International Journal of Research (IJR)

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

Modeling and Forecasting Road Traffic Accident of Addis Ababa Using Self Excited Threshold Autoregressive Model

Hirko Beyene

Department of Statistics, Bule Hora University, Ethiopia

Email addresses: hirkisa06@gmail.com

Abstract

Road traffic accidents are one of the leading causes of injuries and death in both developed and developing countries. According to World Health Organization(WHO), 1.35 million people die each year as a result of road traffic accidents globally. Ethiopia is one of the developing countries and at least 114 people die for every 10,000 vehicle accidents annually. Addis Ababa is a capital city of the federal government of Ethiopia and the road traffic accidents (RTA) occurring from time to time. Assessing the trend of road traffic and forecasting were used to make the appropriate planning for control and prevention of the accidents in upcoming years. Therefore, this study focuses on statistical analysis of road traffic accident using Self-Excited Threshold Autoregressive (SETAR) time series model. The two-regime SETAR model was adopted to accommodate non-linearity. The results show that nonlinear SETAR (2,8,8) model forecast road traffic accident. The out sample forecasted value indicates that, road traffic accident has an increasing trend over the forecasted period from January 2017 to December 2018.

Keywords: Forecasting, Aggregate, white noise, stationary, nonlinear, SETAR

Introduction

Road traffic accident is a major public health problem worldwide. It includes collisions between vehicles and animals, vehicles and pedestrians, vehicles and fixed objects or vehicles and vehicles (Zewude et al., 2016).According to WHO (2018) report, there are 1.35 million deaths per year and road traffic accident is the eighth leading cause of death for all age groups globally.

WHO (2018) investigated the risk of dying in a road traffic accident by continent and Africa is the leading with the chance of 26.6 followed by south east Asia (20.7).Even though the numbers show the large prevalence rate in road traffic accident in developing countries specifically in Africa (including Ethiopia), the issue is still under reported and neglected to be studied and interventions are needed urgently (Samuel et al., 2012).

Abegaz et al. (2019) reveals that, road accident related injuries and fatalities are exceptionally high in Ethiopia and there is paucity of evidence regarding of the accident. Further, the available estimates based on official reports are likely to underestimate the extent of the problem. Addis Ababa is a capital city of the federal government of Ethiopia and the head office of African union is also found in this city and more than 100 diplomatic embassies exist. The city has a total population of 3, 384,569 according to CSA (2016) projected census report for Addis Ababa..

Abdi et al.(2017) reveals that, road traffic accident in Addis Ababa resulting in thousands of physical injuries and costing the economy in millions of dollars. Moreover, the finding of Bekele (2019) emphasis that, road traffic accident continue to be a significant for morbidity and mortality problem in Addis Ababa and requiring urgent attention.

Unless the trend is detained, the social and economic problem of road traffic accident will become more and more serious as the number of vehicle increases. And, access to the information about road traffic accident in a given context is significant to generate evidence to contribute to the prevention and control of context-specific accidents (Yazdani Cherati et al., 2012).Time series models are very useful to enhance one's understanding on traffic accident trends (Van den Bossche et al., 2004).

Nonlinear threshold model was introduced by Tong (1978) and Self-Exciting Threshold Autoregressive (SETAR) model was a special case of the Threshold Autoregressive model which accommodates structural changes in regimes of the data. Moreover, Clements et al. (2004) reviewed, the current state of this model ranging from estimation, evaluation and selection of forecasting models.

Time series analysis related to the road traffic accident has a very important place in revealing the future trends and decision making process (Ghedira et al., 2018).This

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

future trend can help in identifying the feature of road accidents whether it tends to increase or decrease so that preventive measures can be taken. Therefore, the objective of this research is to modeling and forecasting road traffic accident of Addis Ababa using SETAR time series model.

Methodology

Method of data collection

Data for this study were obtained from the Addis Ababa traffic police commission. The site was chosen due to availability of relatively long series of road traffic accident data. Even though, the data were recorded in total hourly (24 hrs) per in each year .The hourly data of road traffic accident were aggregated to monthly for statistical analysis. An aggregation process consists of deriving a low frequency representation of the process from a high frequency formulation; this derivation can be exerted through time. Aggregation across time, also called temporal aggregation, refers to the process by which a low frequency time series is derived from a high frequency time series (Nikolopoulos et al., 2011).

After the data aggregated from total hourly to January 2004 to December 2018, the model forecast performance was conducted by splitting a given data set into in-sample period which were used for initial parameter estimation and model selection; and out–sample period which were used to evaluate forecast performance. Yet, there are no broadly accepted guidelines as to how to select the sample split (Hansen et al., 2012).Instead, researchers adopt a variety of practical approaches (Hansen et al., 2012). In this study, the in-sample period was runs from January 2004 to December 2016(156 observation). And, this was used to model selection and estimation. Whereas, the out-of-sample period runs from January 2017 to December 2018 (24 observation) for evaluation of forecasting purpose.

Method of data analysis

SETAR model

Self-Excited Threshold Autoregressive (SETAR) model is a class of the Threshold Autoregressive (TAR) model proposed by Tong (1978) and further studied in Tong and Lim(1980). The SETAR model is a set of different linear AR models, changing according to the value of the threshold variable which is the lagged values of the series. The process was linear in each regime, but the movement from one regime to the other makes the entire process nonlinear. The two regime SETAR model order (2; $P_{(1)}$, $P_{(2)}$) was given as:

$$y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{P(1)} \phi_i^{(1)} y_{t-i} + \varepsilon_t^{(1)} \quad if \; y_{t-d} \le \tau \\ \\ \phi_0^{(2)} + \sum_{i=1}^{P(2)} \phi_i^{(2)} y_{t-i} + \varepsilon_t^{(2)} \quad \text{if} \; y_{t-d} > \tau \end{cases}$$

………………..……………………..[1]

Where $\phi_i^{(1)}$ and $\phi_i^{(2)}$ are the coefficient in lower and higher regime respectively which needs to be estimated; τ is the threshold value; $P_{(1)}$, and $P_{(2)}$ are the order of the linear AR model in low and high regime respectively. y_{t-d} is the threshold variable that governs the transition between the two regimes, d is the delayed parameter which is a positive integer (d≤ p); $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$ are a sequence of independently and identically distributed random variables with zero mean and constant variance.

Test of non-linearity

This was done using graphical method, Keenan test, Tsay test and likelihood ratio test for threshold nonlinearity.

Graphical method

This method basically gives a clue. If the regression lines were all straight in lagged scatter plots and the density of the plots decrease from the center suggests that the underlying process could be linear. If on the other hand, there were curved regression lines and there exists a hole in the center, then it was suggested that the data could fitted with nonlinear time series model (Cryer and Chan, 2008).

Keenan's One-Degree Test for Nonlinearity

For an observable time series yt, the SETAR model was only be applied if the series under consideration is found to be nonlinear or irregular in nature under the hypothesis:

H0: linearity exists

H1: nonlinearity exists

Beyond testing for nonlinearity, Keenan test suggests the working order (p) of the AR process. This was determined by minimizing the AIC through the AR function. However, where the working order (p) of the AR was known, it can be added in the Keenan test function through the order argument. The partial autocorrelation graph can also suggest the working order (p) of the AR process.

According to Cryer and Chan (2008), Keenan's test was motivated by approximating a nonlinear stationary time series by a second-order Volterra expansion. This can be represented in equation

$$y_t = \mu + \sum_{u=-\infty}^{\infty} \theta_u \, \varepsilon_{t-u} + \sum_{v=-\infty}^{\infty} \sum_{u=-\infty}^{\infty} \theta_{uv} \, \varepsilon_{t-u} \, \varepsilon_{t-v} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots[2]$$

Where, μ is the mean level of nonlinear observation yt , with the error terms ε_{t-u} and ε_{t-v}. ε_t is a sequence of independent and identically distributed with zero-mean random variable. The process yt is linear if the double sum of the right-hand side of (2) does not exist (i.e. $\sum_{v=-\infty}^{\infty} \sum_{u=-\infty}^{\infty} \theta_{uv} \varepsilon_{t-u} \, \varepsilon_{t-v} = 0$) Thus, the researcher was test linearity of the time series by testing whether or not the double sum of (2) was zero. The Keenan's test is equivalent to testing if η=0 in the multiple regression model

$$y_t = \theta_0 + \phi_1 y_{t-1} + \ldots + \phi_m y_{t-m} + \eta \, \hat{y}_t^{\,2} + \varepsilon_t \ldots\ldots\ldots\ldots[3]$$

The Keenan's test statistic for the null hypothesis of linearity (H0: η =0) was given as;

$$\hat{F} = \frac{\eta^2 (n - 2m - 2)}{RSS - \eta^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots[4]$$

Where, m is lag order of the linear autoregressive process, n is same size considered, RSS: the residual sum of squares from the AR(m) process. The null hypothesis of linearity was rejected if the p-value associated with $\hat{F}$ was small (p– value< α).

Tsay's F -test

Tsay (1989) introduced the Tsay test for detecting nonlinearity in an observable time series after improving on the power of the Keenan (1985) test in 1986. Tsay's (Tsay, 1986) linearity test was based on recursive auto regression and destructive term estimators .firstly, the recursive auto regressions were established starting from b observation value in return for the p and the relevant d values with AR level, h is max(1, p+1-d), $\hat{e}_t$ is the estimated residual and then the model was established between the $\hat{e}_t$ values and $(1, y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ .Then, the following test was obtained among the inclusions of the model formed with $\hat{e}_t$

$$\hat{F}(p,d) = \frac{\left(\sum \hat{e}_t - \sum \hat{e}_t^{\,2}\right) / (p+1)}{\sum \hat{e}_t^{\,2} / (n-d-b-p-h)} \text{..........}[5]$$

In the F test of Tsay, the hypothesis was tested as;

H_0: linear AR(p) model vs H_1: non-linear threshold model

Threshold Nonlinearity (Likelihood Ratio) Test

This helps to handle the weakness of Keenan test in detecting threshold nonlinearity (Chan, 1991; and Tong, 1990). Therefore, it becomes important to consider a likelihood ratio test with the threshold model as the specific alternative. The hypothesis of threshold nonlinearity test was given as follows;

H_0: Model is an autoregressive process AR (p) vs

H_1: Model is two-regime TAR model of order p and with constant noise variance

The likelihood ratio test statistic was given by:

$$T_n = (n-p)\log\left\{\frac{\hat{\sigma}^2(H_0)}{\hat{\sigma}^2(H_1)}\right\} \text{...............}[6]$$

Where n-p is the effective sample size, $\hat{\sigma}^2(H_0)$ is the maximum likelihood estimator of the noise variance from the linear AR (p) fit and $\hat{\sigma}^2(H_1)$ from the TAR fit with threshold searched over some finite interval.

Choosing the Threshold Variable and Delay Parameter

Tsay (1989) describe a method of selecting the threshold variable. In context of SETAR model for a given time series y the threshold variable was taken as its own lag value yt−d for some positive integer d called delay parameter providing that d ε{1, 2,. . . , d∗}, where d∗ was the upper bound. The estimated optimal value of d was chosen in such a way that it provides the maximum F-values. Tsay suggests to select an estimate of the delay parameter, such that

$$d = \arg\ \max_{d \le p}\ \hat{F}(p, d_p) \text{.........}[7]$$

Where, $\hat{F}(p,d_p)$ was given in equation (5) the F-statistic value, the estimate of d depends on p. The delay value that gives the highest test of F value for the relevant p-value was selected from the threshold variables and it was suspected to be the delayed parameter for the SETAR model.

Model Selection

The SETAR model has two different AR processes in the two regimes defined by the threshold variable y_{t-d} and threshold value (τ). One important issue in fitting the SETAR model was to select the best subset model. In other words, we need an efficient way to identify values of d and subsets of $\phi_i^{(k)}$,k=1,2 and k=0,…, $P_{(i)}$ that were important , given the maximum AR orders of $P_{(1)}$ and $P_{(2)}$ for the two regimes. So et al.(2003) reveals that identification problem can be highly complicated because it involves a very large number of possible models and there were $d_0 \times 2^{P_1+P_2+2}$ models to consider, where d_0 the maximum possible delay parameter. The grid search method was used to find the potential threshold in the series by minimizing the residual sum of square of as follows:

$$\hat{\theta} = \arg\min_{\theta} Rss(\theta) \quad \text{……………………………[8]}$$

Where θ is threshold parameter. The model which have the smallest residual sum of squares was the most consistent estimate of the delay parameter. Therefore, a threshold value corresponding to the smallest sum square of residuals was efficient. The orders of SETAR models were commonly identified by considering the Akaike information criterion (AIC). For each possible delay parameter, Tong and Lim (1980) use the AIC to estimate threshold value and find suitable autoregressive orders in both regimes of the threshold model.

$$AIC\,(p_1, p_2) = n_1 \ln(\hat{\sigma}_1^{\,2}) + n_2 \ln(\hat{\sigma}_2^{\,2}) + 2\,(p_1+1) + 2(p_2+1) \quad \text{……………………………………… [9]}$$

Where, $n_j\ , j=1,2$ is the number of observations in the j[th] regimes and

$\hat{\sigma}^2{}_j\ , j=1,2$ is the variance of the residuals in the j[th] regimes p_1 and p_2 are the selected lags order in regime 1 and 2 respectively for which the information criterion is minimized.

Parametric Estimation

After the desired model was selected, the next step is to estimate the parameters of the selected model. The parameters were estimated using a sequential conditional least square method. According to Franses and van Dijk (2000), by using this method the resulting estimates were equivalent to maximum likelihood estimates under the additional assumption that the residuals are normally distributed.

Model Diagnostics

After carefully selecting tentative models to be used for forecasting, the residuals of the models were checked. This step is paramount to making any meaningful inferences with the models. So, plot of residual versus time, test of normality and Ljung-Box test of serial correlation were employed in this study.

Time Plot of the Residuals

Time plot of the standardized residuals should not show any structure. It must indicate no trend in the residuals and no changing variance across time.

Normality of the residuals

To investigate whether or not the residuals of the fitted model were normally distributed, the Jarque-Bera test and Q-Q plot were applied. Q-Q plot was a normal probability plot of a plot based on estimated quantiles. The normal Q-Q plots provide a quick way to visually inspect to what extent the pattern of data follows a normal distribution.

Test for Serial Correlation

Ljung and Box (1978), described this test as a diagnostic tool used to check for the presence or absence of serial correlations in the residuals of a fitted model. The test procedure was given as follows; The hypothesis to be tested is;

H_0 : Residuals are uncorrelated up to order k

H_1: Residuals are correlated up to order k

The test statistic is

$$Q_k = n(n+2)\sum_{d=1}^{k}\frac{\hat{\rho}_d^{\,2}}{n-d} \text{........................[10]}$$

Where, $\hat{\rho}_d^{\,2}$ represents residual autocorrelations of the series at lag k, k is the number lags being tested, n is the number of residuals. The model was considered adequate when the p-value associated with Qk is large; otherwise the whole estimation process has to be repeated again in order to get the most adequate model.

Forecasting From SETAR Model

The important aim of considering nonlinear type of model such as SETAR as compare to the linear counterpart was to adequately describe the dynamic behavior of the observable series under consideration and also to produce adequate forecast values .The optimal one step-ahead forecast from the origin is given by:

$$\hat{y}_{t+1|t} = E\left[y_{t+1}\middle|\Omega_t\right] = E\left[F(x_{t+1};\phi)\middle|\Omega_t\right] \text{............[11]}$$

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

Where $\hat{y}_{t+1}$ is the forecast value for the time (t+1), and Ω_t is the history of the time series up to and including the observation at time t. $F(x_t;\phi)$ is the nonlinear function that represent the SETAR model. The next optimal step-ahead forecast is given by:

$$\hat{y}_{t+2|t} = E\left[y_{t+2}|\Omega_t\right] = E\left[F(x_{t+2};\phi)|\,\Omega_t\right] \text{.........[12]}$$

In general, the linear conditional expectation operator E cannot be interchanged with the nonlinear operator F, that is

$$E[F(.)] \neq F(E[.]) \text{.............................[13]}$$

Put differently, the expected value of a nonlinear function is not equal to the function evaluated at the expected value of its argument. Hence,

$$E\left[F(y_{t+1};\phi)|\,\Omega_t\right] \neq F\left(E\left[y_{t+1}|\,\Omega_t\right];\phi\right) = F\left(\hat{y}_{t+h|t};\phi\right)$$

....[14]

The optimal h-step-ahead forecast can be obtained as

$$\hat{y}_{t+h|t} = E\left[y_{t+h}|\Omega_t\right] = F(x_{t+h-1};\phi)$$

...............[15]

Results and discussion

Formal test of nonlinearity

In modeling road traffic accident with the SETAR model, the data should first satisfy the condition of nonlinearity. So, the nonlinearity was checked in this series by specifying the order of the linear AR(p) model. The order of the linear AR(p) was chosen as the AR(p) model based on the maximum lag order with the least value of AIC. The summary of the linearity tests were given in the following Table 1

Table 1: Linearity test for first differenced log road traffic accident data

Test	Test statistic	P- value	Order	Decision
Keenan 1-degree	10.60241	0.00144	12	Reject Linearity
Tsay	3.296	0.0000	12	No threshold nonlinearity rejected

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

Likelihood Ratio Test	3.42983	0.00025	12	The model is SETAR with two regimes

From the above Table 1, Keenan and Tsay tests suggest that the working order, p could be 12 using Akaki Information Criterion. The Keenan test statistic (10.60241) was significant with p-value (0.00144), and Tsay test statistic (3.296) was significant with p-value (0.000). As a result, the null hypothesis was rejected with conclusion that first differenced log road traffic accident data follows a nonlinear process. In addition, the p-value of likelihood ratio test is 0.0025. So, we reject the null hypothesis that the time series follows some AR process and concluded that data follows a two-regime SETAR model of order p with constant noise variance.

Selection of the lag order and nonlinearity test

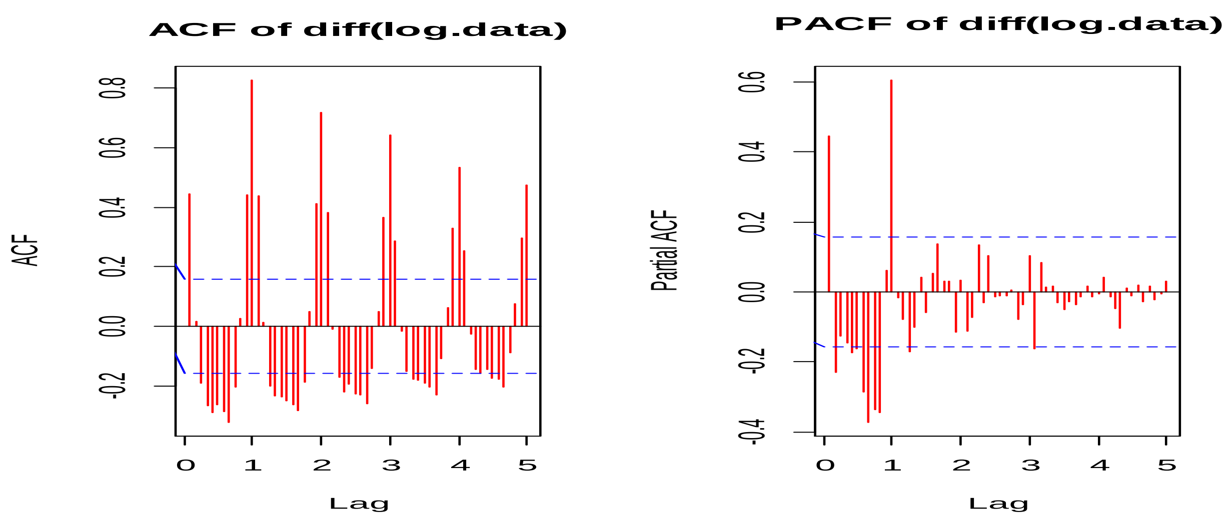

Figure 1: ACF and PACF of first order differenced logarithm of monthly road traffic accident data of Addis Ababa

The above Figure 1 reveals that, the autocorrelation function and partial autocorrelation function were the indication of linear dependence existing between the lags. The structure of the lags indicate the presence of a strong and persistent cycle in

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

the data. From the above Figure 1, there was significant large spike at seasonal lags (12, 24, 36, etc.) revealing that, there was seasonality in the first order differenced logarithm of monthly road traffic accident data of Addis Ababa. The researcher notice a non-seasonal lag cut at lag1 and a seasonal lag cut at lag12 suggesting an AR parameter of order 12 (p=12).This also confirmed by Keenan and Tsay tests from the Table 2. This suggests that, an AR model with each individual lag order from 1 to 12 should be tested using Tsay's F test for threshold nonlinearity.

Table 2: Nonlinearity test (p=12)

Delay (d)	**F-value**	**P-value**
1	**3.6002**	**0.0001***
2	**3.3464**	**0.0003***
3	**3.3005**	**0.0003***
4	1.3898	0.1769
5	1.1934	0.2947
6	1.6188	0.0917
7	**2.6682**	**0.0029***
8	**2.5099**	**0.0050***
9	**2.4310**	**0.0066***

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

10	1.5563	0.1103
11	**2.4720**	**0.00l57***
12	**1.8786**	**0.0411***

The above Table 2 shows that, all tests suggest nonlinearity for d=1,2,3,7,8,9,11 and 12 at the 5% level .However, The p-values of delayed 4, 5,6 and 10 were greater than 0.05. That means, we do not reject the null hypothesis of no threshold nonlinearity for all chosen delay parameter. Then, the researcher tried lower the lag order (p) to 8 by avoiding a non-significant delayed parameters as adopted in (Zivot, E. and Wang, J., 2005). Subsequently, nonlinearity for lag order (p=8) was tested. The summary of the result was given in the following Table 3.

Table 3: Nonlinearity test when p=8

Delay(d)	1	2	3	4	5	6	7	8
F-value	5.2409	8.5895	4.8626	9.4523	10.7204	5.3914	4.1018	4.3857
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.0001	0.0001

The above Table 3 show that, the tests for all delays (d) have a p-values less than 0.05. The null hypothesis of no threshold nonlinearity was actually rejected for all delayed parameters. The optional argument p specifies the AR order to use in the arranged auto regression, and the optional argument d was used to select the delay parameters from

1 to 8. The output was give the F statistics and their corresponding p-values for all chosen values of delayed parameter (d), and shows that the evidence for threshold nonlinearity was strong with the AR(8) specification **.**

Selection of the delayed parameter

For a given AR order p, Tsay suggests to select an estimate of the delay parameter, such that $d = \arg\max_{d \le p} \hat{F}(p, d_p)$.Where, $\hat{F}(p, d_p)$ was the F-statistic value, the estimate of d depends on p. From above Table 3, when d=1, F=5.2409, when d=2, F=8.5895, when d=3, F=4.8626, when d=4, F=9.4523, when d=5, F=10.7204, when d=6, F=5.3914 , when d=7, F=4.1018 , when d=8, F= 4.3857 .The largest test statistic value occurred at d = 5. Consequently, 5 is suspected to be the delayed parameter for the SETAR model.

Model Selection

After confirming that the data were threshold nonlinear with 5 delayed parameter, the specific SETAR model that fit the data was identified. This was done by determining the autoregressive lag order (p) in each regime and the threshold variable y_{t-d} .Where, d represent the delayed parameter. The researcher choose the model with lag order (p) for both regimes and threshold variable with the minimal AIC by performing a grid search on all possible combinations of SETAR models. The selected model using grid search from all possible models combinations were illustrated in Table 4.

Table 4: Grid search for SETAR model using d=5

Grid search for the model using (p=8)						
Threshold delayed (d)	Rank	Order of lower regime	order of upper regime	Threshold Value	AIC	Serial correlation
d=5	**1**	**8**	**8**	**-0.2564485**	**-396.9006**	**No**
	2	8	8	-0.2674560	-395.6086	No
	3	8	8	-0.2244672	-394.1635	No

	4	5	8	-0.2718599	-393.6719	Yes

Thus, from above Table 4, the grid search using delay (d=5) with order of lower regime 8 and order of upper regime 8 with -0.2564485 threshold value have the smallest AIC value. Accordingly, SETAR (2, 8, 8) model with no serial correlation is found to be the selected model that fits road traffic accident data.

Model Estimation

Table 5: Maximum likelihood estimates of SETAR (2, 8, 8) model

	Low Regime				High Regime			
Coefficient	Estimate	Std Error	t- value	p-value	Estimate	Std Error	t- value	p-value
Constant	0.220	0.207	1.059	0.291	0.136	0.033	4.141	0.000
ϕ_1	-0.964	0.206	-4.682	0.000	0.387	0.086	4.490	0.000

ϕ_2	-0.887	0.143	-6.189	0.000	0.169	0.098	1.730	0.085
ϕ_3	-1.122	0.175	-6.399	0.000	-0.211	0.089	-2.364	0.019
ϕ_4	-0.771	0.159	-4.845	0.000	-0.254	0.077	-3.276	0.001
ϕ_5	-0.692	0.196	-3.522	0.000	-0.208	0.071	-2.892	0.004
ϕ_6	-0.806	0.376	-2.143	0.033	-0.310	0.075	-4.107	0.000
ϕ_7	-0.986	0.378	-2.607	0.010	-0.135	0.0621	-2.189	0.030
ϕ_8	-1.351	0.277	-4.8685	0.000	-0.391	0.0619	-6.317	0.000
	Threshold value = -0.2564							
Proportion	25.17%				74.83%			

As indicated in the above Table 5, the numbers of data falling in lower and upper regimes are 25.17% and 74.83% respectively. Hence, the final SETAR (2, 8, 8) model using the estimated value of above Table 5 was written as follows;

$$y_t = \begin{cases} 0.22 - 0.94y_{t-1} - 0.89y_{t-2} - 1.12y_{t-3} - 0.77y_{t-4} - 0.69y_{t-5} - 0.81y_{t-6} - 0.97y_{t-7} - 1.35y_{t-8}, \; y_{t-5} \le -0.2564 \\ 0.14 + 0.39y_{t-1} + 0.10y_{t-2} + 0.09y_{t-3} + 0.08y_{t-4} + 0.07y_{t-5} + 0.81y_{t-6} + 0.06y_{t-7} + 0.062y_{t-8}, \; y_{t-5} > -0.2564 \end{cases}$$

The below Figure 2 depicts nature of the regimes. And, it reveals which data value falls in which regime in the first differenced logarithm of monthly road traffic accident data. Data falling in the lower regime was drawn as black line while high regime was drawn by red line .The estimated threshold was -0.2564 as estimated in the above Table 5.This threshold value was not close to the minimum or the maximum observation . It was the break point of the data.

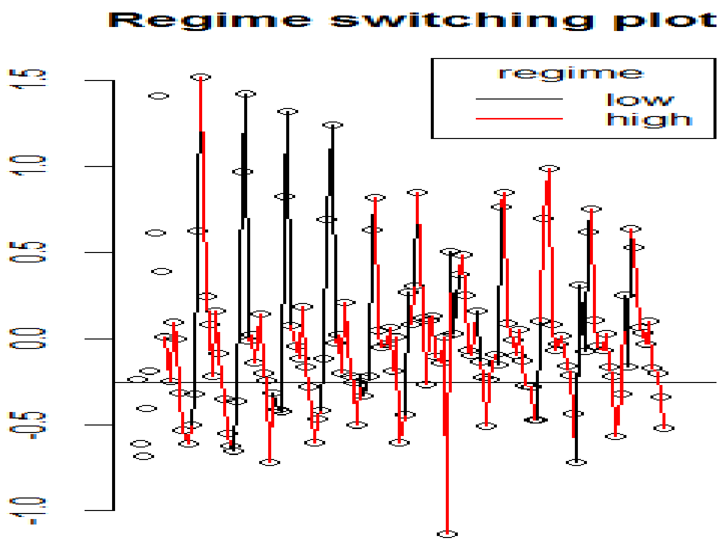

Figure 2: Data falls in the lower and upper regimes of a fitted SETAR (2,8,8) model

Evaluation of forecasting accuracy

Table 6: Evaluation of forecasting accuracy of SETAR (2,8,8)

Accuracy measure	Estimates
ME	0.0008
MSE	0.012
MAE	0.090
RMSE	0.113

Table 6 reveals that, mean error was 0.0008, mean square error was 0.012 and mean absolute error was 0.090 the average magnitude of the errors in a set of forecasts was 0.113. Thus, forecasting error was small.

Forecasting of SETAR (2, 8, 8) Model

Table 7: Actual and forecasted values of the series using SETAR (2, 8,8)

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

Month	Observed value	Forecasted value	Month	Observed value	Forecasted value
Jan 2017	777	578	**Jan 2018**	1575	808
Feb 2017	948	1003	**Feb 2018**	1630	1254
Mar 2017	1551	1347	**Mar 2018**	2146	1688
Apr 2017	2768	2489	**Apr 2018**	3236	2946
May 2017	3124	4703	**May 2018**	3347	3725
Jun 2017	3165	4203	**Jun 2018**	3196	3621
Jul 2017	3089	4285	**Jul 2018**	2973	3335
Aug 2017	3280	2704	**Aug 2018**	2819	2604
Sep 2017	3051	2456	**Sep 2018**	2607	2509
Oct 2017	2382	1928	**Oct 2018**	2142	1845
Nov 2017	1677	1214	**Nov 2018**	1605	1375
Dec 2017	1125	1000	**Dec 2018**	1085	1194

Model diagnosis

After carefully selecting tentative models to be used for forecasting, the researcher check the residuals of the models to ensure that, the model satisfy the assumptions.

Time plot of the SETAR (2, 8, 8) model residuals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

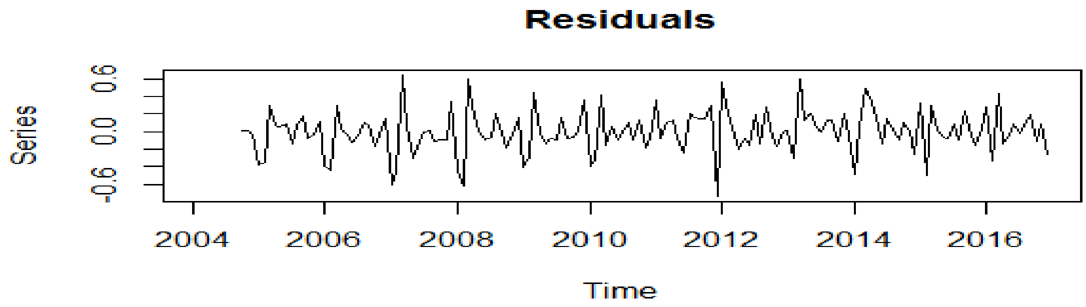

Figure 3: Time plot of the SETAR (2,8,8) model residuals.

As shown in above Figure 3, the standardized residuals plot shows no obvious pattern and looks like an identically and independently distributed of mean zero.

Test of Normality

To investigate whether or not the residuals of the fitted model are normally distributed, the Jarque-Bera test was applied. The test have a null hypothesis that the residual follows a normal distribution and therefore a rejection of the null hypothesis suggests that the residual does not follow a normal distribution.

Table 8: Jarque Bera test for SETAR (2,8,8) model residuals

	X-squared	Df	p-value
Jarque Bera Test	1.1787	2	0.5547

As shown in above Table 8, the p-value for the test was 0.5547 which was greater than 0.05. So, we do not reject the null hypothesis. This test was provide evidence of normality for the standardized residuals. In addition, Figure A5 (Appendix) show that the histogram and QQ plot of SETAR (2,8,8) model residuals. The histogram features provide strong indications of the proper distributional of the model for the data. Also, the QQ-normal plot seem to follow a straight line especially in the extreme values.

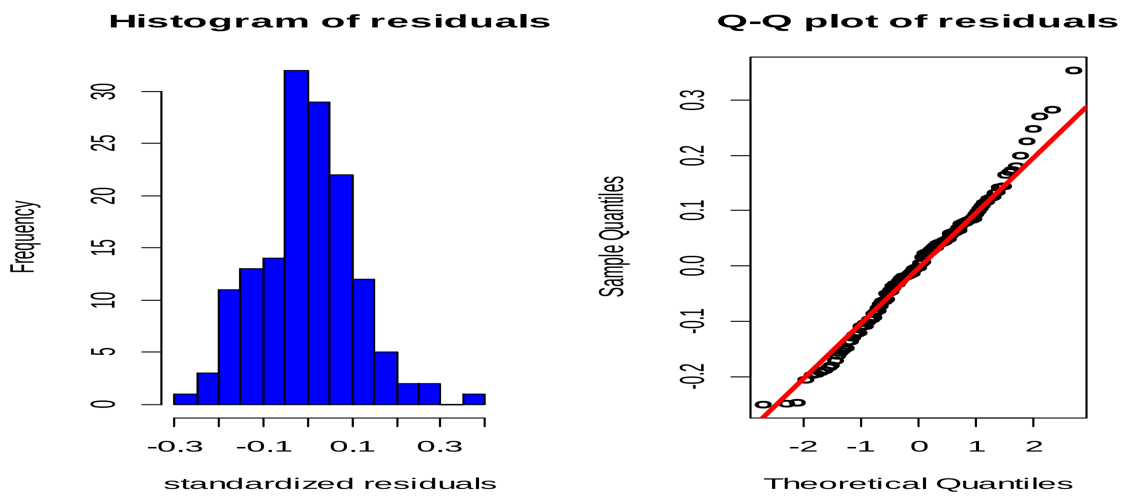

Figure 4: Histogram and Q-Q plot of standardized SETAR (2,8,8) residuals

Test for Serial Correlation

Ljung and Box (1978), described this test as a diagnostic tool used to check for the presence or absence of serial correlations in the residuals of a fitted model.

Table 9: Box-Pierce test for SETAR (2,8,8) model

Box-Ljung test	X-squared	Df	p-value
	0.052	1	0.8196

The above Table 9 shows that the p-value was higher than 0.05. This leads to the conclusion that we don't reject the null hypothesis of no autocorrelation. Therefore, the selected model is an appropriate one for forecasting road traffic accident of Addis Ababa.

Conclusion

This study examined statistical analysis of road traffic accident in Addis Ababa using SETAR model. In the nonlinear SETAR modelling, graphical method, Keenan test, Tsay test and likelihood ratio test were used check non linearity features of the data. A delayed parameter was selected as (d=5) with eight order of lower and upper regime using grid search method. Henceforth, SETAR (2,8, 8) was identified amongst the tentative models and out of sample forecasts were made for 24 months. The result indicates that, an increasing trend of road traffic accident in Addis Ababa.

Recommendation

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

The government should create awareness for drivers, traffic polices and pedestrians about traffic accident and the rules and regulations to reduce the road traffic accidents related morbidity and mortality.

This study showed the increasing pattern of road traffic accident over the forecasted period and recommends policy maker to pay more attention on preventive measures for road traffic accidents so that the burden can be reduced and more lives can be saved.

In this study, the researcher focused on the application of SARIMA and SETAR models in forecasting road

Traffic accident of Addis Ababa. And, one of the finding of this study reveals that, road traffic accident most frequently occurs during the rainy seasons. Hence, further studies may employ Artificial Neural Network (ANN) and Multinomial Logit model to identify the impact of climate change on RTAs using different sets of rainfall data.

REFERENCE

1. Abdi, T. A., Hailu, B. H., Adal, T. A., van Gelder, P. H., Hagenzieker, M. P., & Carbon, C. C. (2017). Road crashes in Addis Ababa, ethiopia: empirical findings between the years 2010 and 2014. *African Research Review*, *11*(2), 1-13.
2. Abegaz, T., & Gebremedhin, S. (2019). Magnitude of road traffic accident related injuries and fatalities in Ethiopia. *PloS one*, *14*(1), e0202240.
3. Bekele, T. G., & Sima, T. H. (2019). Road traffic accident cause and effect on socio economy of Addis Ababa city. *Economics And Social Sciences Academic Journal*, *1*(4), 21-37.
4. Clements, M. P., Franses, P. H., & Swanson, N. R. (2004). Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, *20*(2), 169-183.
5. Cryer, J. D. and K.S. Chan.2008. *Time Series Analysiswith Application in R*. 2nd ed. Springer, New York.
6. Ghédira, A., Kammoun, K., & Saad, C. B. (2018). Temporal analysis of road accidents by ARIMA Model: Case of Tunisia. *International*

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

Journal of Innovation and Applied Studies, *24*(4), 1544-1553.

7. Hansen, P. R., & Timmermann, A. (2012). Choice of sample split in out-of-sample forecast evaluation.
8. Franses, P. H., & Van Dijk, D. (2000). *Non-linear time series models in empirical finance*. Cambridge university press.
9. Keenan, D. M. (1985). A Tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, *72*(1), 39-44.
10. Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., & Assimakopoulos, V. (2011). An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. *Journal of the Operational Research Society*, *62*(3), 544-554.
11. Samuel, J. C., Sankhulani, E., Qureshi, J. S., Baloyi, P., Thupi, C., Lee, C. N., ... & Charles, A. G. (2012). Under-reporting of road traffic mortality in developing countries: application of a capture-recapture statistical model to refine mortality estimates. *PloS one*, *7*(2), e31091.
12. So, M. K., & Chen, C. W. (2003). Subset threshold autoregression. *Journal of Forecasting*, *22*(1), 49-66.
13. Tong, H., & Lim, K. S. (2009). Threshold autoregression, limit cycles and cyclical data. In *Exploration Of A Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics* (pp. 9-56).
14. Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika*, *73*(2), 461-466.
15. Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American statistical association*, *84*(405), 231-240.
16. Tsay, R. S. (2005). *Analysis of financial time series* (Vol. 543). John wiley & sons.
17. Van den Bossche, F., Wets, G., & Brijs, T. (2004). *A regression model with ARIMA errors to investigate the frequency and severity of road traffic accidents* (No. RA-2004-35). LUC.

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 12
December 2021

18. World Health Organization. (2018). *Global status report on road safety 2018: summary* (No. WHO/NMH/NVI/18.20). World Health Organization.
19. Yazdani Cherati, J., Ahmadi Baseri, E., & Ghadami, M. (2013). Mapping of mortality rate in suburban accidents, Mazandaran Province, 2007-2010. *Journal of Mazandaran University of Medical Sciences*, *22*(97), 50-58.
20. Zewude, B. T., & Ashine, K. M. (2016). Statistical Modeling on Determinants of Traffic Fatalities and Injuries in Wolaita Zone, Ethiopia. *Global Journal of Human-Social Science Research*.
21. Zivot, E. and Wang, J., (2005).*Modeling financial series with S-PLUS*, Springer, 2nd edition, .651-676.