car default is reducing at the rate of 0.415 and pedestrian are reducing at rate of 0.447. The 95% confidence interval of this estimate of decreasing rate is (0.142, 0.585) and the p-value of this rate is less than 0.01, showing that the estimated value is significant at 1% level. The regression model for data from Ilala shows that driver error is reducing exponentially at rate of 0.468 a bit higher than those of Kinondoni while Motor cycle decrease at rate of 0.477.

Likewise, regression model for data from Temeke shows much better performance in reducing accidents. The driver error is reducing at rate of 0.562, Motorcycle at rate of 0.572 and car default is reducing at rate of 0.684 per year.

Table 2. Empirical regression summary of selected category

	KINONDONI			
Variable	Intercept (β_0)	Coefficient (β_1)	p-value	Remarks
Driving error	7.729	-0.363	<0.01	
Motor cycle	7.498	-0.508	<0.01	
Car default	6.279	-0.415	<0.01	
Pedestrian	7.292	-0.447	<0.01	
	ILALA			
Variable	Intercept (β_0)	Coefficient (β_1)	p-value	Remarks
Driving error	7.982	-0.468	<0.01	
Motor cycle	7.386	-0.477	<0.01	
Pedestrian	6.082	-0.563		
	TEMEKE			
Variable	Intercept (β_0)	Coefficient (β_1)	p-value	Remarks
Driving error	7.411	-0.562	<0.01	
Motor cycle	7.562	-0.572	<0.01	
Car default	7.154	-0.684	<0.01	

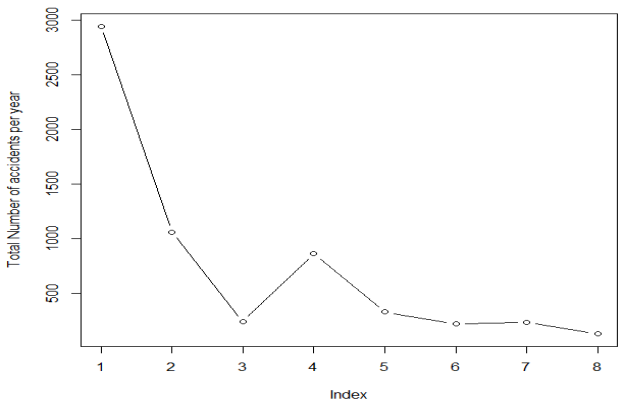

Figure 5. Total number of accidents per year for the category of driving negligence in the Kinondoni region, in Dar es Salaam with accident cases plotted on a linear scale

Figure 5 Depict linear scale plotting while Figure 6 show logarithmic scale plotting for Driving errorin causative of road accident at Kinondoni. The data show a steep drop in first year before it rises up at fourth year. The reducing rate remains almost constant from firth to eighth year. These two drawings depicture the whole picture of the road accident trend in the country.

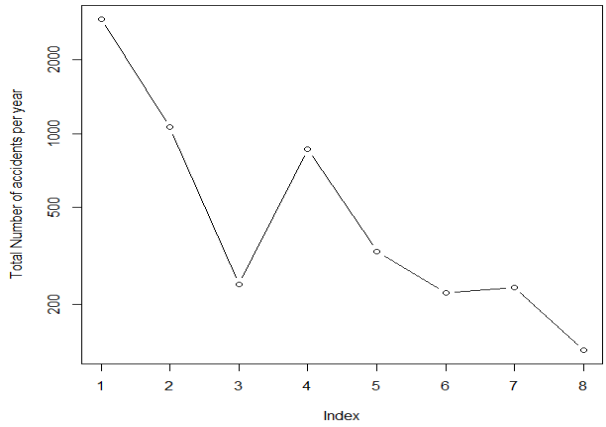

Figure 6. Total number of accidents per year for the category of driving negligence in the Kinondoni region, in Dar es Salaam with accident cases plotted on a logarithmic scale

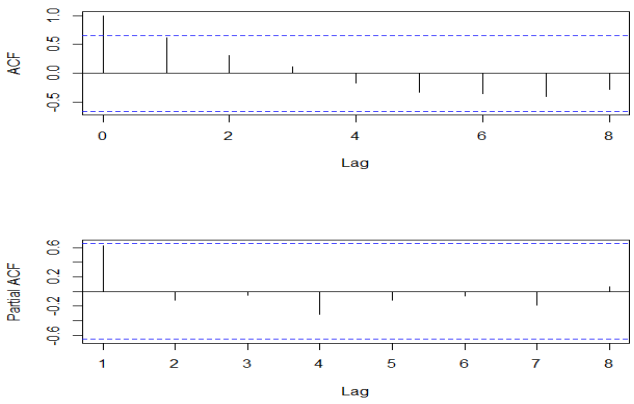

Figure 7. Autocorrelation function (above panel) and partial autocorrelation (below panel) of a transformed number of accidents series

The ARIMA model is implemented on the transformed data $Y_t^* = \ln(Y_t)$. The autocorrelation function (ACF) plot and partial autocorrelation function (PACF) plots (see Figure 7) suggest ARIMA (0,d,0) model. Using "auto.arima" function in R-package, the fitted model is ARIMA (0,1,0). This model is used for forecasting. The forecasted logarithm of the number of accidents for the next nine (9) year is shown in Figure 8. The ARIMA (0,1,0) entails that the annual numbers of traffic accidents in Tanzania can be model using trend component only in equation (2). The trend component for the transformed data is given by equation (6). Transforming equation (6) back to the original scale we have

$$Y_t = e^{10.371-0.36t} \quad (8)$$

for $\boldsymbol{t} = (1,2,3,\dots)$. These time steps could be months, years or weekly. In this study the annual number of accidents are used, hence $\boldsymbol{t} = (1,2,3,\dots)$ are time steps in year. Figure 9 depict the actual time series and the estimated time series of the number of traffic accidents in Tanzania for the period 2013 to 2021.

Equation (8) can be written in a more operational form as

$$Y_t = Y_0 e^{-0.36t} \quad (9)$$

where Y_0 is the number of accidents for reference year.

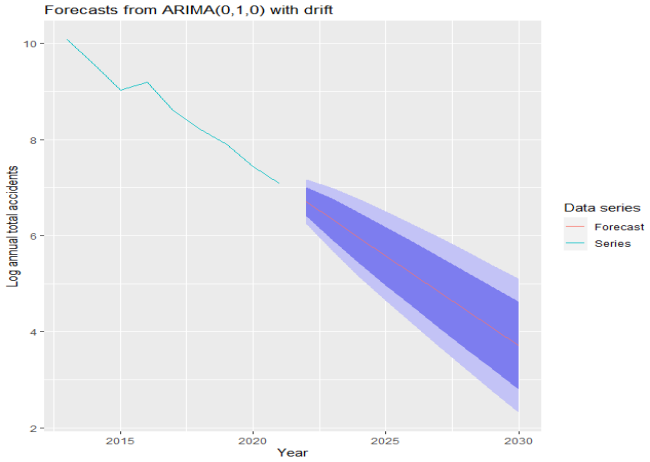

Figure 8. ARIMA(0,1,0) Forecasting plotting

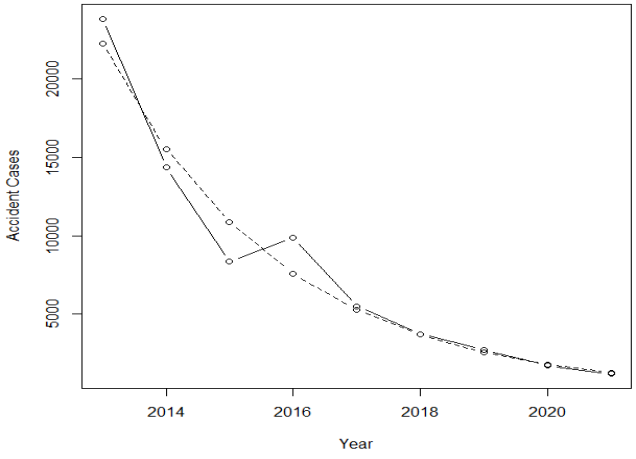

Figure 9. Time series (solid line) and estimated time series (dashed line) of the number of traffic road accidents in Tanzania for the period 2013 to 2021

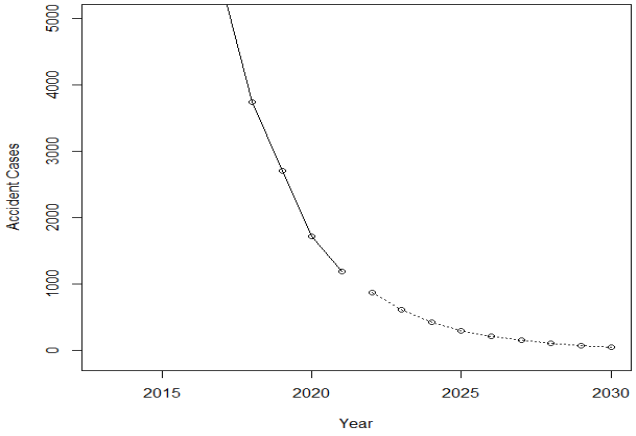

Figure 10. Time series (solid line) and forecasted time series (dashed line) of the number of traffic road accidents in Tanzania for the period 2013 to 2021 and the period 2022 to 2030 respectively

Equation (8) or (9) can be used to predict the number of traffic accidents in the coming year, setting t = 1 at the year 2013 or any referencing year (say 2020) and Y_0 is equal to number accidents recorded in the reference year (1711). For instance, the expected number of traffic accidents in year 2030 is obtained by setting t = 10 in equation (9) with Y_0 = 1711, which gives an estimate of 47 (97% decrease) number of traffic road accidents. Figure 10 depict the number of traffic road accident series (solid line) and the forecasted (dashed line) for the period from 2013 to 2021 and from 2022 to 2030 respectively. From figure 8 it is obvious that the precision of prediction decrease as you the time increase. We observe good prediction between 2020 and 2025 and more discrepancy from 2025 to 230.

4. Conclusions and Recommendations

The ARIMA model was successfully implemented in forecasting the general trends of road accidents in Tanzania. It has been successfully adopted in analysis of road accident data category especially in predict number of accidents by 2030. Our findings lead to a conclusion that road accidents are decreasing exponentially at a rate of 0.360 per year. It is forecasted that the number of the road accidents can decrease to 47 (97%) by 2030 which is above the preset target of 50%. The findings also conclude that with the annual number of road traffic accidents, the time series does not have significant seasonality and moving average components. Furthermore, even though road engineering design (Road condition) category of accident remains moderate it is claimed to be the source of other accident categories. Intervention on major three categories of causal for road accidents including driver error, motorcyclist and driving speed must be paid attention and given more priority. We propose further study on modeling and prediction of impact of road design and road asset planning in determining traffic flow based on the increasing number of cars on road traffic. Other more studies can be conducted on traffic congestion simulation in relation to pedestrian characterization and its impact on national economy.

ACKNOWLEDGEMENTS

We acknowledge the department of Road Safety Squad in allowing and providing such data for this study. We extend our acknowledgement to Dar es Salam Institute of Technology for supporting fund for publication ofthis result.

Declaration

Until the time of submitting this job we declared no conflict of interest in our work.

REFERENCES

[1] Abdulhafedh, A. (2017). Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, *07*(02), 190–205. https://doi.org/10.4236/jtts.2017.72014.

[2] Asalor, J. O. (1984). A general model of road traffic accidents. *Applied Mathematical Modelling*, *8*(2), 133–138. https://doi.org/10.1016/0307-904X(84)90066-0.

[3] Avuglah, R. K., & Harris, E. (2014). Application of ARIMA Models to Road Traffic Accident Cases in Ghana. *International Journal of Statistics and Applications*, *4*(5), 233–239. https://doi.org/10.5923/j.statistics.20140405.03.

[4] ERIC. (2019). *Introduction to the Fundamentals of Time Series Data and Analysis*. UPTECH. https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/.

[5] Harris, E. (2013). Modeling annual Coffee production in Ghana using ARIMA time series Model. *Modeling Annual Coffee Production in Ghana Using ARIMA Time Series Model*, *2*(7), 175–186. https://doi.org/10.18533/ijbsr.v2i7.129.

[6] Kumar, S., & Toshniwal, D. (2016). A novel framework to analyze road accident time series data. *Journal of Big Data*, *3*(1). https://doi.org/10.1186/s40537-016-0044-5.

[7] Meißner, K., & Rieck, J. (2021). Multivariate Forecasting of Road Accidents Based on Geographically Separated Data. *Vietnam Journal of Computer Science*, *8*(3), 433–454. https://doi.org/10.1142/S2196888821500196.

[8] Peden, M., & Sminkey, L. (2004). World Health Organization dedicates World Health Day to road safety. *Injury Prevention*, *10*(2), 67. https://doi.org/10.1136/ip.2004.005405.

[9] Peixeiro, M. (2019). *The Complete Guide to Time Series Analysis and Forecasting | by Marco Peixeiro | Towards Data Science*. https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775.

[10] Wang, Y., Wei, F., Sun, C., & Li, Q. (2016). The Research of Improved Grey GM (1, 1) Model to Predict the Postprandial Glucose in Type 2 Diabetes. *BioMed Research International*, *2016*. https://doi.org/10.1155/2016/6837052.

[11] Yang, X., Zou, J., Kong, D., & Jiang, G. (2018). The analysis of GM (1, 1) grey model to predict the incidence trend of typhoid and paratyphoid fevers in Wuhan City, China. *Medicine (United States)*, *97*(34). https://doi.org/10.1097/MD.0000000000011787.

[12] Abdulhafedh, A. (2017). Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, *07*(02), 190–205.https://doi.org/10.4236/jtts.2017.72014.

[13] Asalor, J. O. (1984). A general model of road traffic accidents. *Applied Mathematical Modelling*, *8*(2), 133–138. https://doi.org/10.1016/0307-904X(84)90066-0.

[14] Avuglah, R. K., & Harris, E. (2014). Application of ARIMA Models to Road Traffic Accident Cases in Ghana. *International Journal of Statistics and Applications*, *4*(5), 233–239. https://doi.org/10.5923/j.statistics.20140405.03.

[15] ERIC. (2019). *Introduction to the Fundamentals of Time Series Data and Analysis*. UPTECH. https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/.

[16] Harris, E. (2013). Modeling annual Coffee production in Ghana using ARIMA time series Model. *Modeling Annual Coffee Production in Ghana Using ARIMA Time Series Model*, *2*(7), 175–186. https://doi.org/10.18533/ijbsr.v2i7.129.

[17] Kumar, S., & Toshniwal, D. (2016). A novel framework to analyze road accident time series data. *Journal of Big Data*, *3*(1). https://doi.org/10.1186/s40537-016-0044-5.

[18] Meißner, K., & Rieck, J. (2021). Multivariate Forecasting of Road Accidents Based on Geographically Separated Data. *Vietnam Journal of Computer Science*, *8*(3), 433–454. https://doi.org/10.1142/S2196888821500196.

[19] Peden, M., & Sminkey, L. (2004). World Health Organization dedicates World Health Day to road safety. *Injury Prevention*, *10*(2), 67. https://doi.org/10.1136/ip.2004.005405.

[20] Peixeiro, M. (2019). *The Complete Guide to Time Series Analysis and Forecasting | by Marco Peixeiro | Towards Data Science*. https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775.

[21] Wang, Y., Wei, F., Sun, C., & Li, Q. (2016). The Research of Improved Grey GM (1, 1) Model to Predict the Postprandial Glucose in Type 2 Diabetes. *BioMed Research International*, *2016*. https://doi.org/10.1155/2016/6837052.

[22] Yang, X., Zou, J., Kong, D., & Jiang, G. (2018). The analysis of GM (1, 1) grey model to predict the incidence trend of typhoid and paratyphoid fevers in Wuhan City, China. *Medicine (United States)*, *97*(34). https://doi.org/10.1097/MD.0000000000011787.