Book of Proceedings, 14th Nigeria Association of Hydrological Sciences Conference (Okitipupa 2024) held at Olusegun Agagu University of Science and Technology, Okitipupa, Ondo State, Nigeria, November 5 - 8, 2024

TIME SERIES ANALYSIS AND FORECASTING OF LONG-MEMORY DATA: AN APPLICATION TO CLIMATE CHANGE

K. I. EKERIKEVWE[1] and T.O. OLATAYO[2]

[1]Department of Statistics, Delta State Polytechnic, Otefe-Oghara, Delta State, Nigeria

[2]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye

Email: kennedyekerikevwe@gmail.com.

ABSTRACT

This study is designed to develop a hybrid time series model. A statistical model is said to be hybrid if it combines two or more existing models for a better and efficient performance. Our new hybrid model will be used to model time series data such as those generated by climate change and environmental agents. Data generated by climate change and environmental agents are usually not normally distributed hence they are characterised as heavy. Literature showed that quite a number of researchers have studied ARIMAX associated with exogenous covariate (s), using different short-memory frequency data, with little or no strength to capture long-memory (high frequency) observations with heavy tailed traits. Having in mind that conventional ARIMAX model has been rarely applied to any of the climate change and environmental agents which are the most cognate agent with associated exogenous variables and are usually characterized by kurtosis, skewness, outliers, long memory (high frequency) and large fluctuation series; this study, therefore, proposes a more robust and sufficient model that would be needed for modeling time series observational data with heavy tailed traits.

Keywords: Time, Long-Memory Data, Distributions, Model, Tailed, Traits.

1.0 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 (correlated and ordered) or independency that are continuous or discrete time variant uniformly interval series or observations (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 endogenous 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 both 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, otherwise known to be ARIMAX model as dynamic regression.

The general overview of this research is an extension and modifications of ARMA model. The lognormal distribution will be introduced as the error term (white noise) to the ready established ARIMAX model (that is, regression like time series model) with additional inputs. The initiation was sourced for from literature and findings that ARIMAX with standardized white noise from Gaussian supported by short memory-time varying series.

2.0 REVIEW OF RELATED LITERATURE

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 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 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 Auto–Regressive 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 can help 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 is 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 satisfy 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: (i) a ProbSparse Self-attention mechanism, which achieves (Llog L) in time complexity and memory usage, and has comparable performance on sequences' dependency alignment. (ii) the self-attention distilling highlights dominating attention by halving cascading layer input, and efficiently handles extreme long input sequences. (iii) the generative style decoder, while conceptually simple, predicts the long time-series sequences at one forward operation rather than a step-by-step way, which drastically improves the inference speed of long-sequence predictions. Extensive experiments on four large-scale datasets demonstrate that Informer significantly outperforms existing methods and provides a new solution to the LSTF problem.

3.0 MATERIALS AND METHODS

According to Chen (2019) and Jonathan and Kung-Sik (2008), Autoregressive (AR) model is a time –varying model that changes variable that regresses on it order or lagged. That is,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots\cdots + \phi_p y_{t-p} + \varepsilon_t \qquad (3.1)$$

$$= \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \qquad (3.2)$$

Moving Average (MA) is a time series model that merges the dependency component between time observational series and its white noise (residual noise) via time lagged (Patel et al, 2017). That is,

$$\varepsilon_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots\cdots + \theta_q \varepsilon_{t-q} \qquad (3.3)$$

$$= \theta_0 + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (3.4)$$

Autoregressive Moving Average (ARMA) model is considered the mixture of AR and MA .That is,

$$Y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \qquad (3.5)$$

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

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

Otherwise,

$$\varphi_p(B) Y_t = \mu + \theta_q(B) \varepsilon_t \qquad (3.6)$$

Or

$$\varphi_p(B) \nabla^d Y_t = \varphi_0 + \theta_q(B) \varepsilon_t \qquad (3.7)$$

For Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) respectively

where;

ε_t is the white noise at time $t \ni \varepsilon_t \sim N(0, \sigma^2)$

B is the background shift operator defined as $B^p Y_t = Y_{t-p}$

$\nabla^d = (1 - B)^d$ is the differencing operator of order "d"

$\varphi(B) = \left(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p\right)$ is Autoregressive polynomial of order "p".

$\theta(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\right)$ is Moving Average polynomial of order "q".

μ is the constant parameter.

It is necessary for estimation to assume that all the roots of $\varphi(B)$ and $\theta(B)$ lie outside the unit circle. Equations (3.8) and (3.9) can be simplify as

ARMA, for $$Y_t = \varphi_0 + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)} \quad (3.8)$$

ARIMA, for $$\nabla^d Y_t = \varphi_0 + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)} \quad (3.9)$$

For the Differencing,

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

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

$$= (1 - 2B + B^2) Y_t \quad (3.10)$$

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

$$\nabla^3 Y_t = (1 - B)^3 Y_t = Y_t - 3Y_{t-2} + Y_{t-3} \quad (3.11)$$

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 (dependents) and forecasting.

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

$$Y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots\cdots + \varphi_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (3.12)$$

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

Otherwise,

$$\varphi_p(B) Y_t = \varphi_0 + \varphi(B) x_t + \theta_q(B) \varepsilon_t \quad (3.13)$$

$$\varphi_p(B) \nabla^d Y_t = \varphi_0 + \varphi(B) x_t + \theta_q(B) \varepsilon_t \quad (3.14)$$

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.

$$\varphi(B)x_t = [\varphi_0 + \varphi_1 B + \varphi_2 B^2 + \cdots\cdots]X_t \quad (3.15)$$

$\varphi_1, \varphi_2, \ldots$ in eq. (3.15) 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.

$$\varphi(B)x_t = \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t \quad (3.16)$$

So, $$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)} \quad \text{(ARMAX)} \quad (3.17)$$

where;

$$\sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t = \left(\sum_{i=0}^{\infty} \varphi(B)x_t\right)B^b = \sum_{i=0}^{\infty} (\varphi_i B^i)B^b \quad (3.18)$$

$$= \varphi_0 B^b + \varphi_1 B^{b+1} + \varphi_2 B^{b+2} + \varphi_3 B^{b+3} + \cdots\cdots \quad (3.19)$$

Equation (3.17) could be written in terms of Integrated, that is, in terms of ARIMAX as;

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t + \frac{\theta_q(B)\varepsilon_t}{\nabla^d \varphi_p(B)} \quad \text{(ARIMAX)} \quad (3.20)$$

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

$$f(y_t) = \frac{1}{y_t\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{(\ln(y_t))^2}{2\sigma^2}\right)\right] y_t > 0 \quad (3.21)$$

or

$$f(\varepsilon_t) = \frac{1}{\varepsilon_t\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{(\ln(\varepsilon_t))^2}{2\sigma^2}\right)\right] \varepsilon_t > 0 \quad (3.22)$$

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

$$\text{With } y_t \sim \varepsilon_t \sim N\left[\exp\left(\frac{\sigma^2}{2}\right), \exp(2\sigma^2) - \exp(\sigma^2)\right] \quad (3.23)$$

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)} \sim N\left[\exp\left(\frac{\sigma^2}{2}\right), \exp(2\sigma^2) - \exp(\sigma^2)\right] \quad (3.24)$$

For log-ARMAX
and,

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\nabla^d \varphi_p(B)} \sim N\left[\exp\left(\frac{\sigma^2}{2}\right), \exp(2\sigma^2) - \exp(\sigma^2)\right] \qquad (3.25)$$

For log-ARIMAX

Table 3.1: Data of Oil Spills from Four Companies

Years	BG (Metric tons)	BP(Metric tons)	CNE(Metric tons)	TLW(Metric tons)
2005	361.75	429.91	168.44	156.75
2006	408.25	469.21	189.96	178
2007	411.25	457.94	176.17	173.75
2008	405.25	445.83	175.56	160.25
2009	416.75	467.3	184.05	167
2010	459	492.74	204.19	186.5
2011	471.25	533.87	227.22	191.25
2012	499.5	535.99	268.27	215.5
2013	538	571.19	297.05	260
2014	496	529.21	261.6	242.5
2015	540.5	537.69	277.21	264
2016	574.5	524.97	290.84	270
2017	635	576.26	291.9	313.5
2018	667.5	541.96	292.36	298.5
2019	719.5	568.63	321.89	339
2020	737	581.96	352.34	411.5

Source: NBS Yearly Bulletin, 2020

4.0 RESULTS AND DISCUSSION

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

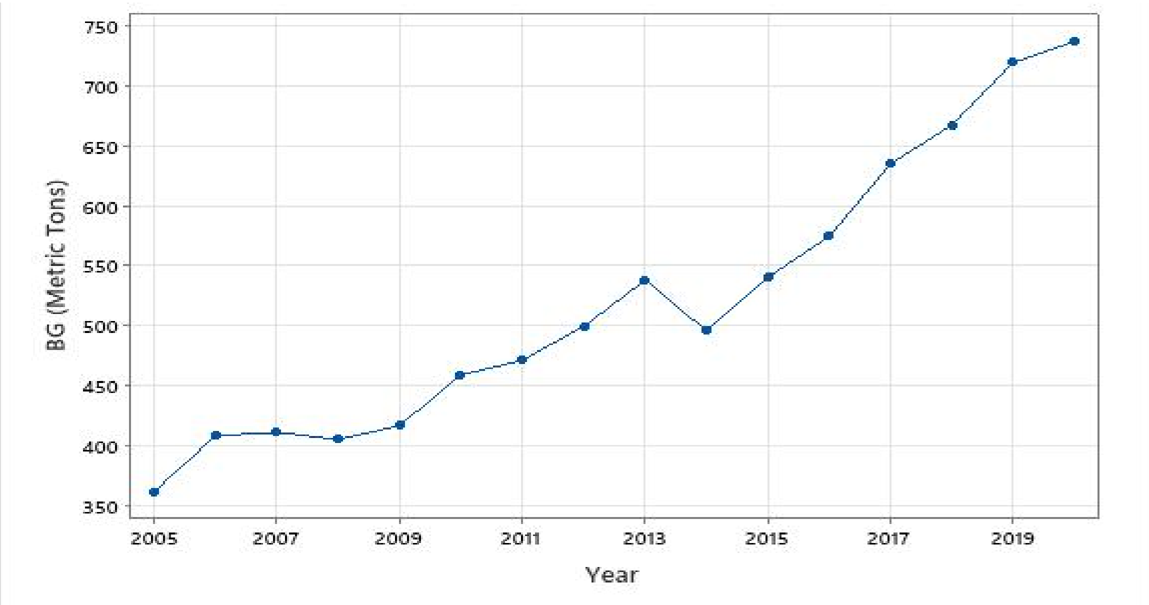

Fig 4.1: A time plot of the observed data for BG (2005-2020)

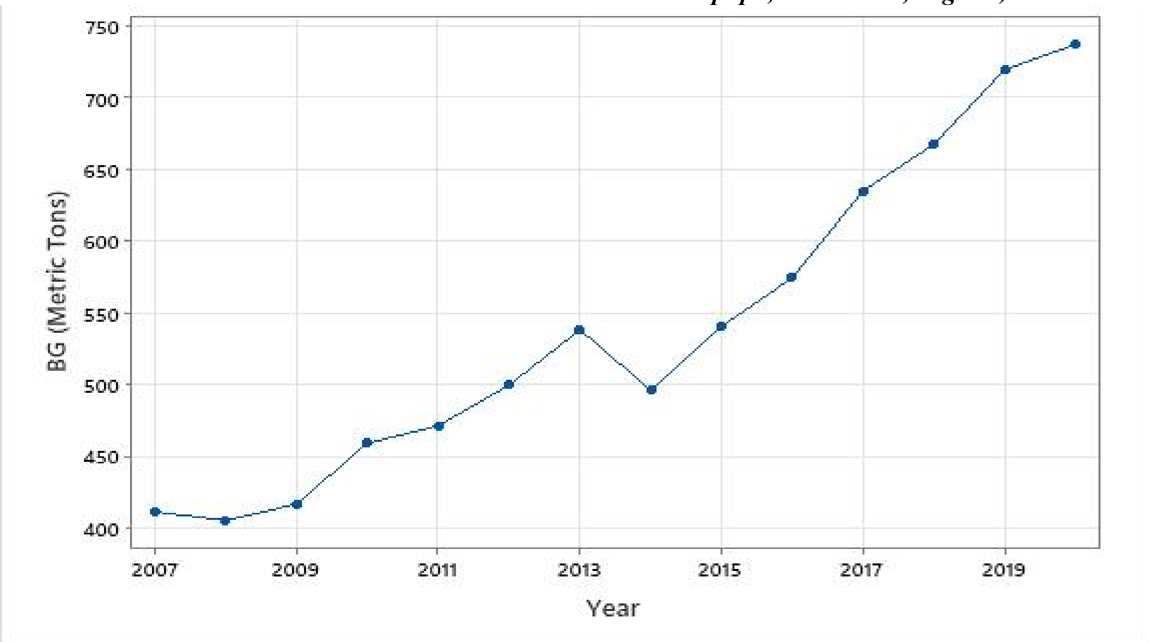

Fig. 4.2: A Time Plot of the Observed Data for BG (2007-2020)

Table 4.1: Correlation Matrix –First Time Horizon (2005 - 2020)

	BG	BP	CNE	TLW
BG	1.0000	-0.2490	0.4511	0.2114
BP	-0.2490	1.0000	0.0043	-0.2619
CNE	0.5511	0.0043	1.0000	0.6301
TLW	0.0114	0.2619	0.1301	1.0000

Table 4.2: Correlation Matrix –Second Time Horizon (2007 - 2020)

	BG	BP	CNE	TLW
BG	1.0000	-0.0678	0.6543	0.4106
BP	-0.0678	1.0000	-0.0789	-0.0601
CNE	0.6543	-0.0789	1.0000	0.755
TLW	0.2106	-0.0601	0.755	1.0000

Table 4.3: 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 4.4: 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 4.4.

Table 4.5: 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 4.6: 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 4.7: Results Summary (BG)

Model	ARIMAX	LOG-ARIMAX	ARIMAX *	LOG-ARIMAX *

L.COR	0.911	- 0.249	+ 0.811	- 0.068
AIC	781.65	765.72	533.38	525.53
MAE	44.77	49.82	53.44	42.60
RMSE	56.55	49.82	65.29	54.01
MSE	3198.18	2482.30	4262.757	2917.39

4.2 Discussion of Results

From the analysis above, Fig 4.1-Fig 4.8 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.

Also, from the analysis above, Tables 4.1 and 4.2 show the linear correlation between the considered oil spills of the four oil companiesin 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 4.3 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. Tables4.4 and 4.5 show estimation of model parameters and error metrics (forecast accuracy measures) respectively. The values of the error metrics, in terms of MAE, RMSE, and MSE, show that the LOG-ARIMAX model gives better forecasting accuracy than the traditional ARIMAX model.

5.0 SUMMARY AND CONCLUSION

5.1 Summary

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 thesis, 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, Ω_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.

5.2 Conclusion

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. The results of the analysis show that the hybridization of Logarithm and ARIMAX (LOG-ARIMAX) as propounded in this work is more robust, efficient, sufficient and reliable in forecasting long-memory data characterized by heavy tailed traits.

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

Akouemo H.N., Povinelli R.J. (2017). Data Improving In Time Series Using ARX and ANN Models IEEE Transactions on Power Systems, 32 (5), pp. 3352-3359

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.

Chen, J. (2019). *Autoregressive Integrated Moving Average (ARIMA)*. Advanced Technical Analysis concept.

Farhana I. And Monzur A. I. (2020) Use of Teleconnections to Predict Western Australian Seasonal Rainfall Using ARIMAX Model 2020. Hydrology 7(3):52

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

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(4): 29-40. doi:10.13189/ms.2019.070705.

Nimish, J.Shraddha, S.B. Rajanarayan, P. (2021) Performance Comparison of Two Statistical Parametric Methods for Outlier Detection and Correction. IFAC-PapersOnLine, Volume 54, Issue 16, 168-174

Ugoh, C. I., Uzuke, C. A. and Ugoh, D.O. (2021). Application of ARIMAX Model on Forecasting Nigeria's GDP. American Journal of Theoretical and Applied Statistics10(5):216

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

Lang, J.H., Peltonen, T. A., and Sarlin, P. (2017). *A framework for early-warning modeling withan application to banks*.Working Paper Series. European Central Bank. No 2182.