International Journal of Development Mathematics Vol 2 Issue 2 Page No. 314 - 327

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Transformed-Arimax Model for Heavy Tailed Distributions

Kennedy I. Ekerikevwe[a*] and Tayo K. Oyeleke[b]

[a]*Department of Statistics, Delta State Polytechnic, Otefe-Oghara, Delta State.*

[b]*National Bureau of Statistics, Abuja*

ARTICLE INFO

Article history:
Received 20 March 2025
Received in revised form 10 May 2025
Accepted 20 June 2025

Keywords:
Transformed-ARIMAX, Time Series Analysis, Performance, Forecastability, Hybrid.

MSC 2020 Subject classification:
62-XX, 62M10, 62E17.

ABSTRACT

This study develops a Transformed Autoregressive Integrated Moving Average with covariate X by taking the logarithm of Arimax (Log-ARIMAX) model for high frequency time series data that is coupled with external time-varying covariate(s) with heavy tailed distributional lognormal form of a residual structure. This study also evaluates both the in-sample and out-sample forecasting accuracy of two forecasting models namely ARIMAX and Log-ARIMAX. A Log-ARIMAX model for time series data with heavy-tailed trait to analyse the obtained data is proposed. The Generalised Linear Method (GLM) was used to estimate the parameters of the proposed model. The oil spill data used was collected both monthly and yearly and was derived from four Oil and Gas companies from 2005-2020 with a total of 64 observations. The Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) serve as the error metrics (Forecast Accuracy Measures) in evaluating the forecastability of the models. Also, Diebold and Mariano test of Accuracy was employed to test the significance of the models. The results of analysis show that Log-ARIMAX models performed better as compared to ARIMAX models in both time regimes. The study recommended that the Transformed-Arimax model is good for a time series data with heavy tailed distributions.

1. Introduction

Statistical methods and models are either linear or non-linear based on some assumptions theoretically and analytically. These assumptions led to the splitting of approach of dealing with time varying observations (time series) models into two approaches; Time domain (otherwise known as probabilistic approach) and frequency domain (spectral function) analyzes. The time domain approach relies solely on either dependence or independency that are continuous or discrete time variant uniformly interval series or daily, hourly, weekly, monthly, quarterly, yearly or bi-annually series (Akouemo and Povinelli, 2014).

ARIMAX model comes in when time series are affected by special events such as environmental regulations, legislative activities, policy changes, and similar events, which might be referred to as augment, supportive or intervention events. One or more exogenous variables (Xs) can be incorporated in the time series model to be able to predict the value of another series by using a transfer function. The Transfer functions can be used to model and forecast the response series, and to analyze the impact of the intervention. The general transfer function can be employed by the ARIMA procedure discussed by (Box and Tiao, 1975). When an ARIMA model includes other time series as input variables, the model is sometimes referred to as an ARIMAX model. ARIMAX models could also be referred to as dynamic regression models.

The general overview of this research is an extension and modifications of ARMA model. This study introduces

* Corresponding author. Tel.: +2348064647455
E-mail address: kennedyekerikevwe@gmail.com. (Kennedy I. Ekerikevwe)
https://doi.org/10.62054/ijdm/0202.18

a lognormal distribution as the error term (white noise) to the established ARIMAX model, resulting in a regression-like time series model with additional inputs. According to the literature, several scientist have examined ARIMAX related to exogenous covariate(s) by employing a variety of short-memory frequency data. Iqbal et al. (2005) looked into Pakistan's wheat output. The forecasting of Pakistan's wheat production through the year 2022 is based on the study of wheat production. They employed ARIMA technique and production of wheat is the factor considered. Afterward, Chen (2008) utilized ARIMA technique for making transient estimate with property related misconduct for one city of China. One hundred and ten Computer Aided Dispatch records from a community police department were used to create the used dataset; there were fifty weeks of property crime data in the dataset. In addition, Fan et al. (2009) looked at how to apply multiple time series to ARIMAX representation, which may be used for systems like economic, engineering, and biological ones. They were familiar with the ARIMAX's structure and model building, which they maintained. They used a dataset from China's statistical year that included the production values of tertiary industries from 1978 to 2007. The results showed that, in comparison to ARIMA, the applied and modeled multivariate ARIMAX provided a distinct and precise analysis and forecast of yield worth for tertiary firm. The models ARIMA, ARIMAX, SARIMA, and SARIMAX were utilized by Cools et al. (2009) for examining everyday holdup tally. For effective model generalization and predictions, they placed an emphasis on investigating and identifying seasonality effects on everyday traffic figures and implications of holidays at various areas. They have prior knowledge of cyclical patterns and spectral analysis, which they used to calculate the daily traffic count's periodicities. Kinley et al. (2010) used ARIMAX and time series data to create methods for estimating and projecting the prevalence of malaria in Bhutan's falciparum-endemic regions. The study was conducted retrospectively using climatic information provided by Meteorological Unit. Bruce et al. (2013) investigated the long-term disability in social security disability insurance benefit claims and insurance insurers' short-run ahead grade for long-run disability benefit averment. Their work demonstrated ARIMAX and ARIMA methodologies because of their abilities to produce accurate four-quarter forecasts. Anggraeni et al. (2015) stated that, particularly during the Eid holidays, there will likely be an increase in demand for Moslem children dressed in HabibahBusana. Their study compared the ARIMAX multivariate method's univariate data forecast to the ARIMAX multivariate method's independent variables, such as the various Eid holidays observed annually. Their findings revealed that the accuracy of the testing, training, and subsequent time forecasting processes was superior to that of the ARIMA method. They found that there are something like fourteen factors should have been added as exogenous factors in the ARIMAX model to make exactness level not a diminishing one. They certified that ARIMAX model performed better compared to ARIMA technique as far as contrasting and model execution record of AIC and conjecture mistakes of MAPE and RMSE. ARIMAX is better off because it was able to combine the almanac influence variation using dummy parameters and produce forecasting results with a higher level of accuracy than the ARIMA method. Tamuke et al. (2018) forecasted the Headline Consumer Price Index in Sierra Leone and conducted empirical research on both the ARIMA and ARIMAX methodologies. The main objective of evaluating results for a time before and ahead of forecasting is addressed utilizing a static technique was answered.

A heavy tailed distribution is the one that has a tail that is heavier than an exponential distribution. A distribution with heavy tailed goes to zero slower than the one with exponential tails. Heavy tailed distributions tend to have many outliers with very high values. These high values tend to skew your sample statistics; the mean would be very misleading while the sample variance will probably be very large and the sample mean usually underestimates the population mean (Bryson, 1974). Mathematical expectations for heavy-tailed distributions are infinite.

In time series analysis, a time regime also known as time horizon refers to distinct periods or states within a time series where the underlying statistical properties of the data are significantly different. Consequently, the behaviours of a time series model considered in two different time regime can shift between the different time states. Dead time on the other hand, refers to a period where a system is unable to respond to input or events due to processing or recovery. It is a lag in a system response as a result of delay between an input and a measurable output. This delay can be attributable to physical limitations of the system, like the time it takes to transport mass, energy, or data, or by the time it takes for a sensor, controller, or other component to process sets of data.

Laili et al. (2019) made an estimate of the total departure of ship passengers in the main port of Makassar using the ARIMAX method with the effects of calendar variations. They opined that the ARIMAX method is a method that can be used when there are exogenous variables, where in this case the exogenous variable is in the form of variable dummy which is Eid holidays. Their forecasting results show that the ARIMAX method has a relatively small accuracy with the Mean Absolute Percentage Error (MAPE) value.

Ling *et al*. (2019) developed an Autoregressive Integrated Moving Average with external variables (ARIMAX) model which tries to account the effects due to the climatic influencing factors, to forecast the weekly cocoa black pod disease incidence. With respect to the performance measures, it is found that the proposed ARIMAX model improves the traditional Autoregressive Integrated Moving Average (ARIMA) model. The results of this forecasting can provide benefits especially for the development of decision support system in determine the right timing of action to be taken in controlling the cocoa black pod disease.

Farhana and Monzur (2020) studied the development of Autoregressive Integrated Moving Average models with exogenous input (ARIMAX) to forecast autumn rainfall in the South West Division (SWD) of Western Australia (WA). The developed ARIMAX model is found helpful to overcome the difficulty in seasonal rainfall prediction as well as its application can make an invaluable contribution to stakeholders' economic preparedness plans. Nimish et al. (2021) in their paper compared both methods' preprocessing performance when applied to seasonal time series data with varying time resolutions and complex trend patterns for different content of outliers through detailed result analyses. Further, a new metric to measure outlier correction capability was suggested.

Abdallah (2021) used Gross Domestic Product (GDP) and consumer price index (CPI) as significant indicators to describe and evaluate economic activity and levels of development. His paper aimed at modeling and predicting GDP and CPI in Jordan. In order to achieve this goal, their study applied the Box- Jenkins (JB) methodology for the period 1976-2019.

Ugoh (2021) proposed an appropriate ARIMAX model that is used to forecast the Nigeria's GDP. The data used for the study is sourced from the World Bank for a period of 1990-2019. The ARIMA model is fitted on the residuals using Box-Jenkins approach. The Bayesian Information Criterion (BIC) is adopted to assess the adequacy of the models. The raw data satisfied the assumption of multicollinearity when export is eliminated and the residual series is stationary after the first differencing. This study shows that import is a significant exogenous variable for the GDP dynamics. Zhou et al. (2021) design an efficient transformer-based model for LSTF, named Informer, with three distinctive characteristics; a ProbSparse Self-attention mechanism, which achieves in time complexity and memory usage, a self-attention distilling highlights dominating attention by halving cascading layer input and efficiently handles extreme long input sequences, and a generative style decoder. Extensive experiments on four large-scale datasets demonstrate that the informer significantly outperforms existing methods and provides a new solution to the LSTF problem.

Some of these literatures reviewed considered ARIMAX model for short-memory data only. The motivation to propose and formulate a hybrid model whose distributional form would be robust and sufficient in capturing and accommodating both the external covariate (s) and the heavy-tailed properties of long-memory (high range) observational time series events becomes necessary. In view of this, this study, therefore, proposes a Transformed-ARIMAX model to capture and accommodate both the external covariate(s) and the heavy-tailed properties of observational time series events using secondary datasets of the long memory types of oil spillage and temperatures. This review, subsequently, has shown that the conventional ARIMAX model has little or no strength to manage environmental change and ecological data that are normally impacted by kurtosis, skewness, exceptions, high recurrence or long memory observational time occasions; hence the gap to propose and formulate a hybrid time series model to capture and accommodate long memory or heavy fluctuation series with heavy tailed distribution. To model time series data with heavy-tailed characteristic, a transformation of ARIMAX would be propounded to achieve improved forecasting accuracy.

2. Methods

A secondary dataset of oil spills was used in this study. The oil spill data used was collected both monthly and yearly and was derived from four Oil and Gas companies from 2005-2020 with a total of 64 observations The monthly oil spillage information recorded by four oil and gas organisations from January 2005 to December 2020 was obtained from Bunge Petroleum (BG), Base Petroleum (BP), Caine Energy (CNE), and Tullow Oil & Gas (TLW). The data analysis was carried out using SPSS and R software.

2.1 Autoregressive Model

An autoregressive model is a cycle used to foresee what was in store in light of gathered information from an earlier time. Because there is a connection between the two, it is possible. Any random procedure whose output is dependent on any previous values which is denoted for this model. First-order autoregressive representation makes assumption about present worth is determined with immediately preceding worth. On the other hand, there are instances in which the present value may be dependent on two previous values. As a result, time series play a significant role in an autoregressive model and are utilised in accordance with the circumstance and desired outcome.

The model is given by,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots\ldots\ldots\ldots\ldots\ldots\ldots \phi_p y_{t-p} + \ell_t \tag{2.1}$$

Stationarity Condition: $\phi_1 + \phi_2 < 1, \qquad \phi_2 - \phi_1 < 1, \qquad |\phi_2| < 1$

where;

y_t is a single-variate component having periodic measurements over a continuous span of duration.

y_{t-i} is a single-variate parameter including fixed duration measurements including lag*i.*

ϕ_0 is the parameter of the autoregressive.

ϕ_i is the coefficient of autoregressive for various lags and has to be rigidly below 1. Such that the error term ε_t follows a standardized normally distributed variate, that is $\varepsilon_t \sim (0, \sigma^2)$.

2.2 Moving Average Model

A time series representation which takes into consideration relatively short-term correlations is known as moving average model. In essence, this mean of all previous observations will be the next observation. The MA(q) is obtained specifically by checking the ACF graph which is given as

$$X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_1\varepsilon_{t-q} \quad (2.2)$$

$$= \theta_0 + \sum_{i=1}^{q} \theta_i\varepsilon_{t-i} \quad (2.3)$$

where,

ε_t is the residual noise for fixed span of periodic measurements.

ε_{t-i} a residual noise for fixed span of periodic measurement for lag i.

θ_0 a constant for moving average.

θ_i is the coefficient of moving average for various lag and has to be rigidly below 1.

2.3 Autoregressive Moving Average Model

The AR and MA models are combined in the Autoregressive Moving Average (ARMA) model with desirable accuracy in precision, forecasting, and parsimonious in parameterization. It is widely used because of its flexibility and widely used application. It was predicated on the idea that the residual will consist of a series of independently and identically distributed random parameters having zero mean and variance(σ^2).

That is,

$$Y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1}$$

$$-\theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \quad (2.4)$$

White noise $\varepsilon_t \sim (0, \sigma^2)$

$$E(\varepsilon_t) = 0 \quad (2.5)$$

$$E(\varepsilon_t\varepsilon_T) = \begin{cases} \sigma^2 & \text{for t} = T \\ 0 & \text{otherwise} \end{cases} \quad (2.6)$$

With parameters $(\varphi_1, \varphi_2, \ldots, \varphi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)$ to be estimated based on the observational time varying values of Y. It is necessary for estimation to assume that all the roots of $\varphi_1, \varphi_2, \ldots, \varphi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ lies outside the unit circle, that is all the coefficients of AR and MA to be estimated must be less than unity, that is, $\varphi_1, \varphi_2, \ldots, \varphi_p > 1$ and $\theta_1, \theta_2, \ldots, \theta_q > 1$ to satisfy the standard regularity stationarity condition.

For the Differencing,

$$\nabla^1 \mathrm{Y_t} = (1 - \mathrm{B})^1 \mathrm{Y_t} \nabla^1 \mathrm{Y_t} = (1 - \mathrm{B})^1 \mathrm{Y_t} = \mathrm{Y_t} - \mathrm{Y_{t_{-1}}}$$

$$\nabla^2 \mathrm{Y_t} = (1 - \mathrm{B})^2 \mathrm{Y_t}$$

$$= (1 - 2\mathrm{B} + \mathrm{B}^2)\mathrm{Y_t} \quad (2.7)$$

$$= \nabla^1 \mathrm{Y_t} - \nabla^1 \mathrm{Y_{t-1}} = (\mathrm{Y_t} - \mathrm{Y_{t-1}}) - (\mathrm{Y_{t-1}} - \mathrm{Y_{t-2}})$$

$$\nabla^3 Y_t = (1-B)^3 Y_t = Y_t - 3Y_{t-2} + Y_{t-3} \quad (2.8)$$

2.4 Autoregressive Integrated Moving Average with Covariates "X"

According to Yang and Wang (2017), Autoregressive Integrated Moving Average with Covariates "X" (ARIMAX) model which is an improved version of the ARMA makes up the room for incorporating exogenous variables or covariates in order to improve comprehensiveness, supportive items (dependency) and forecasting.

The abstraction of reality of the ARIMAX can be defined as:

$$Y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots\cdots + \varphi_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (2.9)$$

Where $x_t \cdots\cdots x_{t-p}$ are the p-lagged period of the exogenous covariates (x_{t-p}) with errors that are independently and identically distributed with mean zero, variance (σ^2) and covariance of zero.

Otherwise,

$$\varphi_p(B) Y_t = \varphi_0 + \varphi(B) x_t + \theta_q(B) \varepsilon_t \quad (2.10)$$

$$\varphi_p(B) \nabla^d Y_t = \varphi_0 + \varphi(B) x_t + \theta_q(B) \varepsilon_t \quad (2.11)$$

For ARMAX and ARIMAX respectively, such that

$Y_t \Rightarrow$ The output observational series (in regression, term as dependent variable)

$x_t \Rightarrow$ The input observational series (in regression, term as independent variable/covariates)

$\varepsilon_t \Rightarrow$ The series noise or stochastic disturbance, it is to be noted that it is independent of the input series

$\varphi(B)x_t \Rightarrow$ is known as the transfer function (otherwise called link function or impulse response function) that link x_t to y_t through distributed lag.

$$\varphi(B) x_t = [\varphi_0 + \varphi_1 B + \varphi_2 B^2 + \cdots\cdots] X_t \quad (2.12)$$

$\varphi_1, \varphi_2, \ldots$ in eq. (2.12) are regarded as the infinite coefficients of the regression impulse weights of the responses that could be a non-negative or negative. Suppose the number of the impulse weights is equal to "b" (known as dead time) and rewriting the link function as ratio of distributed lag polynomial time of a finite lag to a low ordered polynomial lag in B.

$$\varphi(B) x_t = \frac{\eta_h(B) B^b}{\lambda_r(B)} X_t \quad (2.13)$$

$$\text{So, } Y_t = \sum_{j=1}^{n} \frac{\eta_h(B) B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B) \varepsilon_t}{\varphi_p(B)} \quad \text{(ARMAX)} \quad (2.14)$$

where;

$$\sum_{j=1}^{n} \frac{\eta_h(B) B^b}{\lambda_r(B)} X_t = \left(\sum_{i=0}^{\infty} \varphi(B) x_t\right) B^b = \sum_{i=0}^{\infty} (\varphi_i B^i) B^b \quad (2.15)$$

$$= \varphi_0 B^b + \varphi_1 B^{b+1} + \varphi_2 B^{b+2} + \varphi_3 B^{b+3} + \cdots\cdots \quad (2.16)$$

Equation (2.16) could be written in terms of Integrated, that is, in terms of ARIMAX as;

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\nabla^d \varphi_p(B)} \quad \text{(ARIMAX)} \tag{2.17}$$

2.5 Transformed-Autoregressive Integrated Moving Average-X (Log-ARIMAX)

For long-memory (highly frequency) observational series, ARMAX or ARIMAX, the distributional form of (ε_t) is then given as

$$f(y_t) = \frac{1}{y_t \sigma\sqrt{2\pi}} \exp\left[-\left(\frac{(\ln(y_t))^2}{2\sigma^2}\right)\right], \; y_t > 0 \tag{2.18}$$

or

$$f(\varepsilon_t) = \frac{1}{\varepsilon_t \sigma\sqrt{2\pi}} \exp\left[-\left(\frac{(\ln(\varepsilon_t))^2}{2\sigma^2}\right)\right], \; \varepsilon_t > 0 \tag{2.19}$$

Because, the error term and the observational series share the same distributional form

$$\text{With } y_t \sim \varepsilon_t \sim N\left[\exp\left(\frac{\sigma^2}{2}\right), \exp(2\sigma^2) - \exp(\sigma^2)\right] \tag{2.20}$$

For log-ARMAX,

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)} \sim N\left[\exp\left(\frac{\sigma^2}{2}\right), \exp(2\sigma^2) - \exp(\sigma^2)\right] \tag{2.21}$$

For log-ARIMAX,

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\nabla^d \varphi_p(B)} \sim N\left[\exp\left(\frac{\sigma^2}{2}\right), \exp(2\sigma^2) - \exp(\sigma^2)\right] \tag{2.22}$$

The Transformed-ARIMAX model is the proposed as shown in equation (2.22) above.

2.6 Optimal Model Selection Criteria

For determining an aim order, numerous variables are brought into play in earlier studies. They involve Akaike's information, Schwarz-Rissanen, Bayesian estimation, Hannan-Qiunn criterion and others. One of the most recent options for model criterion is the Akaike's information corrected criteria (AICC), which was developed by (Hurvich and Tsai, 1989).

$$AIC = -2\ln Likelihood\left(\hat{\varphi}, \hat{\theta}, \hat{\sigma}^2\right) + 2(p+q+1) \tag{2.23}$$

$$BIC = (n-p-q)\ln\frac{n\hat{\sigma}^2}{n-p-q} + n\left(1+\ln\sqrt{2\pi}\right) + (p+q)\ln\left[\frac{\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}^2}{p+q}\right] \tag{2.24}$$

$$BEC = \hat{\sigma}^2 + (p_x+q_x)\hat{\sigma}_x^2 \ln\frac{n}{n-p_x-q_x} \tag{2.25}$$

Akaike's information criterion, Bayesian information criteria and Bayesian estimation criteria respectively.

2.7 Forecasting Accuracy Measures

If the actual values of the series to be forecasted are observed, forecasts can be evaluated once they have been made. The precision of forecasts can be measured in a few different ways. These are the Mean Absolute Percentage Error (MAPE), the Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE). According to Olatayo and

Taiwo (2016), the following are the forecast accuracy measures.

$$MAE = \frac{1}{h+1}\sum_{t=s}^{h+s}\left(\hat{X}_t - X_t\right)^2 \quad (2.26)$$

$$RMSE = \sqrt{\frac{1}{h+1}\sum_{t=s}^{h+s}\left(\hat{X}_t - X_t\right)^2} \quad (2.27)$$

$$MAPE = \frac{100}{h+s}\sum_{t=s}^{h+s}\left|\frac{\hat{X}_t - X_t}{\hat{X}_t}\right| \quad (2.28)$$

Diebold-Mariano (DM): $S^* = \left[\frac{n+1-2h+n^{-1}h(h-1)}{n}\right]^{1/2} S$ (2.29)

3.0 RESULTS

The graph below shows the time plots of the observed data in two different time horizons.

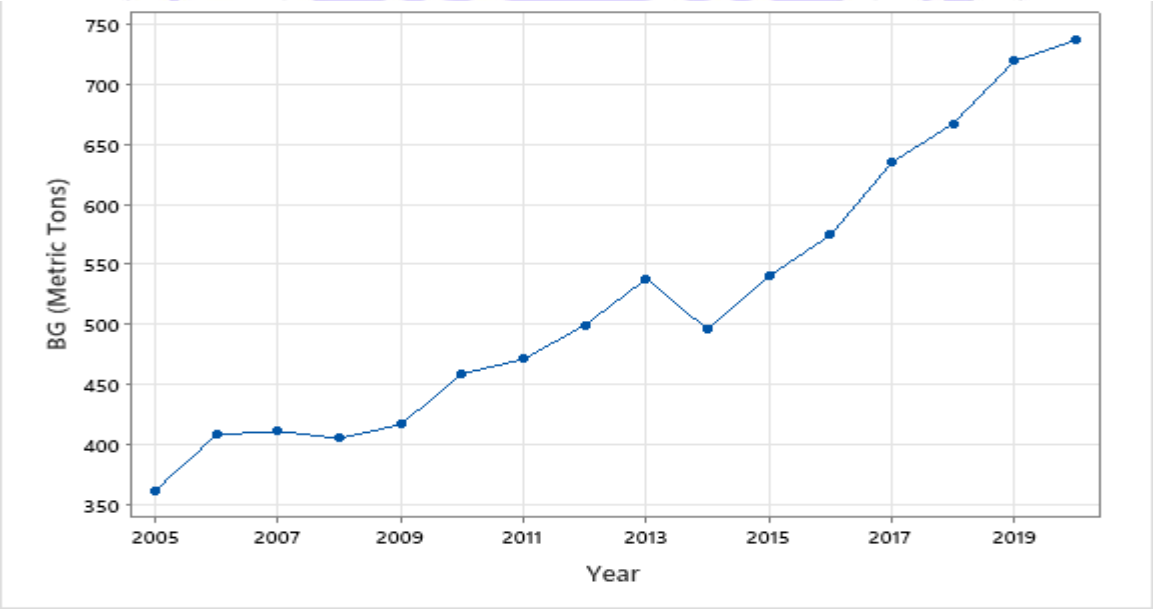

Figure 1: A time plot of the observed data for BG (2005-2020)

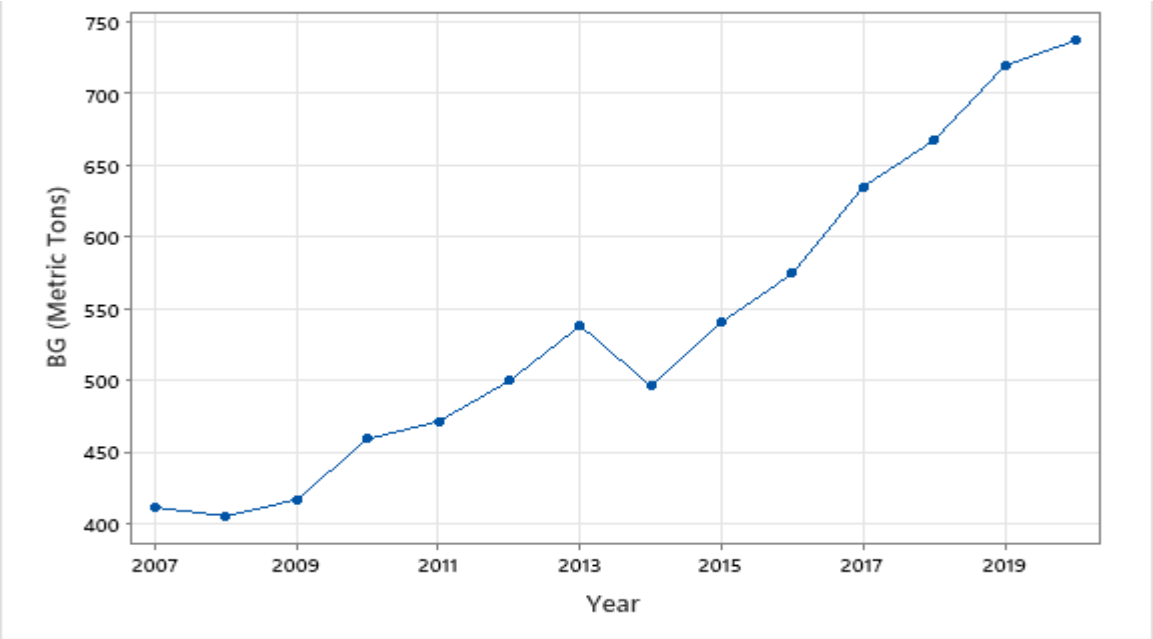

Figure 2: A Time Plot of the Observed Data for BG (2007-2020)

Table 1: Correlation Matrix –First Time Horizon (2005 - 2020)

	BG	BP	CNE	TLW
BG	**1.0000**	**-0.2390**	**0.4411**	**0.2112**
BP	**-0.2390**	**1.0000**	**0.0043**	**-0.2619**
CNE	**0.4411**	**0.0043**	**1.0000**	**0.6301**
TLW	**0.2112**	**-0.2619**	**0.6301**	**1.0000**

Table 2: Correlation Matrix –Second Time Horizon (2007 - 2020)

	BG	BP	CNE	TLW
BG	**1.0000**	**-0.0578**	**0.5543**	**0.6106**
BP	**-0.0578**	**1.0000**	**-0.0789**	**-0.0601**
CNE	**0.5543**	**-0.0789**	**1.0000**	**0.755**
TLW	**0.6106**	**-0.0601**	**0.755**	**1.0000**

Table 3: Results for ARIMAX and LOG-ARIMAX Models Selection (BG)

Ticker	Model Type	Selected Model	AIC
DBG	ARIMAX	(0,1,2)	781.65
DBG	LOG-ARIMAX	(0,1,2)	765.72
DBG*	ARIMAX	(0,1,2)	533.38
DBG*	LOG-ARIMAX	(0,1,2)	525.53

Table 4: Estimation of Model Parameters (BG)

ESTIMATES	ARIMAX	LOG-ARIMAX	ARIMAX*	LOG-ARIMAX*
b	-	0.6785	-	0.6366
AR (1)	-	-	-	-
AR (2)	-	-	-	-
AR (3)	-	-	-	-
MA (1)	-0.0199	-0.0297	-0.022	-0.033
MA (2)	0.2916	0.3322	0.3103	0.3367
MA (3)	-	-	-	-

The parameter estimates in both time regimes is presented in Table 4.

Table 5: Error Metrics (Forecast Accuracy Measures) (BG)

	Test Type		
Ticker	MAE	RMSE	MSE
ARIMAX	44.7725	56.5525	3198.1830
LOG-ARIMAX	36.80373	49.8227	2482.3040
ARIMAX*	53.4383	65.2898	4262.7570
LOG-ARIMAX*	42.6022	54.0129	2917.3950

Table 6: Diebold-Mariano Test for Comparing Models (BG)

Ticker	Test Type	p – value
ARIMAX	DM	**< 0.0018**
LOG-ARIMAX	DM	**< 0.0001**
ARIMAX *	DM	**< 0.0049**
LOG-ARIMAX *	**DM**	**< 0.0001**

Table 7: Results Summary (BG)

Mod	ARIMAX	LOG-ARIMAX	ARIMAX *	LOG-ARIMAX *
L.C	0.911	- 0.249	+ 0.811	**- 0.068**
AIC	781.65	765.72	533.38	**525.53**
MA	44.77	49.82	53.44	**42.60**
RM	56.55	49.82	65.29	**54.01**
MS	**3198.18**	**2482.30**	**4262.757**	**2917.39**

4. Discussion

From the analysis above, Figure 1 and Figure 2 are the time plots of the observed data for the two different time horizons which show an upward pattern of growth in the oil spill data from BG, BP, CNE and TLW. Besides, the graphs depict heavy fluctuations and outliers in the observed oil spill data in the two time horizons. The diagram shows that every new model performed well because each model followed the time plot of the observed data. However, candidate models followed the observed data much more closely in the first regime (2007–2020) than in the second regime; the error metrics for the second time regime are significantly higher than those for the first time regime. The diagram in the second time system isn't generally so minimized as the one in the initial time system. Essentially, the Akaike Information Criterion (AIC) and the linear correlation between the oil spills that are being considered, taking into account the respective data histories, are fundamental measures that will play a significant role in the candidate methods' forecastability. Results from the analysis above, Tables 1 and 2 show the linear correlation between the considered oil spills of the four oil companies in the two time zones of 2005-2020 and 2007-2020 respectively. The results show that the volumes of oil spills from the four oil companies are not significantly correlated. None of the random walk test of all the considered oil spills in the Oil and Gas Industry was significant both with homoskedastic and heteroskedastic errors. Table 3 shows that the Log-ARIMAX model has the least AIC in the two time horizon as compared to the classical ARIMAX model. This implies that the LOG-ARIMAX model has a better forecasting strength and accuracy as compare to that of ARIMAX model. Tables 4 and 5 show estimation of model parameters and error metrics (forecast accuracy measures) respectively. The error metrics (MAE, RMSE, and MSE) indicate that the LOG-ARIMAX model provides better forecasting accuracy than the traditional ARIMAX model. These results are in line with the research carried out by Shumway and Stoffar (2010) and Olatayo and Taiwo (2016). The outcome of the Diebold and Mariano test in this research is in consonant with the assertion by Diebold and Mariano (2002). A

Diebold and Mariano test of a model of the univariate variable with an exogenous variable identified them as having distinct forecasting abilities in each case. The Diebold-Mariano test as depicted in Table 6 demonstrates that, in the first regime, LOG-ARIMAX models outperform conventional ARIMAX models when it comes to forecasting. Likewise, in the subsequent system, the test shows that LOG-ARIMAX has a superior estimating strength when contrasted with ARIMAX models. The summary results are clearly comparable as shown in Table 7.

5. Conclusion

With reference to the first objective of this thesis, it is empirically evident that ARIMAX model with an exogenous variable (LOG-ARIMAX) performed creditably well in all cases and scenarios as outlined in chapter four. This emphasizes that, when improving the in – sample forecasting accuracy of oil spills using the Box – Jenkins model, it is in order to incorporate an exogenous variable to further augment the accuracy of the in – sample forecast. In this paper, historical adjusted oil spills recorded by four Oil and Gas companies in Nigeria were use as possible exogenous variable or as public information.

On the other hand, linear correlation between the ARIMAX model with exogenous variable did very little to improve the in-sample forecasting accuracy of all the considered scenarios in this thesis. In most cases, the high and low linear correlation between oil spills of candidate models only gave signal to the corresponding Akaike Information Criterion (AIC) value. High correlation in most cases gave a lower value of the AIC and vice-versa. However, this assertion was not consistent. Evidently, the Diebold and Mariano test of accuracy is dependent AIC of the candidate models. However, in most cases smaller AIC values turn to minimize the considered error metrics (i.e., MAE, RMSE and MSE) and vice versa. This is evident throughout the results. The linear correlation on the hand had little or no impact on the performing models.

The Box-Jenkins Method with/without an exogenous variable supports the semi – strong form of EMH. Thus, the information, $\Omega_t$set comprising of the past and current oil spills and all publicly available information supports the Efficient Market Hypothesis (EMH) in its semi-strong form. Timmermann and Granger (2004) in their paper "Efficient market hypothesis and forecasting" argued that traditional time series forecasting methods relying on individual forecasting models or stable combinations of these are not likely to be useful. This in one way or the other confirms our findings that even though log-ARIMAX model is an improvement of an ARIMAX model in most cases.

This study proposes a hybrid ARIMAX model to capture and accommodate both the external covariate(s) and the heavy-tailed properties of observational time series events using secondary datasets of the long memory types of oil spillage. As part of contribution to knowledge, this study has been able to propose a hybrid time series model termed as Transformed-ARIMAX model. This model is more robust, efficient and gives reliable predictions when used to model high range or long-memory observational time series events which have external covariate (s) and heavy-tailed properties as compare to the classical ARIMAX models. Since this work is limited to one exogenous variable; we therefore suggests that more than one exogenous variables be imposed into the developed model for further studies.

References

Abdallah, G. (2021). Applying the ARIMA model to the process of forecasting GDP and CPI in the Jordanian Economy. International Journal of Financial Research 12(3):70-86

Akouemo, H.N., & Povinelli, R.J. (2014).Time series outlier detection and imputation. IEEE PES General Meeting—Conference & Exposition, 1–5.

Box, G.E.P. & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*. 70(349): 70-79.

Bryson, M. (1974). Heavy Tailed Distributions: Properties and Tests. Technometrics 16(1):61-68

Bruce, H., Andrews, M. D., Dean, R. S. & Caroline, C. (2013). Building ARIMA and ARIMAX models for predicting long-term disability benefit application rates in the public/private sectors. Journal of Actuaries Society 15(2):16-28

Chen, P., Yuan, H., & Shu, X. (2008). Forecasting crime using the arima model, fifth international conference on fuzzy systems and knowledge discovery. 978-0-7695-3305- 6/08, © 2008, IEEE, Doi:10.1109/FSKD.2008.222.

Cools, M., Moons, E., & Wets, G. (2009). Investigating the variability in daily traffic counts through use of arimax and sarimax models for assessing the effect of holidays on two site locations. *Journal of the Transportation Research Board*, *Washington D.C*., 2136: 57–66.

Diebold, F. & Mariano, R. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20. 134-44.

Fan, J., Shan, R., Cao, X. & Li, P. (2009). The analysis to tertiary-industry with ARIMAX model. *Journal of Mathematics Research*, 1(2): 156-163.

Farhana I. & Monzur A. I. (2020) Use of Teleconnections to predict Western Australian seasonal rainfall using ARIMAX model 2020. Hydrology7(3):52

Anggraeni, W.,Vinarti, R. A., & Kurniawati, Y. D.(2015). Performance Comparisons between Arima and Arimax Method in Moslem kids' clothes demand forecasting: Case study. Procedia Computer Science, Third Information Systems International Conference, 72: 630–637.

Iqbal, N., Bakhsh, K., Maqbool, A., & Ahmad, A. S. (2005). Use of the ARIMA model for forecasting wheat area and production in Pakistan. *Journal of Agriculture & Social Sciences*, 1(2):120-122. 5.

Kinley W., Pratap S., Tassanee S., Saranath L., Nicholas, J. W. & Jaranit, K. (2010). Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. *Malaria Journal*, 9**(**251), 23-78.

Laili, N. H. & Khairi, K. F. (2019). A longitudinal study of audit quality differences among independent auditors. *Manh Dung Tran. Journal of Economics and Development*. ISSN: 1859-0020.

Ling, A. S. C., Darmesah, G., Chong, K. P., Ho, C. M. (2019). Application of ARIMAX model to Forecast Weekly Cocoa black Pod disease Incidence. Mathematics and Statistics, 7(4A):29-40

Nimish, J., Shraddha, S.B. & Rajanarayan, P. (2021). Performance comparison of two statistical parametric methods for outlier detection and correction. IFAC-Papers OnLine, 54(16), 168-174.

Olatayo, T. O. and Taiwo, A. I. (2016) modelling and evaluating performance with neural network using climate time series data. *Nigerian Journal of Mathematics and Applications,* 25; 205-216.

Stoffer, D.S. & Shumway, R. H. (2010). Time series analysis and its applications with R examples. Springer. 3rd (Ed.).

Tamuke, E. Jackson, E. A. & Sillah, A. (2018). Forecasting inflation in Sierra Leone using ARIMA and ARIMAX: A comparative evaluation, model building and analysis team (2018). *Journal of* **Theoretical and Practical Research in the Economic Fields**, **1**(17).

Timmermann, A., & Granger C.W.J. (2004). Efficient market hypothesis and forecasting. *International Journal of*

Forecasting, 20(1), 15–27.

Ugoh, C. I., Uzuke, C. A. & Ugoh, D.O. (2021). Application of ARIMAX model on forecasting Nigeria's GDP. America *Journal of Theoretical and* Applied Statistics, 10(5):216.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of AAAI.