Proceedings of the 1st International Conference on Statistical Sciences and Data Analytics

HYBRID MODELING APPROACH FOR ANALYZING LONG-MEMORY ENVIRONMENTAL DATA

Kennedy I. Ekerikevwe[1] and Tayo K. Oyeleke[2]

[1]Department of Statistics, Delta State Polytechnic, Otefe-Oghara, Delta State.
[2]Statistician, National Bureau of Statistics, Abuja

*Corresponding author Email: kennedyekerikevwe@gmail.com

ABSTRACT

Traditionally Autoregressive Integrated Moving Average with covariates X (ARIMAX) model is rarely applied to climate change and environmental data, which are the most cognate data with associated exogenous variables and are typically characterised by kurtosis, skewness, outliers, long memory (high frequency), and large fluctuation series. In order to neutralise this model for a better and enhanced prediction of the system, a distributional form of the error term that is robust and sufficient in capturing and accommodating both the external covariate(s) and long memory (high frequency) data is required. This study is to develop a hybrid Autoregressive Integrated Moving Average with covariate X by taking the logarithm of Arimax (Log-ARIMAX) model for long memory (high frequency) time series data that is coupled with external time-varying covariate(s) with heavy tailed distributional lognormal form of a residual structure. The Generalized Linear Method (GLM) was used to estimate the parameters of the proposed model. The results of the analysis show MAE =49.82, RMSE = 49.82 and MSE = 2482.30); these imply that the hybrid model has a better forecasting strength than the classical models.

Keywords: Hybridisation, time Series, Long Memory, Analysis, Performance, Forecast.

1. INTRODUCTION

The sequence of observations obtained through repeated measurements over time is termed time series, especially at equal intervals (Wiwik, 2017). The time spans can be hourly, day to day, week by week, month to month and so on. Time series include things like the annual number of students admitted to a school, steel production, daily stock prices on a stock exchange, monthly rainfall in a given location, etc. despite the fact that, time series are some of the time explored simply due to intrigue ever, the significant aim in examination of tome series align with concentration for connection between factors through time with the end goal of determining future upsides of the series. (Ekerikevwe and Odior, 2020). According to Yucesan (2018), hybridization is a combination of different mathematical models to obtain new and better model (s). Hybridization is a model that attempts to remedy the shortcomings of existing model (s). Olatayo and Ekerikevwe (2022), opined that 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 (Ling *et al.*, 2019). As a result, time series play a significant role in an autoregressive model and are utilized in

accordance with the circumstance and desired outcome. In the work of Uyodhu and Didi (2016), 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 moving average model order q can be obtained specifically by checking the ACF graph of the time series. Stationarity is essentially a random cycle that it characteristics remain unchanged with movement over a period. It is defined as having a mean and variance that remain constant throughout the time series. A period series should be fixed for it to make great expectation. In Lang *et al.*, (2017), by looking at the autocorrelation function, stationarity can be determined. Essentially, autocorrelation is to measure how similar two observations are in relation to the time lag between them. When analyzing time series data, possibility of encountering a situation known as long memory is evident. Long-range dependence is another name for this phenomenon. It basically refers to the statistical degree of dependence between two-time series points. They are often observed in large volumes. Ugoh, *et al.*, (2021), simply put forecasting as the process of utilizing previous data values to predict future data values. If you want to make predictions based on accurate information, stationarity of time series is required. Forecasting can be used in numerous fields worldwide. In the business and financial world, forecasting is often used by businesses to predict their profits or expenses. Stock prices can also be predicted using forecasting! It can be seen in environmental issues like global warming. Time series and forecasting are also heavily used in the economics field to predict how societies will behave. These are just a few of the many real-world applications of time series and forecasting (Ekerikevwe and Olatayo, 2022).

Domenico *et al.*, (2021) proposed a straightforward econometric model that may be useful for predicting COVID-2019's spread. We applied the Auto Regressive Integrated Moving Average (ARIMA) model to the epidemiological data from Johns Hopkins in order to forecast the trend in the prevalence and incidence of COVID-2019. For purposes of comparison or perspective, case definition and data collection must be maintained in real time. Using Excel 2019 and the daily prevalence statistics for COVID-2019 from January 20, 2020, to February 10, 2020, which were downloaded from the university's official website, a time-series database was created.22 number judgements made up the dataset, which was subjected to the ARIMA model. Their results show that COVID-2019 has an increasing general prevalence and a plateauing epidemic. A non-steady increase in the number of confirmed cases was evident when comparing instances from one day to occurrences from the previous day using the formula (Xn-Xn-1). To minimize bias, the incidence of new COVID-2019 confirmed cases was assessed using descriptive data analysis.

As per Munir and Mayfield (2021), the straight auto-backward coordinated moving normal (ARIMA) outflanked its nonlinear partner with regards to execution. All things considered, Rachasak *et al.*, (2022), using the records of 3685 patients admitted to Thailand's College field medical clinic, established a multiple times series model for estimating Coronavirus cases. The autoregressive coordinated moving normal (ARIMA), its extensions to exogenous components (ARIMAX), and affiliation rule mining (ARM) are the methodologies used for the review. The study showed that the ARIMAX model has the potential to forecast the number of COVID-19 cases by combining the most related prognostic parameters discovered by ARM methods into the ARIMA model. This might be used to plan for pandemics and manage hospital resources in the best way possible. It was evident from the reviewed and available literature that the ARIMAX model has never been modified or extended to capture stabilize means of the time series (via switching of cyclical traits), seasonality signals, if any, when characterized with high or long-memory series by heavy-tailed distributional traits, and outliers in the subjected series. Long-memory, fluctuations, cyclical, and seasonal characteristics typically originate

from the series itself or an additional exogenous variable or variables. To mitigate the threat posed by long-memory traits, however, a neutralize function is required. To put it another way, the natural logarithm would be added to the ARIMAX to alter the Gaussian random noises attached to and assumed by both the ARIMA and the regression error term to the exogenous. Having said that, in order to carve out the transformation function, it would be necessary to switch from the general linear model technique to the generalized linear model because the Gaussian distributional noise would be altered.

In a nutshell, the ARIMAX model would be transformed into Log-ARIMAX strictly negative or positive continuous type observations by employing a Log-Gaussian distributional general random noise. The driven function for parameters estimation through the Generalized Linear Model with adjusted model performances (AIC, BIC, HQNIN, etc.) would be the transform function that would be the necessity component of the non-linear Log-ARIMAX and forecast measure indexes that are mostly accurate (such as MSE, MAD, RMSE and MAE). In addition, after critically examining the majority of the applications of ARIMAX, it was determined that the exogenous variable(s) regarded as one of GDP, per capita income, macroeconomic variables, consumer price index, unemployment rate, etc., was the predominant exogenous variable(s) on either daily, weekly, monthly, quarterly, or annual records of financial returns. The ARIMAX model has been used infrequently to any of the climate change agents, especially temperature, which is the most closely linked agent with associated exogenous factors like intense dry spells, torrential rains, erosions, and flooding; more frost and heat). Since temperature and other agents are regarded as rare events, extreme values, and large fluctuation series, ARIMAX was only used sparingly or not at all. Being viewed as vacillation series makes the Gaussian arbitrary commotion (repetitive sound) detonates asymptotically. Environmental agent series, which are typically influenced by kurtosis, skewness, outliers, high frequency, and lengthy memory, are comparable to climate change agents' series. Logarithmic-ARIMAX is a rival time series model for analyzing and modeling oil spillage temperature and environmental data without model explosion and with robustness. This review, subsequently, proposed a robust ARIMAX model to manage environmental change and ecological data that are normally impacted by kurtosis, skewness, exceptions, high recurrence or long memory observational time series data.

2. MATERIALS AND METHODS

2.1 Autoregressive Integrated Moving Average with Covariates

According to Yang and Wang (2018) and Ekerikevwe and Oyeleke (2025). Autoregressive Integrated Moving Average with Covariates "X" is known as ARIMAX. This model is an improved version of the ARMA because it 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:

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

Where x_t ... are the p-lagged period of the exogenous covariates (x_{t-p}) with errors that are independentand identical in distribution having zero mean, variance (σ^2) and covariance of zero.

Otherwise,

$$\varphi_p(B)Y_t = \varphi_0 + \varphi(B)x_t + \theta_q(B)\varepsilon_t \quad (2)$$

$$\varphi_p(B)\nabla^d Y_t = \varphi_0 + \varphi(B)x_t + \theta_q(B)\varepsilon_t \quad (3)$$

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.

$$\varphi(B)x_t = [\varphi_0 + \varphi_1 B + \varphi_2 B^2 + \cdots\cdots]X_t \quad (4)$$

$\varphi_1, \varphi_2, \ldots$ in eq. (4) 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.

$$\varphi(B)x_t = \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t \quad (5)$$

where;

$$\eta_h(B) = \eta_0 + \eta_1 B^1 + \eta_2 B^2 + \cdots\cdots + \eta_h B^h; \quad (6)$$

$$\lambda_r(B) = 1 - \lambda_1 B^1 - \lambda_2 B^2 - \cdots\cdots - \lambda_r B^r; \quad (7)$$

$h \rightarrow$The +1 number of items of the independent variables

$r \rightarrow$The +1 number of items of the dependent variables

$b \rightarrow$The impulse weight or dead time

So, $Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)}$ (ARMAX) (8)

where;

$$\sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)}X_t = (\sum_{i=0}^{\infty} \varphi(B)x_t)B^b = \sum_{i=0}^{\infty}(\varphi_i B^i)B^b \quad (9)$$

$$= \varphi_0 B^b + \varphi_1 B^{b+1} + \varphi_2 B^{b+2} + \varphi_3 B^{b+3} + \cdots\cdots \quad (10)$$

Equation (3.24) could be written in terms of Integrated, that is, in terms of ARIMAX as;

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\nabla^d \varphi_p(B)} \quad \text{(ARIMAX)} \tag{11}$$

2.2 Logarithmic Autoregressive Integrated Moving Average-X (Log-ARIMAX)

In long-memory (highly frequency) observational series, ARMAX or ARIMAX might unable to dissolve the characterized long-memory in both the exogenous covariates and series or in any of the two observational. In category of such long-memory data are high valued financial returns, climate indexes recorded data, sea wave measurements, evolutional data, inflation, GDP etc. However, the stochastic disturbance would be less in power to dissolve high constant flexibility. An ideal distributional form of the stochastic disturbance (ε_t) via the lognormal variate shall be introduced.

The distributional form of (ε_t)is then given as

$$f(y_t) = \frac{1}{y_t \sigma\sqrt{2\pi}} exp\left[-\left(\frac{(ln(y_t))^2}{2\sigma^2}\right)\right] \quad y_t > 0 \tag{12}$$

or

$$f(\varepsilon_t) = \frac{1}{\varepsilon_t \sigma\sqrt{2\pi}} exp\left[-\left(\frac{(ln(\varepsilon_t))^2}{2\sigma^2}\right)\right] \quad \varepsilon_t > 0 \tag{13}$$

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

$$\text{With } y_t \sim \varepsilon_t \sim N\left[exp\left(\frac{\sigma^2}{2}\right), exp(2\sigma^2) - exp(\sigma^2)\right] \tag{14}$$

Therefore, for logarithmic ARMAX, we have;

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\varphi_p(B)} \sim N\left[exp\left(\frac{\sigma^2}{2}\right), exp(2\sigma^2) - exp(\sigma^2)\right] \tag{15}$$

And for logarithmic ARIMAX, we have;

$$Y_t = \sum_{j=1}^{n} \frac{\eta_h(B)B^b}{\lambda_r(B)} X_t + \frac{\theta_q(B)\varepsilon_t}{\nabla^d \varphi_p(B)} \sim N\left[exp\left(\frac{\sigma^2}{2}\right), exp(2\sigma^2) - exp(\sigma^2)\right] \tag{16}$$

Where;

$$\theta_1^2 = \frac{1+\theta_2 \sum_{i=1}^{n} y_{t-2} y_{t-1}}{\sum_{i=1}^{n} y_{t-1}^2} \tag{17}$$

$$\theta_i = \sqrt{\frac{1+\theta_2 \sum_{i=1}^{n} y_{t-2} y_{t-1}}{\sum_{i=1}^{n} y_{t-1}^2}} \tag{18}$$

Equation 16 is our proposed hybridization equation to be used in the analysis of the environmental data, while equations 17 and 18 are the estimator of the parameter and the square root of the estimator respectively.

3. RESULTS AND DISCUSSION

The data used for the validation of our hybrid model are oil spillage data obtained from four oil and gas organisations from January 2005 to December 2024. 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' forecasting ability.

Table 1: Results for ARIMAX and LOG-ARIMAX Models Selection

Ticker	Model Type	Selected Model	AIC
DCNE	ARIMAX	(0,1,1)	692.36
DCNE	LOG-ARIMAX	(0,1,1)	680.95
DCNE*	ARIMAX	(1,1,0)	469.66
DCNE*	LOG-ARIMAX	(0,1,3)	465.77

Table 2: Estimation of Model Parameters

ESTIMATES	ARIMAX	LOG-ARIMAX	ARIMAX*	LOG-ARIMAX*
□	-	0.3303	-	0.3711
AR (1)	-	-	-	-
AR (2)	-	-	-	-
AR (3)	-	-	-	-
MA (1)	0.0025	0.1153	-	0.1656
MA (2)	-	-	-	0.0479
MA (3)	-	-	-	-

Table 3: Error Metrics (Forecast Accuracy Measures)

	Metrics Type		
Ticker	MAE	RMSE	MSE
ARIMAX	22.7411	30.6196	937.5589
LOG-ARIMAX	20.7456	27.8590	776.1221
ARIMAX*	17.3169	24.4696	598.7617
LOG-ARIMAX*	24.5394	33.9308	1151.2960

Table 4: Diebold-Mariano Test for Comparing Models

Ticker	Test Type	p – value
ARIMAX	**DM**	< 0.0685
LOG-ARIMAX	**DM**	< 0.0001
ARIMAX *	**DM**	< 0.0708
LOG-ARIMAX *	**DM**	< 0.0003

Table 5: Results Summary

Model	ARIMAX	LOG-ARIMAX	ARIMAX *	LOG-ARIMAX *
L.COR	0.911	- 0.249	+ 0.811	*- 0.068
AIC	692.36	680.95	469.66	465.77
MAE	44.77	49.82	53.44	42.60
RMSE	56.55	49.89	65.29	54.01
MSE	3198.18	2482.30	4262.757	2917.39

3.1 Discussion of Results

Table 1 shows that the new ARIMAX model has the least AIC in the two time horizon as compared to the classical ARIMAX model. This implies that the new ARIMAX model has a better forecasting strength and accuracy as compare to that of ARIMAX model. Tables 2 and 3 show estimation of model parameters and error metrics (forecast accuracy measures) respectively. The values of the forecast

accuracy metrics, in terms of MAE, RMSE, and MSE, show that the new ARIMAX model gives better predictive accuracy than the old ARIMAX model.

4. CONCLUSION

With reference to the objective of this paper, it is empirically evident that ARIMAX model with an exogenous variable know as logarithmic ARIMAX performed creditably well in all cases and scenarios as outlined in the above analysis. This emphasizes that, when improving the in – sample forecasting accuracy of environmental data 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 firms in Nigeria were use as possible exogenous variable or as public information and for the purpose of model validation. Evidently, the Diebold and Mariano test of accuracy is dependent on the AIC of the candidate models. However, in most cases smaller AIC values turn to minimize the considered error metrics (i.e., MAE =49.82, RMSE = 49.82 and MSE = 2482.30) and vice versa. This is evident throughout the results. 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 robust time series model known as Logarithmic-ARIMAX model to model the external covariate(s) of long memory types of environmental data. The results of the analysis show that the new ARIMAX (Logarithmic-ARIMAX) as proposed in this work is useful, more robust, and efficient in forecasting long-memory environmental data. The results of these study are collaborate earlier researches conducted by (Olatayo and Taiwo, 2016; Ekerikevwe and Oyeleke, 2025).

REFERENCES

Domenico B, Marta G, Lazzaro V, Silvia A, Massimo C (2021). Application of the ARIMA Model On The Covid-2019 Epidemic Dataset. Elsevier Volume 29, April 2020, 105340

Ekerikevwe, K. I. and Odior, K. A. (2020) "Seasonally Adjusted Trend Forecast of Road Traffic Crashes along Lagos-Benin Highway of Nigeria" International Journal of Maritime and Interdisciplinary Researches (IJMIR), Published by Nigeria Maritime University, Vol1 (1) Page16-29.

Ekerikevwe, K. I. and Olatayo, T. O. (2022) "Parameter Estimation of Log-Arimax Model via Generalized Linear Method of Exponential Form" by *ABACUS (The Journal of Mathematical Association of Nigeria (MAN), Mathematics Sciences Series),* Vol 49, No 2, July, Pg 162-173)

Ekerikevwe, K. I. and Oyeleke, K. T. (2022). Transformed-Arimax Model for Heavy Tailed Distributions, International Journal of Development Mathematics Vol 2 Issue 2 | 314 – 327. https://doi.org/10.62054/ijdm/0202.18

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

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.

Munir, S., & Mayfield, M. (2021).Application of density plots and time series modelling to the analysis of nitrogen dioxides measured by low-cost and reference sensors in urban areas. Nitrogen, 2(2), 167-195.

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.

Olatayo, T.O. and Ekerikevwe, K. I. (2022) "Performance Measures For Evaluating The Accuracy Of Time Series Hybrid Model Using High Frequency Data" Published by *Britain International of Exact Sciences (BIoEx) Journal. ISSN: 2686-1208 (Online), 2686-1216 (Print,)* Vol. 4, No. 3, September, Page: 244- 259, DOI: https://doi.org/10.33258/bioex.v4i3.760

Rachasak, S., Warin, K., Amasiri, W. *et al.*, (2022) Forecasting COVID-19 cases using time series modeling and association rule mining. *BMC Med Res Methodol* 22, 281.

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

Uyodhu, A.V. Didi, E. (2016). Autoregressive Integrated Moving Average with Exogenous Variable (ARIMAX) Model for Nigerian Non-Oil Export. *European Journal of Business and Management*, *8*(36), www.iiste.orgISSN 2222-1905 (Paper) ISSN 2222-2839.

Wiwik A., Kuntoro B. A., Sumaryanto, F. M (2017). The Performance of ARIMAX Model and Vector Autoregressive (VAR) Model in Forecasting Strategic Commodity Price in Indonesia. Procedia Computer Science, Volume 124, Pages 189-196

Yucesan, M., Gul, M.,& Celik, E. (2018).Performance Comparison between ARIMAX, ANN and ARIMAX-ANN Hybridization in Sales Forecasting for Furniture Industry.DrvnaIndustrija, 69 (4): 357-370.