Modelling the Number of Road Accidents in Kenya Using Time Series Analysis: A SARIMA Approach

Abstract

Road transport is a vital mode of mobility in Kenya. Despite existing research and numerous interventions, road accidents continue to pose a significant challenge. This study investigates the trend and future projection of road accidents using a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. Secondary data from the National Transport and Safety Authority (NTSA) from 2019 to 2023 was analyzed using R (version 4.3.1). The optimal model, SARIMA (0,1,1)(1,0,0)[12], demonstrated high forecasting accuracy. The study provides forecasts for the next 24 months and offers insights for policymakers and safety planners to mitigate accident risks.

1 Introduction

Road accidents in Kenya have been a persistent public safety concern, contributing to considerable socioeconomic losses. Time series forecasting is crucial for informed policy and strategic planning. Road accident means an event that occurs during the movement, and with the participation of a vehicle on a road, in which people are killed or injured, vehicles, equipments or goods are damaged, or any other material damage is caused. Techniques of analysing time series data have been widely applied on different areas (sectors) namely; tourism, climate, GDP, road accidents, crop yields

among other. However, focus on the impact of road accidents a time Series based road traffic accidents forecasting via SARIMA and Facebook prophet model with potential changepoints in Ghana[2]. They found that besides the safety controls, experience of the drivers, the state of the roads is a critical factor towards occurrence of road accidents.A comparison between SARIMA based on time series model for forecasting road accidents in Pakistan. The researcher found that persistent cause of road accidents was weather conditions which altered create tension on the surface[3]. The roads in Kenya plays a vital role in the society not only to accomplish day to day activities but also the dynamic growth of the economy in the country. Besides this, it also has negative impacts that has resulted to loss of life among many people who use road as their mode of transport. The report from National Transport and safety Authority (NTSA) [6] shows that nearly 3000 people die per annum, however, they have not clearly stated the main cause of road accidents. Despite the efforts to improve road infrastructure, road accidents have continued to occur almost with haste, even with the introduction of speed control devices, safety belts and other reforms in the sector. This project focused on assessing the trend of road accidents for the period January 2019 to December 2023 and forecast them for the next 24 months. Time-series is a collection of data points over a set period. Models for continuous data are very well developed. Real-valued time series models, such as the auto regressive integrated moving average (ARIMA) model, introduced by Box and Jenkins have been used to model time series count data in many applications over the last few decades Therefore ARIMA models can account for various patterns, such as linear or nonlinear trends, constant or varying volatility, and seasonal or non-seasonal fluctuations. It is also easy to implement and interpret, as they only require a few parameters and assumptions.The SARIMA model is denoted as SARIMA(p,d,q)(P,D,Q,s), where the parameters encapsulate both non-seasonal and seasonal aspects of the data. Here, 'p', 'd', and 'q' correspond to the autoregressive order, the degree of non-seasonal differencing, and the moving average order, respectively. These parameters function analogously to those in the ARIMA model, capturing the non-seasonal dependencies and trends within the data. The seasonal components are represented by 'P', 'D', and 'Q', which denote the seasonal autoregressive order, the degree of seasonal differencing, and the seasonal moving average order, respectively. The parameter 's' signifies the length of the seasonal cycle, such as 12 for monthly data with annual seasonality.The SARIMA model is a widely used methodology in time series analysis to identify trends and make forecasts based on seasonal data. This approach demonstrates superior precision in forecasting seasonal time series data. The SARIMA model has been effectively utilized in several domains over the last thirty years; nevertheless, it includes specific limitations. The applicability of the SARIMA model is limited to linear time

series data models, as it cannot effectively handle nonlinear patternse, according to [8] Houston and Richardson [5], 2002; Noland et al.,[9] 2006). The research on the road traffic accidents predictions based on SARIMA-LSTM model [4], [10] was greatly attributed towards this research. By comparing the performance of three models, SARIMA, LSTM and SARIMA-LSTM, in RTAS prediction in Jilin Province, we find that the SARIMA-LSTM model significantly outperforms the SARIMA and LSTM models alone. For example, using 2019 data, the RMSE, MAE, and MAPE of the SARIMA-LSTM model improved by 51.79[11] Time series modelling of road traffic accidents in West Arsi Ethiopia.The mean of 16.83 and standard deviation of 5.764 for the total accident was served the minimum and maximum record of road traffic accidents is 4 and 33 respectively. By differencing data one time,(2, 1, 3) model was fitted for making a two-year ahead forecast. Proper model adequacy checking was done. Two-year ahead forecasts showed that October, January, and April 2021 are the months with the most prominent values. Even if the trend in total accidents was decreasing there is still a need to pay more attention in order to prevent the occurrence of accidents related to road traffic accidents. [1]Despite substantial progress in road safety, road traffic fatalities (RTFs) continue to be a persistent issue in Australia. This study aims to forecast RTFs trends up to 2050 by analyzing factors such as geographic location, age, gender, speed limits, and time of occurrence. Utilizing historical data from 1989 to 2024, fatalities were categorized by road user type, demographics, and day of the week. The Facebook Prophet time series model, incorporating categorical variables like region, age, and speed limits, was employed to predict future trends. This study analyzed the trend of road accidents in Kenya and fitted a suitable SARIMA model for the road accidents data in Kenya. .And also aims to identify the trend and predict road accident frequencies using robust time series models.

2 Research Methods

This study employs a quantitative research design to model and forecast the number of road accidents in Kenya. The secondary data was obtained from the National Transport and Safety Authority (NTSA), which maintains comprehensive records of road accident statistics across the country. A time series is a sequence of observations recorded at regular time intervals, often used to track the evolution of phenomena over time. In this study, the time series comprises monthly records of road accidents in Kenya, as provided by the National Transport and Safety Authority (NTSA). Time series analysis aims to understand patterns such as trends and seasonality

within the data and to produce accurate forecasts. To achieve this, the study employs the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, a powerful extension of the ARIMA model [7]. The analysis begins with an exploratory examination of the data. Plotting the series reveals visible trends and possible seasonality, which suggests that the series is non-stationary. Stationarity—where the statistical properties of the series like mean and variance remain constant over time—is a crucial prerequisite for SARIMA modeling. To formally test for stationarity, the Augmented Dickey-Fuller (ADF) test is applied.[2].The test indicates that the series is non-stationary, prompting the application of first-order differencing (d=1) to eliminate the trend. Seasonal differencing is not required (D=0), as the seasonal pattern is effectively captured by the seasonal autoregressive structure of the model. . The first difference process The first difference process

$\bigtriangledown X_t$ consists of $\bigtriangledown X_t = X_t - X_{t-1}$ The second difference process

$\bigtriangledown^2 X_t$ consists of $\bigtriangledown^2 X_t = \bigtriangledown(\bigtriangledown X_t) = \bigtriangledown X_t - \bigtriangledown X_{t-1} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}$ In general, the d^{th} difference process $\bigtriangledown^d X_t$ consists of

$$\bigtriangledown^d X_t = \bigtriangledown(\bigtriangledown^{d-1} X_t) = \bigtriangledown^{d-1} X_t - \bigtriangledown^{d-1} X_{t-1} for$$

d = 1, 2, ... processes appeared as a generalization model of ARIMA, including a seasonal part. Generally, as $s_1, ..., s_n$, n integers, then a process X_t is a SARIMA(p,d,q) integrated auto regressive seasonal moving average process following the equation below

$\phi_p(B)\Phi_p(B^s)^{dD}{}_S X_t = \theta_q(B)\Theta_Q(B^S)Z_t$
(1)

for all $t \geq 0$, $\phi_0 = 1, \theta_0 = 1$ where, $\phi(B), \theta(B)$ are polynomials whose roots are of modulus higher than 1. This form includes the ARIMA models as it is enough to take n=d and $s_1 = ... = s_n = 1$.
For any $t \geq 0$ where only one seasonal factor(s) intervenes, either applied to an ARMA process in the first case, or applied to an ARIMA process in the second case. where:

P (AR): Seasonal Autoregressive order
D (I): The degree of seasonal differencing
Q (MA): Seasonal Moving average order
S: Length of the seasonal cycle
p: autoregressive order
d: The number of times the raw observations are differenced; also known as the degree of non- seasonal differencing.
q: Moving average order
The below indicated model is as follows:
$SARIMA(P, D, Q, S) \times (p, d, q)$ Where p, d, q are non-seasonal parameters; P, D, Q are seasonal parameters; and s is the seasonal period (12 for monthly data) [12]. Following stationarity, model identification is performed by examining the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. A strong spike in the ACF at lag 1 indicates the presence of a non-seasonal moving average component (q=1), while the PACF reveals a significant seasonal spike at lag 12, suggesting the need for a seasonal autoregressive term (P=1). Based on these insights, a SARIMA Parameter estimation for the model is conducted using the Maximum Likelihood Estimation (MLE) method. To validate the model, diagnostic checks are performed. Residuals are analyzed to ensure they resemble white noise—that is, they should exhibit constant variance, no autocorrelation, and a mean close to zero. Residual plots show no noticeable patterns or heteroskedasticity. The ACF and PACF plots of the residuals support the absence of significant autocorrelation. Furthermore, the Ljung-Box test confirms the independence of residuals, and normality is verified through the Shapiro-Wilk test and Q-Q plots.For model selection, the Akaike Information Criterion (AIC) is used to compare alternative models. The analysis was conducted using R statistical software, version 4.3.1, leveraging its robust capabilities for time series modeling, particularly the SARIMA approach.

3 RESULT AND DISCUSSION

3.1 Model Identification

The ACF plot of the series showed strong positive autocorrelations at multiple lags and a seasonal pattern, suggesting non-stationarity and the presence of seasonality. After applying seasonal differencing, the ACF values declined substantially, as shown in figure 1 indicating that stationarity was achieved. The ACF of the residuals from the SARIMA model showed no significant spikes, confirming that the model effectively captured the underlying structure of the time series. The PACF plot of

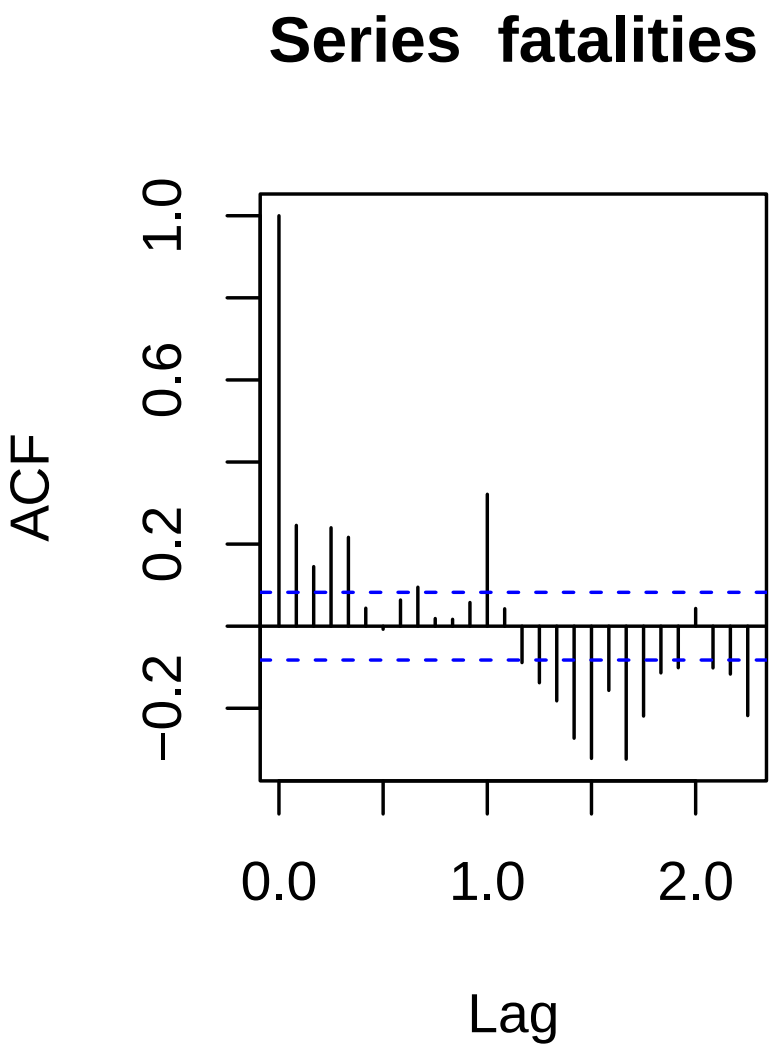

Figure 1: ACF Plot after differencing

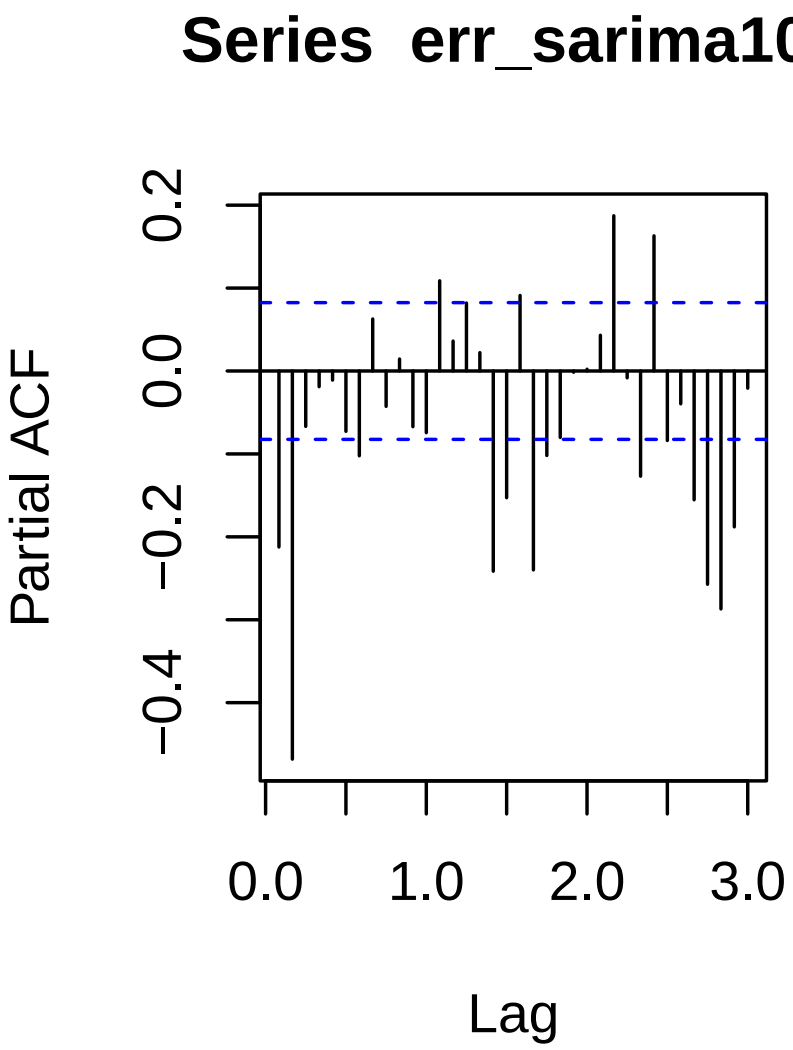

Figure 2: PACF plot after differencing

the original series showed significant spikes at lags 1 and 2, suggesting the presence of both short-term and seasonal autoregressive components. After applying seasonal differencing, the PACF pattern simplified, with a significant spike only at lag 1. This supported the inclusion of non-seasonal AR(1) and seasonal AR(1) terms in the SARIMA model. The PACF of the residuals exhibited no significant lags, indicating that the model adequately accounted for autocorrelations. Selection of the best model for forecasting number of people involved in road accidents.

3.2 Selection and Parameter Estimation

Table 1: ARIMA Model AIC Comparison

Model	**AIC**
ARIMA(0,1,0) with drift	857.0032
ARIMA(1,1,0)(1,0,0)[12] with drift	839.213
ARIMA(0,1,0)	853.4417
ARIMA(1,1,0) with drift	851.7776
ARIMA(1,1,0)(0,0,1)[12] with drift	839.4082
ARIMA(0,1,0)(1,0,0)[12] with drift	851.3594
ARIMA(2,1,0)(1,0,0)[12] with drift	841.1989
ARIMA(1,1,0)(1,0,0)[12]	835.3515
ARIMA(1,1,0)	848.3919
ARIMA(1,1,0)(0,0,1)[12]	835.7279
ARIMA(0,1,0)(1,0,0)[12]	847.4882
ARIMA(2,1,0)(1,0,0)[12]	837.4102
ARIMA(1,1,1)(1,0,0)[12]	829.5711
ARIMA(1,1,1)	841.5117
ARIMA(1,1,1)(0,0,1)[12]	832.594
ARIMA(0,1,1)(1,0,0)[12]	**825.5133**
ARIMA(0,1,1)	837.6644
ARIMA(0,1,1)(0,0,1)[12]	828.5169
ARIMA(0,1,2)(1,0,0)[12]	829.5746
ARIMA(1,1,2)(1,0,0)[12]	832.7979

Best model: ARIMA(0,1,1)(1,0,0)[12]

The road accidents data exhibits seasonality, hence extending the ARIMA model to SARIMA. The SARIMA (0,1,1)(1,0,0)[12] developed in this study was used to make

forecasts.
From the above table 1 it was seen that AIC for SARIMA (0,1,1)(1,0,0)[12] was 825.5133. The AIC was cued for all the models but they favored SARIMA (0,1,1)(1,0,0)[12], model which had the least AIC. From the discussion above, it was clear that SARIMA (0,1,1)(1,0,0)[12] model was the best model for forecasting the number of road accidents which will occur in Kenya for the next 24 months. SARIMA (0,1,1)(1,0,0)[12] is the best model for forecasting for the number of people involved in road accidents cases in Kenya.
The model, therefore, is given as:
$\phi_p(B)\Phi_p(B^s)^{dD}{}_S X_t = \theta_q(B)\Theta_Q(B^S)Z_t$
$\phi_0(B)\Phi_1(B^{12})^{10}{}_{12}X_t = \theta_0(B)\Theta_0(B^{12})Z_t$ The SARIMA$(0,1,1)(1,0,0)_{12}$ model was estimated using the Maximum Likelihood Estimation (MLE) technique. This method seeks to determine the model parameters by maximizing the likelihood function under the assumption of normally distributed residuals. The estimated SARIMA$(0,1,1)(1,0,0)_{12}$ model is given by:

$$(1 - \Phi_1 B^{12})(1 - B)X_t = (1 + \theta_1 B)\epsilon_t \tag{2}$$

where:

- B is the backshift operator,
- X_t is the observed time series,
- Φ_1 is the seasonal autoregressive coefficient at lag 12,
- θ_1 is the non-seasonal moving average coefficient,
- ϵ_t is the white noise error term.

The estimated model parameters using the modification of an existing model are as follows

- Non-seasonal MA(1): $\hat{\theta}_1 = -0.45$
- Seasonal AR(12): $\hat{\Phi}_1 = 0.60$
- Residual variance: $\hat{\sigma}^2 = 12.34$

$$(1 - 0.6_1 B^{12})(1 - B)X_t = (1 - 0.45_1 B)12.34_t \tag{3}$$

These estimates indicate a moderate moving average effect and strong seasonal autoregressive behavior. Diagnostic checks of residuals confirmed the adequacy of the model, supporting its use for forecasting.

3.3 Model Diagnostic Checking

The test for stationarity of the number of road accident cases was conducted using ADF test in the R software. The null hypothesis was road accidents was non-stationary and the alternative hypothesis was stationary

H_0: Road accidents was non-stationary

H_1: Road accidents was stationary

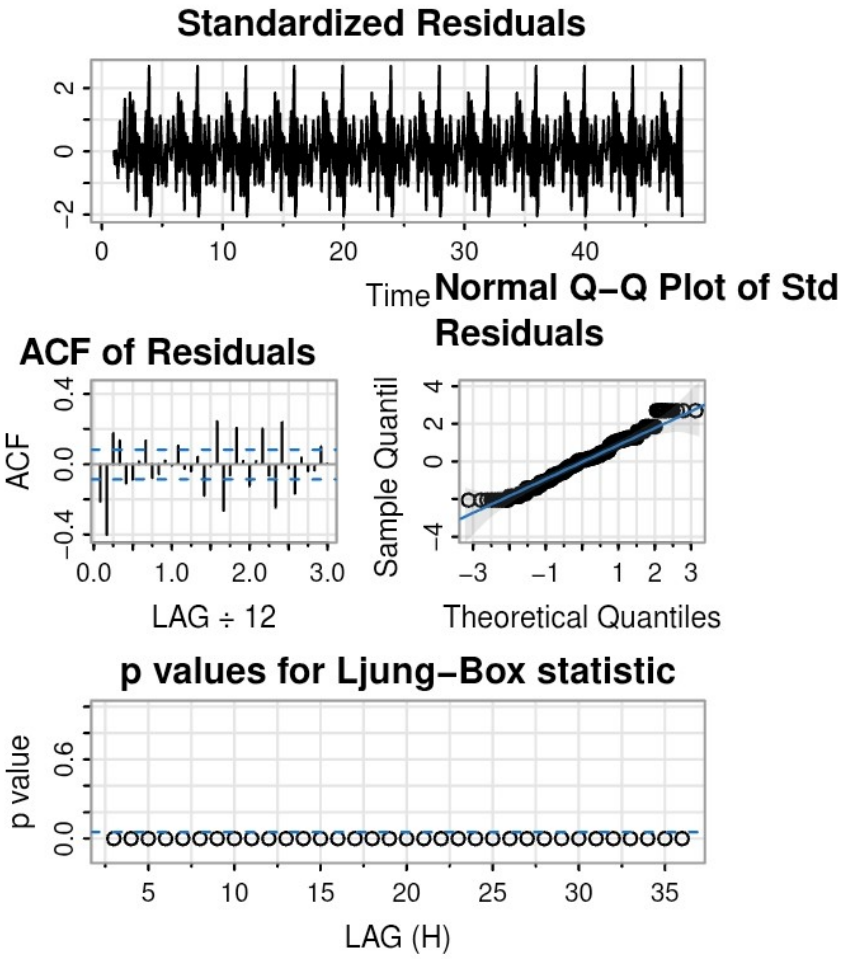

Figure 3: Q-Q Plot and Residual Diagnostics

To validate the model assumptions, a Q-Q plot of the residuals was examined. The plot indicated that the residuals were approximately normally distributed, as the points closely followed the reference line. This supports the appropriateness of the SARIMA model and suggests that the model's inferences are statistically reliable. Table 2 Shapiro-Wilk test is used to test for normality at 5% level of significance for the number of road accidents occurring in Kenya for the data from NTSA. The p-value is greater than alpha level($0.2869 > 0.05$), the road accident data is normally distributed.

Table 2: Test for Normality

Shapiro Wilk statistic	p-value
0.97612	0.2869

newpage

3.4 Forecasted road accidents

The table 3 shows the point forecast in the next 24 months. It is clear that the of road accidents is increasing at an alarming rate.

Table 3: Forecasted Road Accidents

Months	Forecast
Jan 2024	1949.882
Feb 2024	1828.136
Mar 2024	1975.773
Apr 2024	1993.402
May 2024	1895.344
June 2024	2012.683
July 2024	2016.539
Aug 2024	2228.676
Sept 2024	1975.773
Oct 2024	1996.707
Nov 2024	1761.479
Dec 2024	2445.128
Jan 2025	2035.204
Feb 2025	1968.136
Mar 2025	2049.467
Apr 2025	2059.178
May 2025	2005.160
June 2025	2069.800
July 2025	2071.924
Aug 2025	2221.841
Sept 2025	2049.467
Oct 2025	2060.999
Nov 2025	1931.416
Dec 2025	2308.028

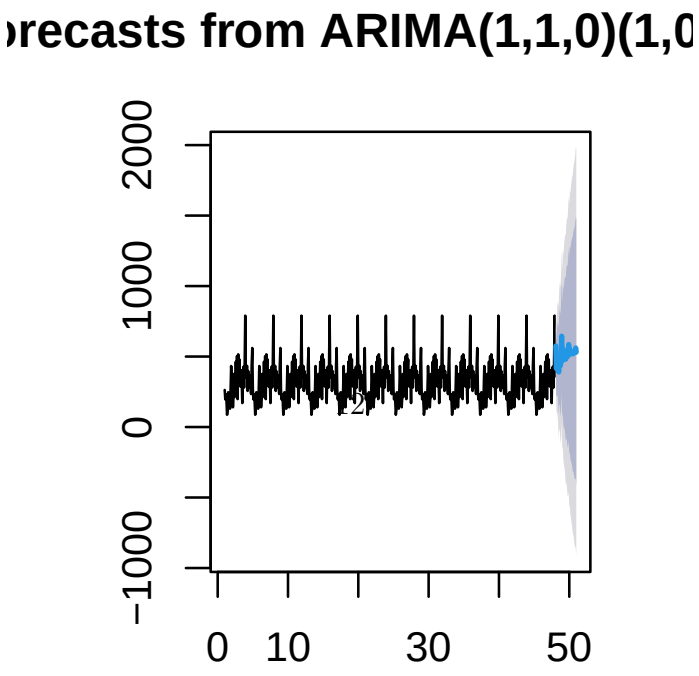

Figure 4 shows forecasted values for the period January 2024 to December 2025 indicate an overall increasing trend in the number of road accidents. While there are monthly fluctuations, the general upward movement suggests a gradual rise in accident frequency over time. This trend emphasizes the need for proactive policy interventions, improved road safety measures, and continuous monitoring to mitigate the projected increase in road accidents.

4 Conclusion and Recommendations

4.1 Conclusion

The main objective of this study was to assess the trend and forecast the number of road accidents in Kenya which was achieved in chapter 3 as the number of road accidents was forecasted in the next 24 months. The best model found was SARIMA (0,1,1)(1,0,0)[12].From the findings, it is clear that the number of road accidents in Kenya is increasing at an alarming rate and has raised major concerns. The SARIMA (0,1,1)(1,0,0)[12] model provides accurate short-term forecasts for road accidents in Kenya. The study confirms the increasing trend of accidents, underlining the urgency of responsive policy measures. In line with the findings of this study, the best model is SARIMA (0,1,1)(1,0,0)[12]. It is therefore recommended for researchers to use the same model. However the following precautionary measures should be taken into consideration to prevent the increasing forecast values of these models: Enforcement of traffic safety campaigns, proper maintenance of the roads, strict adherence to road traffic rules. The model should not be used to forecast a long time ahead because long periods could lead to arbitrary large forecast values. Finally, it is also recommended that further study should be done to look for more appropriate models that can take care of the drastic government interventions.

4.2 Recommendations

- Use SARIMA for regular short-term forecasts
- Implement targeted awareness campaigns during high-risk months.
- Invest in infrastructure and driver training.
- Avoid long-term extrapolations due to forecast uncertainty.

References

[1] Time-series projecting road traffic fatalities in australia: Insights for targeted safety interventions. *Injury*, 56(3):112166, 2025. ISSN 0020-1383. doi: https://doi.org/10.1016/j.injury.2025.112166.

[2] Edmund F Agyemang, Joseph A Mensah, Eric Ocran, Enock Opoku, and Ezekiel NN Nortey. Time series based road traffic accidents forecasting via sarima and facebook prophet model with potential changepoints. *Heliyon*, 9(12), 2023.

[3] Abdulgafoor M Bachani, Pranali Koradia, Hadley K Herbert, Stephen Mogere, Daniel Akungah, Jackim Nyamari, Eric Osoro, William Maina, and Kent A Stevens. Road traffic injuries in kenya: the health burden and risk factors in two districts. *Traffic injury prevention*, 13(sup1):24–30, 2012.

[4] Tong Cheng. Research on the road traffic accident prediction based on sarima-lstm model. In Xiantao Xiao and Jia Yao, editors, *Eighth International Conference on Traffic Engineering and Transportation System (ICTETS 2024)*, volume 13421, page 134213Z. International Society for Optics and Photonics, SPIE, 2024. doi: 10.1117/12.3054553. URL `https://doi.org/10.1117/12.3054553`.

[5] David J Houston and Lilliard E Richardson Jr. Traffic safety and the switch to a primary seat belt law: the california experience. *Accident Analysis Prevention*, 34(6):743–751, 2002.

[6] Humanitarian Data Exchange. Kenya accidents database.xlsx, 2023. `https://data.humdata.org/dataset/8288bf4a-1ec3-454d-a201-3b7e4c623063/resource/bcd9ef77-cf9f-4dc0-b3f8-75ad238fb433/download/kenya-accidents-database.xlsx`.

[7] M Manikandan, Vishnu Prasad, Amit Kumar Mishra, Rajesh Kumar Konduru, and A Newtonraj. Forecasting road traffic accident deaths in india using seasonal autoregressive integrated moving average model. *International Journal of Community Medicine and Public Health*, 5(9):3962, 2018.

[8] Nadia K Naqvi, Mohammed Quddus, and Marcus Enoch. Modelling the effects of fuel price changes on road traffic collisions in the european union using panel data. *Accident Analysis & Prevention*, 191:107196, 2023.

[9] Mohammed A Quddus, Robert B Noland, and Washington Y Ochieng. A high accuracy fuzzy logic based map matching algorithm for road transport. *Journal of Intelligent Transportation Systems*, 10(3):103–115, 2006.

[10] Muhammad Babar Ali Rabbani, Muhammad Ali Musarat, Wesam Salah Alaloul, Muhammad Shoaib Rabbani, Ahsen Maqsoom, Saba Ayub, Hamna Bukhari, and Muhammad Altaf. A comparison between seasonal autoregressive integrated moving average (sarima) and exponential smoothing (es) based on time series model for forecasting road accidents. *Arabian Journal for Science and Engineering*, 46(11):11113–11138, 2021.

[11] Sodano Sheiso and Desta. Time series modelling of road traffic accidents in west arsi, ethiopia, August 20 2024. Available at SSRN.

[12] Ho Jen Sim, Choo Wei Chong, Khairil Anwar Abu Kassim, Ching Siew Mooi, and Zhang Yuruixian. Forecasting road traffic fatalities in malaysia using seasonal autoregressive integrated moving average (sarima) model. *Pertanika Journal of Science & Technology*, 30(2), 2022.