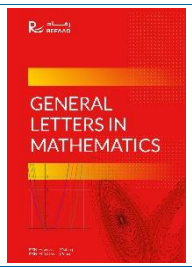
Available online at www.refaad.com
Gen. Lett. Math., 13(1) (2023), 5- 17
Research Article

General Letters in Mathematics (GLM)

Journal Homepage: https://www.refaad.com/Journal/Index/1

ISSN: 2519-9277 (Online) 2519-9269 (Print)



A hybrid Modeling and Forecasting of Carbon dioxide Emissions in Tanzania

Twahil Hemed Shakiru[a,*], Xiaohui Liu[b], and Qing Liu [c]

[a] PhD candidate School of Statistics, Jiangxi University of Finance and Economics, China.
[b] School of Statistics, Jiangxi University of Finance and Economics, China.
[c] School of Statistics, Jiangxi University of Finance and Economics, China.
Email: [a] hemedtwahil@gmail.com

Abstract

Carbon dioxide (CO_2) emissions is among of global environmental pollutants contributing to climate change. The current study aims to create an Autoregressive Integrated Moving Average with external factors (ARIMAX) model to predict CO_2 emissions in Tanzania. In this study, an Autoregressive Integrated Moving Average (ARIMA) model is first created. Then it is combined with the influencing factors using multiple linear regression to fit an ARIMAX model. There is a high possibility of both under- and overestimation because the ARIMAX employing multiple linear regression (ARIMA-MLR) model only generates mean forecasts. A hybrid ARIMA-Quantile Regression (ARIMA-QR) is created to forecast high and low quantiles. The ARIMA-QR model mainly predicts the quantiles instead of extrapolating from the mean point of the ARIMA-MLR model, which bases more on the assumption of normality. The established ARIMA-MLR and ARIMA-QR were used to forecast and model annual data on CO_2 emissions in Tanzania. The findings reveal that both ARIMA-MLR and ARIMA-QR models outperform the traditional ARIMA model in terms of forecasting accuracy with the least mean absolute percentage error (MAPE) and root mean square error (RMSE).

Keywords: CO_2 emissions, Hybrid ARIMA, Quantile Regression.
2020 MSC: 62J05,62J20,62P20.

1. Introduction

Carbon dioxide (CO_2) is among environmental pollutants contributing to climate change, accounting for 58.8% of all greenhouse gases (GHG). The single main source of carbon dioxide and total GHG emissions is the burning of fossil fuels. Due to the recent global economic expansion, their impact has grown at the quickest rate of any critical source since 1970.

African countries must simultaneously address climate change's broad and complicated effects while addressing their need for economic development without becoming further dependent on fossil fuels or wasteful technologies [35]. Most nations have seen rising pollution levels due to financial incentives to expand industrial output, and if current development patterns continue, this trend is likely to continue [36].

* Corresponding author
Email addresses: hemedtwahil@gmail.com (Twahil Hemed Shakiru).
doi: https://doi.org/10.31559/glm2023.13.1.2
Received: 18 Jan 2023 Revised: 14 Feb 2023 Accepted: 25 Feb 2023

Tanzania is a prime example of a developing country that has emerged since the 21st century. In addition to experiencing an incredible economic miracle, Tanzania has also seen a rapid increase in energy use and carbon emissions. [32] assert that Tanzania is one of the world's most politically and economically vibrant nations. At the beginning of the twenty-first century, Tanzania, like all other East African nations, was agrarian and pursued a range of revival methods for their economies.

In an era with highly developed technology pathways for lowering emissions, many wealthy countries have achieved or will soon reach their emission peaks [23]. Within the next generation, developing countries and even less developed countries, particularly Tanzania, will become the main theatre of conflict for future emission reductions [17]. Although Tanzania has no recognized fossil fuel deposits, its numerous energy sources analysis demonstrates that petroleum-based fuels are the main drivers of its domestic economy. The average annual usage of fossil fuels in Tanzania is about 2.3 million tons. Since fossil fuels emit massive volumes of emissions into the atmosphere when burned, it causes air pollution. Figure (1) below shows the sources of carbon dioxide emissions in Tanzania by source. Carbon dioxide is the main significant component of greenhouse gases.

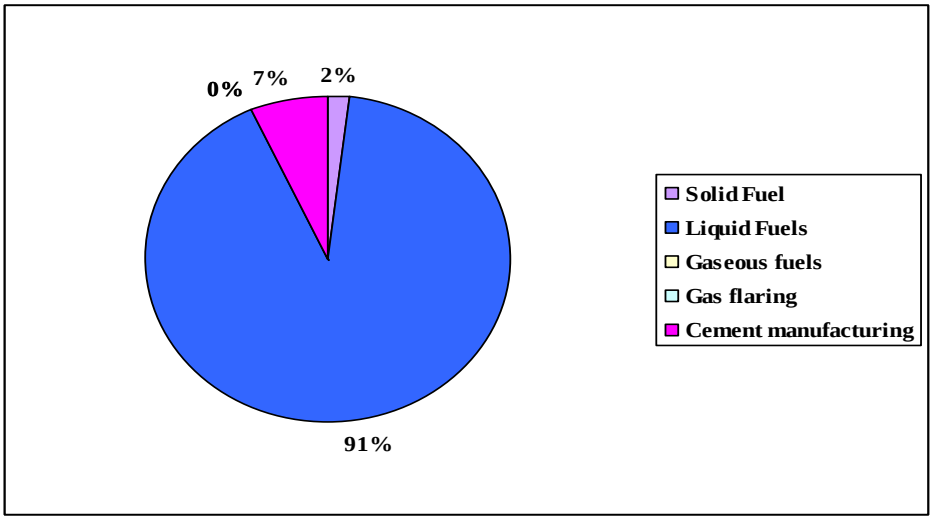


Figure (1): Carbon Emissions by Source in Tanzania (in thousands of metric tons of CO_2)
Source: *World Resources Institute, Earth Trends*

Based on Figure (1), it is evident that fossil fuels are Tanzania's primary source of energy and carbon emissions. Since Tanzania now relies more on fossil fuels as its primary energy source, which urgently has to be replaced by new, cutting-edge energy sources. However, due to financial limitations and technological lag, underdeveloped nations like Tanzania cannot fast develop sustainable energy sources [30]. Thus, the study and forecasting of CO_2 emissions using external variables like electricity consumption and economic growth are essential to the clean energy economy in Tanzania.

In general, quantitative methods can be used to forecast CO_2 emissions. Two examples of quantitative approaches are building mathematical models containing causative factors or extrapolating historical environmental data. The quantitative techniques are divided into hybrid, causal, and time series models. The causal models depend on the causal relationships among the factors influencing CO_2 emissions. Time series models mostly depend on the previous environmental observations collected systematically to forecast future pollution. Hybrid models incorporate both time series and causal models. Proper forecasting is necessary for precise investment planning in energy manufacturing and supply, environmental management, and economic growth. Numerous research has used causal models to analyze and predict energy consumption and CO_2 emissions [10], [28]. The disadvantage of causal models is that they rely on the reliability and accuracy of data on explanatory variables across the forecasting period, requiring additional data collection and computation efforts. Univariate time-series analysis, like that used in ARIMA, is another modeling strategy that requires the variable's past data to forecast its future behavior. The univariate ARIMA analysis proposed by [7] has been widely employed in numerous fields to provide accurate forecasting conclusions; many historical observations have been required. Artificial neural networks (ANN) [4]; [19], fuzzy regression [25],[5], and other intelligent nonlinear forecasting techniques have been used to anticipate energy demand more accurately. However, the standard approach struggles to attain high precision due to the nonlinear nature of environmental data. As a result, several researchers started researching intelligent models to improve forecast accuracy. However, univariate and casual models may have limits. Thus, more researchers are beginning to enhance forecasts by combining two or more models into a hybrid model to predict energy consumption, pricing, and other concerns [27], [31]. Forecasting CO_2 emissions

must consider the influencing elements to improve prediction accuracy. The ultimate goal of a forecasting model is to place an order based on the forecast's outcome.

A certain amount of forecast uncertainty always affects the judgments, even when the forecasting model is successful [12]. Since CO_2 is the greenhouse gas shown to have the most significant effects on environmental problems, forecasting CO_2 emissions has become a global concern. Forecasting CO_2 emissions is another crucial step in raising public awareness of environmental issues. Keeping these concerns in mind, the main objective of this study is to develop a hybrid time series model with external variables utilizing a hybrid ARIMA- Q.R. model to forecast CO_2 emissions. This study's remaining sections are organized as follows: Brief methodology details are provided in Section 2. The application of the created model is described in Section 3. The results and discussion are covered in Section 4. Lastly, the conclusion is presented in Section 5.

2. Methodology

2.1 ARIMA Model

Univariate time series data with equal spacing are analyzed and predicted using the ARIMA model. It predicts by linearly mixing the initial values and errors in a response time series. The three steps that make up the analysis performed by the ARIMA technique are identification, estimation and diagnostic checking, and forecasting. These three steps match those listed by Box and Jenkins. Statistics time series are described by traditional Box-Jenkins models. Therefore, before attempting to determine a Box-Jenkins model, initially, we convert the time series into a stationary time series by differencing the series. The ARIMA model is commonly written as (p, d, q), and it is created by combining three building components, namely, p for autoregressive (A.R.), d for integration order term (I), and q for moving average (M.A.) to simulate the serial correlation in the disturbance term.

The parameterization of Autoregressive (A.R. (p)) of order p is written in equation (1):

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \varepsilon_t \tag{1}$$

The parameterization of the moving average (M.A. (q)) is expressed in equation (2).

$x_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_h \varepsilon_{t-h}$

$x_t = \sum_{j=0}^{h} \beta_j \varepsilon_{t-j}$

Where, $\beta_0 = 1 \; and \; \varepsilon_t$ is white noise i.e

$E(\varepsilon_t) = 0$

$E(\varepsilon_t \varepsilon_s) = \sigma^2 \; for \; t = s$

$$E(\varepsilon_t \varepsilon_s) = 0 \; for \; t \neq s \tag{2}$$

Thus, the general ARMA (p, q) is written as:

$$x_t + p_1 x_{t-1} + \cdots + p_p x_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_h \varepsilon_{t-h} \tag{3}$$

Hence, the ARIMA model of order (p, d, q) can be used to postulate the backward shift operator as follows:

$\phi(Z)\Delta^d x_t = \varphi + \theta(Z)\varepsilon_t$

Where, $\phi(Z) = 1 - \phi_1 Z - \phi_2 Z^2 - \cdots - \phi_p Z^p$

$$\theta(Z) = 1 - \theta_1 Z - \theta_2 Z^2 - \cdots - \theta_q Z^q \tag{4}$$

The symbol denotes the time series is x_t; the backward operator is denoted by the letter Z, d shows the number of differences; lastly, the symbols ε_t, and ε_{t-1} represents independent disturbance terms. It is assumed that the series ε_t is a white noise process, and the polynomials $\phi(Z)$ and $\theta(Z)$ in Z are of orders p and q, respectively. The polynomial roots of $\phi(Z)$ and $\theta(Z)$ always lie outside the unit circle.

2.2 ARIMAX

A time series Y_t can be explained by external factors employing a linear regression model rather than just modeling it with a collection of lagged values. It is expected that errors in linear regression analysis are random. However, the autocorrelation property of the data frequently violates it in the case of time series [3]. Therefore, ARIMAX corrects this problem and explains autocorrelation. The ARIMA model is enhanced to become an ARIMA model with explanatory variables known as ARIMAX (p, d, q) X, where X denotes the vector of explanatory variables. It is possible to express the general ARIMAX model as follows:

$$y_t = \gamma_0 + \gamma_1 x_{1,t} + \gamma_2 x_{2,t} + \cdots + \gamma_s x_{s,t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d}\varepsilon_t \tag{5}$$

Where $x_{1,t}\ x_{2,t} \ldots, x_{s,t}$ are the explanatory variables/ external variables with regression coefficients denoted as $\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_s$.

2.3 Quantile Regression

Researchers [24] pointed out that point forecasting is weak support for decision-making. The efficiency of point forecasting decreases in the presence of severe and undefined circumstances. [9] discussed the significance of delivering interval forecasts, including assessing future uncertainty, developing various tactics for a range of potential outcomes, comparing forecasts, and investigating various scenarios by considering several assumptions. Estimating the prediction interval is crucial for anticipating environmental data and wise decision-making. In most older studies, this issue was solved by converting point forecasts into interval forecasts. The confidence intervals for 10% and 90% were proposed by [15] as $x_t \pm 1.282\sigma_t$. The prediction's confidence interval range is expected to include the mean response for the model's specified values of the explanatory variables. However, the range of a model's prediction interval is expected to contain the average predicted response for the provided value of the explanatory variables. [16] noted the computation of forecasting intervals by employing theoretical mathematical equations that are substantially similar. Equations (6) and (7) can be used to express the forecasting interval formula at $(1-\alpha)100\%$ for the value k phases ahead:

$$\widehat{x_t} \pm Z\alpha_{/2}\sqrt{var[\varepsilon_n(k)]} \tag{6}$$

Where $\varepsilon_n(k)$ denotes the forecast disturbance term. Another formula was suggested by [37].

$$\widehat{x_t} \pm Z\alpha_{/2}\hat{\sigma}\sqrt{1+{}^{1}/_{n}} \tag{7}$$

The term $\hat{\sigma}$ represents the standard deviation of the disturbance term.

These techniques are only applicable if the residuals are normal and uncorrelated. In place of a point forecast based on prediction intervals, [33] recommended using quantiles in forecasting. Under average conditional function $E(Y|X)$, the traditional mean regression analysis illustrates how the average dependent variable (y) evolves with the explanatory vector variable (x). [22] expanded the traditional mean regression analysis to incorporate conditional quantiles of the dependent variables. Quantile regression (QR) takes conditional quantiles function$Q_\tau(Y|X)$ into account, in contrast, to the mean regression model. The link between the response and explanatory variables is thoroughly viewed using the quantile regression technique.

The sum of squared errors is minimized to determine the regression coefficients (γ) in a conditional mean model [6]:

$$\min_{\gamma\in\Re} \sum_{i=1}^{n}(y_i - x_i'\gamma)^2 \tag{8}$$

The regression coefficients in a conditional median model are determined by minimizing the sum of absolute deviations:

$$\min_{\gamma\in\Re} \sum_{i=1}^{n}|y_i - x_i'\gamma| \tag{9}$$

The quantile regression in equation (10) reduces the objective function for the quantile and then applies a 50 percent median regression to each remaining quantile [21]:

$$\min_{\gamma\in\Re} \sum_{i:y_i\geq x_i'\gamma}^{n}\tau|y_i - x_i'\gamma| + \sum_{i:y_i<x_i'\gamma}^{n}(\tau-1)\,|y_i - x_i'\gamma_\tau| \tag{10}$$

where $\tau \in (0,1)$ is the existing quantile, the symbol γ_τ represents τth regression quantile. Median regression is well described when $\tau = 0.5$. We can analyze the properties of the dependent variable beyond the Gaussian distribution's mean using the QR technique. Consequently, using QR one may create prediction intervals [26]. Compared to looking solely at the dependent variable's average, different inferences may be drawn by studying the properties of the dependent variable at various quantiles by considering its explanatory variables.

2.4 Performance Assessment

The accuracy of the forecasting performance target is to outperform a targeted model and continuously enhance better prediction procedures [14]. They recommended a better method to scale the forecasting performance objective using a straightforward naive forecasting technique as a targeted model. The suggested model was compared against the naive model's forecast accuracy as a baseline. Furthermore, [13] suggested using forecast value assessed (FVA) analysis to contrast the prediction approaches. He claims that FVA analysis can be

applied to find and remove procedures that aren't improving the forecast while producing improved projected outcomes.

The mean errors of the out-of-sample data set may be employed to gauge the model's forecast accuracy. Performance measurements employed in this research include root mean square error (RMSE) and MAPE. This provides the foundation for prediction accuracy using the ARIMA and the developed hybrid models. Equation (11) and (12) represents MAPE and RMSE, respectively

$$MAPE = {}^{1}/_{n} \sum_{t=1}^{n} \left| \frac{X_t - \widehat{X_t}}{X_t} \right| \tag{11}$$

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(X_t - \widehat{X_t})^2}{n}} \tag{12}$$

Where $\widehat{X_t}$ and n are forecasted time series and the number of observations, respectively.

2.5 Description of Data and Variables

The research adopted for this work is a quantitative research design, and data for this study was collected through reliable secondary sources like world bank publications. The key variables shown in Table (1) are CO_2 emissions, electricity generation, consumption and GDP Per Capita, urban population, and labor force. In this study, the annual data from 1989 to 2020 was used.

Table (1): Variable measurement

Variables	Measurements (unit)
CO_2 emissions	Dollar (Us)
GDP Per Capita	Megatons (Mt)
Electricity generation	Gigawatt hour (GWh)
Consumption	Gigawatt hour (GWh)
Urban population	Percentage (% of the total population)
Labor force	Count

Source: Author

3. Application of ARIMAX Models

Two steps can be utilized while performing the ARIMAX modeling:
Firstly, establishing an ARIMA model, and secondly, using MLR and Q.R. to incorporate external factors and the expected values of the ARIMA model as input variables. The residuals are then tested for white noise using parameter significance testing and diagnosis.

3.1 Developing of ARIMA Model

Identifying a model entails examining the characteristics of ACF and PACF values or their graphs. The first stage in the ARIMAX process is establishing an ARIMA model using the provided data set. The complete dataset has been divided into two parts to facilitate model comparison: a training part with 70% of the sample size and a validation part with 30% of the sample size. Figures (2a and 2b) display the annual CO_2 emissions autocorrelation and partial autocorrelation plots, respectively. It can be seen that the ACF plot exhibits a geometric decline and one significant lag in the PACF plot. This shows that the time series depends only on the previous observation.

Furthermore, the Phillips-Perron method shows that the time series is stationary after the first differencing**.** The training part autocorrelation and partial autocorrelation were examined to identify the order of the ARIMA model. Figures (2a and 2b) show that the ARIMA (1,0,0) model is the most effective. STATA and SPSS version 23 produced the ARIMA model's outcomes. Table (2) displays the results of evaluating the performance of the ARIMA model using MAPE and RMSE for both in-sample (training) and out-of-sample (testing) data sets. The ARIMA model's results demonstrate that the model's performance is the same during training and testing.

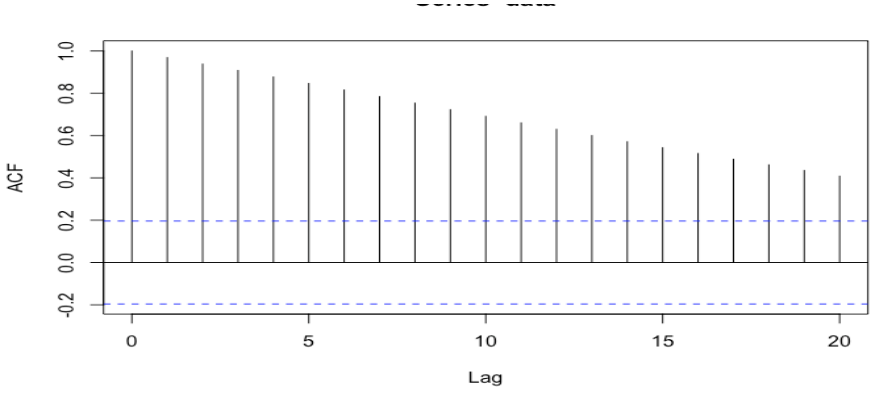


Figure (2 a): ACF Plot

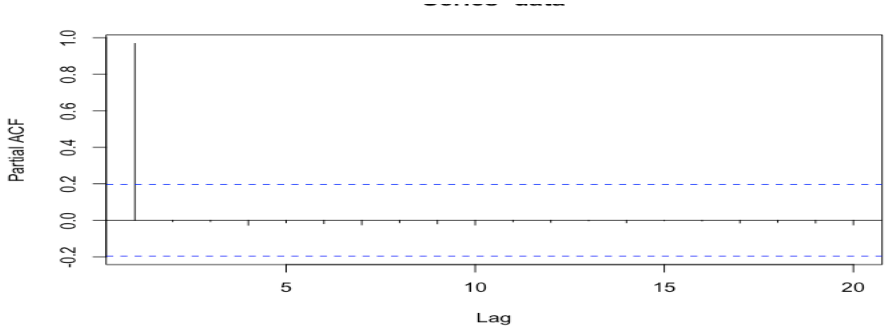


Figure (2 b): PACF Plot

Table (2): Performance of ARIMA for training and test data set of CO_2 emission

Performance measure	Training (in-sample)	Test (out of sample)
MAPE	25.31	22.71
RMSE	24.43	19.86

3.2 ARIMA-MLR Model

The next step of our methodology is to apply a multiple linear regression model with CO_2 emission as a response variable and predicted ARIMA CO_2 emission and other external variables served as explanatory variables. The external factors/variables include labor force, electricity consumption, GDP per capita, and urban population. The Variance Inflation Factor (VIF) is utilized as a post-diagnostic method to assess the effects of multicollinearity in multiple linear regression analysis. Multicollinearity may cause misleading or counterfeit results when an expert or analyst examines how well each explanatory variable can most efficiently be used to forecast or know the response variable in the analytical model. Unfortunately, even under high multicollinearity, the OLS assumptions do not violate. The variance inflation factors greater than 5 indicate the presence of severe multicollinearity within the independent variables. The findings in Table 3 demonstrate that the explanatory variables did not exhibit multicollinearity.

Furthermore, the results of the MLR model in Table (3) show that all six jointly explanatory variables, such as predicted CO_2 emission using ARIMA, GDP per capita, electricity consumption, Labour force, and urban population, are statistically significant at a 1% level of significance. Predicted CO_2 emission and GDP per capita, used as the proxy of economic growth and electricity consumption, are all statistically significant at a 5% significance level. Urban population and labor force are statistically significant at a 1% significance level while electricity generation is statistically insignificant. The adjusted coefficient of determination (adjusted R^2) is 0.72, which means that 72% of the total variation is accounted for by six explanatory variables included in the model while other factors outside the analytical model can account for the remaining 28%. The model intercept is interpreted as the estimated function of the CO_2 emission with no effect on GDP per capita, electricity consumption, generation, labor force, urban population, and predicted CO_2 emission from ARIMA. The ARIMAX model's MAPE and RMSE for the testing data set are 21.92 and 17.22, respectively (see Table 5), demonstrating that it performs better than the ARIMA model.

Table (3): Combined ARIMA-MLR for CO_2 emission

C02 Emission	**Coefficient Estimate**	**Standard Error**	**Test statistic**	**P-value**	**VIF**
GDP Per Capita	0.253	0.012	-1.97	0.049	2.342
Predicted C02 emission from ARIMA	0.145	0.020	3.547	0.003	2.112
Electricity generation	-0.321	0.000	-0.71	0.475	2.675
Consumption	0.413	0.000	2.46	0.014	2.764
Urban population	1.726	0.310	5.56	0.000	2.084
Labor force	0.653	0.027	2.02	0.043	1.904
Intercept	0.040	0.149	0.27	0.788	-
Model P-value	0.000				

3.3 ARIMA-Quantile Regression (ARIMA-QR)

In the estimates of different covariates on the conditional mean of CO_2 emission, it has been acknowledged that the classical OLS regression method, which was used in ARIMAX, did not accurately capture the size and nature of the covariate effects on the lower and upper tail of the CO_2 emission distribution. Quantile functions can fully portray the covariate effects. Thus, quantile regression is used to regress the dependent variable with covariates. Figure (3 a) shows the CO_2 emissions against quantiles starting from 0.05 to 0.90 for all explanatory variables and an intercept. The shaded grey area offers the 90% confidence interval (CI) for the quantile regression estimates, the red line depicts the regression estimate of the average effect, and the two dotted red lines represent the 90% confidence intervals for the least square estimation. It can be seen that each of the estimated falls with the grey region, which is the 90% confidence interval, suggesting that the fitted quantile regression model is stable. Furthermore, Figure (3 b) predicts that CO_2 emission using ARIMA-QR is constant over time and that there is no cause for alarm. Table (4) displays all the covariate coefficients and standard errors for each quantile under consideration.

Table (4): Estimates of the quantile regression coefficients (ARIMA-QR)

C02 emission										
Quantiles	5th	10th	20th	30th	40th	50th	60th	70th	80th	90th
GDP Per Capita	0.347*** (0.028)	0.321*** (0.024)	0.296*** (0.022)	0.368*** (0.021)	0.375*** (0.024)	0.377*** (0.021)	0.362*** (0.019)	0.365*** (0.027)	0.381*** (0.017)	0.421*** (0.016)
Electricity generation	0.137 (0.063)	0.187 (0.067)	0.217* (0.042)	0.351*** (0.021)	0.345*** (0.021)	0.301 (0.086)	0.323 (0.086)	0.294 (0.074)	0.387* (0.027)	0.412*** (0.023)
C02 emission from ARIMA	0.267*** (0.045)	0.281*** (0.039)	0.312*** (0.051)	0.324*** (0.054)	0.253*** (0.061)	0.371*** (0.052)	0.295*** (0.041)	0.301*** (0.064)	0.276*** (0.042)	0.205*** (0.046)
Consumption	0.023* (0.011)	0.052* (0.012)	0.034 (0.013)	0.013 (0.023)	0.013 (0.023)	0.055*** (0.021)	0.041*** (0.031)	0.019*** (0.025)	0.033*** (0.035)	0.035*** (0.041)
Urban population	-0.117*** (0.031)	-0.137*** (0.036)	-0.338*** (0.038)	-0.381*** (0.034)	-0.404*** (0.029)	-0.401*** (0.027)	-0.432*** (0.024)	-0.454*** (0.028)	-0.384*** (0.023)	-0.302*** (0.026)
Labor force	0.351*** (0.011)	0.374*** (0.018)	0.405*** (0.015)	0.464*** (0.017)	0.564*** (0.025)	0.403*** (0.021)	0.487*** (0.028)	0.503*** (0.030)	0.514*** (0.032)	0.525*** (0.035)
Constant	0.521** (0.032)	0.491** (0.026)	0.472* (0.022)	0.502*** (0.025)	0.623 (0.037)	0.646 (0.041)	0.573 (0.039)	0.413*** (0.016)	0.458*** (0.016)	0.406*** (0.014)
Model P-value	**0.0000*** **									

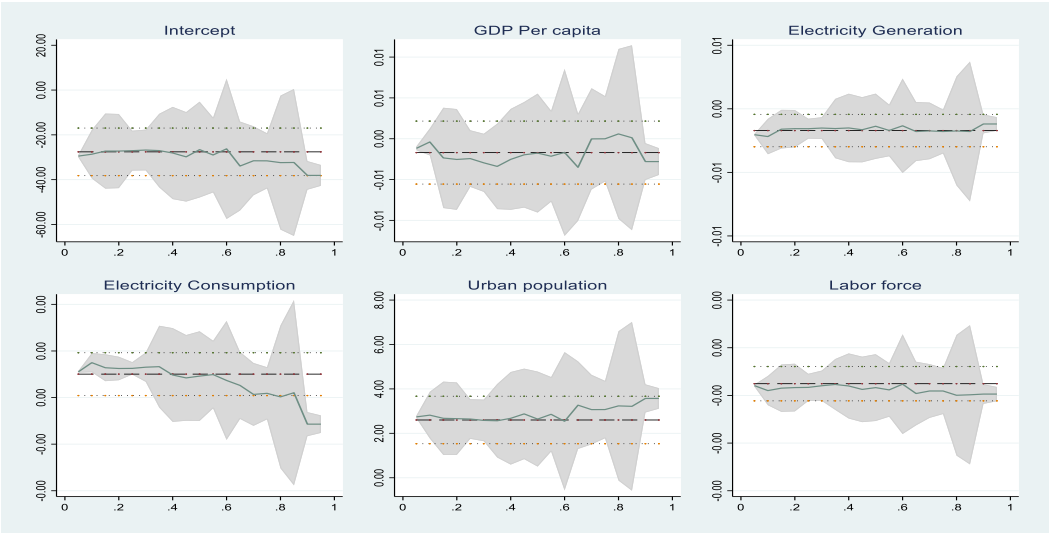


Figure (3 a): Ordinary Least Square and Quantile Regression Plots

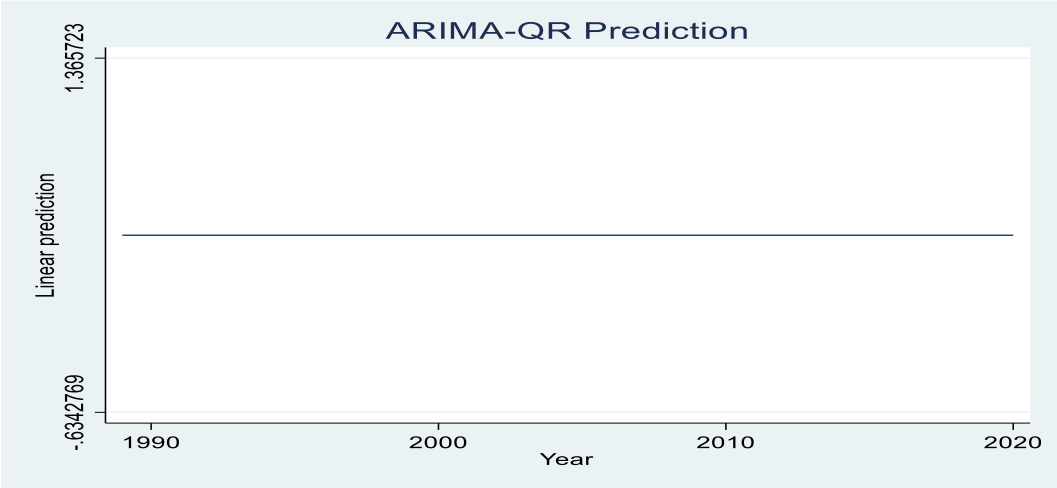


Figure (3 b): ARIMA-QR Forecasting

3.4 Performance Assessments Results

Table (5) shows the performance assessment results with the level of accuracy of carbon dioxide (CO_2) emission forecast with each model. ARIMA-MLR and ARIMA-QR have superior predictions than traditional ARIMA because they have lower mean absolute percentage error (MAPE) and root mean square error (RMSE) than standard ARIMA (see Table 5). This is due to the integration modification and use of influencing external factors. ARIMA-QR outperforms ARIMA-MLR in terms of forecasting accuracy because ARIMA-MLR produces only a mean forecast. In contrast, the ARIMA-QR model directly forecasts the quantiles rather than extrapolating them from a mean point. Section four explains the fundamental importance of ARIMA-QR over ARIMA-MLR.

Table (5): Performance Assessments Results for the out-of-sample dataset of CO_2 emission in all the forecasting models

Forecasting Method	MAPE	RMSE
ARIMA	22.71	19.86
ARIMA-MLR	21.91	17.22
ARIMA-QR	19.55	16.17

3.5 Cumulative Sum for Parameter Stability Test Procedure

According to [8], who presented the test statistic method based on the cusum of recursive residuals, the recursive residuals are shown to be independent and identically distributed as normal, with a zero mean and constant variance under the null hypothesis. The recursive cusum process graph will deviate from the predicted value of 0 if the coefficients change after a certain period. It relies on its conclusion on whether the time series undergoes sudden changes that the model doesn't anticipate. Technically, it examines the residuals structural break.

A Brownian movement can approximate the limiting distribution of the recursive cusum statistic's sequence. A linear function approximates the Brownian process limits at a certain significance level. It is recommended to reject the null hypothesis when the sample cusum process crosses the theoretical bounds at any point in time. This can be done by looking at a graph that displays the borders and the recursive cusum statistic. Furthermore, a null hypothesis can be tested using a test statistic method built on the maximum of the recursive cusum statistic. We can see that the recursive plot in figure (4) falls between the 95% confidence band, indicating that the ARIMAX model's mean is stable at a 5% significance level.

Additionally, the plot offers details about the stability of the regression model's coefficients. Note that the predicted value of the cusum of recursive residuals under the null hypothesis of no parameter instability is 0. The cusum of the recursive residuals in figure (4) below, which also shows that the regression model's parameters become stable over that time, falls inside the 95% confidence intervals in the middle of the sample.

The cusum of OLS residuals is used to calculate a similar test statistic. The OLS residuals are heteroskedastic and correlated under the null hypothesis. The OLS residuals have an anticipated value of 0 if there is an intercept or if the underlying process has a mean of 0 without an intercept. Unlike the recursive cusum process, which would drift away from 0 following a structural break, their cumulative sum always returns to 0. A Brownian bridge process can approximate the limiting distribution of the OLS cusum statistic under the null hypothesis [29].

Moreover, if a structural break occurs at a specific time, the OLS cusum statistic's absolute value peaks before reverting to its predicted value of zero. The parameter stability is tested using a graph having a constant upper and lower bound at a given significance level. Reject the null hypothesis of parameter stability if the cusum of the OLS residual plot exceeds the boundaries. The plot further explains the timing of the structural break. The OLS cusum statistic's absolute value can also be employed to test the null hypothesis at a particular significance level. The OLS cusum process crossing within the 95% band in Figure (4 b) below indicates that the null hypothesis is rejected at a 5% significance level, supporting the stability of the ARIMAX mean.

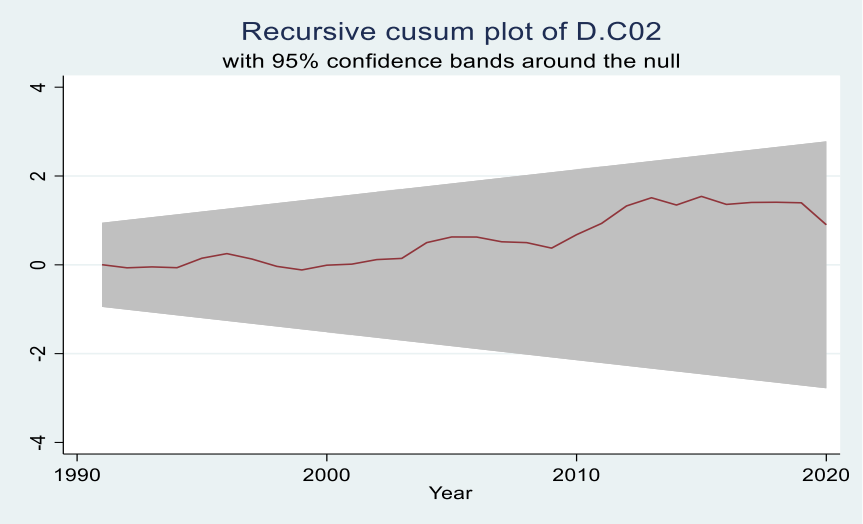


Figure (4 a): Recursive Cusum Plot

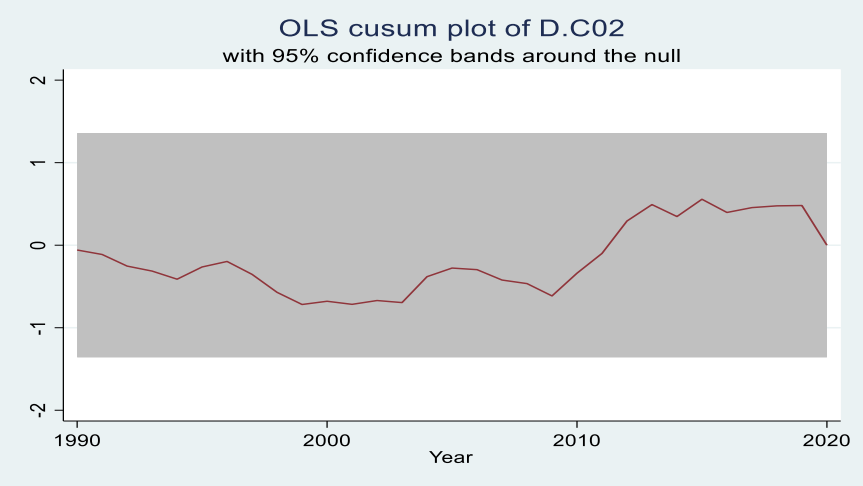


Figure (4 b): OLS Cusum Plot

4. Results and discussion

4.1 ARIMA-MLR and Q.R. Results

The results of the ordinary least square regression technique, which were employed in the ARIMA-MLR model, have been acknowledged as failing to adequately reflect the size and type of the effects of the covariates on the lower and higher part of the CO_2 emission distribution. Multiple linear regression findings in Table 3 reveal that the intercept (c) is not different from zero, but it is statistically significant from the 5th to 30th quantiles and from the 70th to 90th quantiles in the quantile regression model (QR). The predicted CO_2 emission from the ARIMA model is significant in low, middle, and upper quantiles (see table 4). MLR and Q.R. show that predicted CO_2 emission from ARIMA is an increasing function of CO_2 emission in Tanzania. The computed conditional quantile function demonstrates that the covariate's effect also increases as the quantiles increase.

Furthermore, similar effects are seen for both conditional mean and median. The coefficient of electricity generation is positive and significant in the 20th and 80th quantiles at a 10% significance level. Similarly, at a 1% level of significance, the coefficient of electricity generation is positive and statistically significant in the 30th and 90th quantiles. Additionally, the electricity generation coefficient is positive and significant at a 5% significance level in the 40^{th} quantile. The ordinary least square (OLS) results show that electricity generation has no impact on CO_2 emission. This is a perfect example of how inaccurate results of conventional least squares predictions appear.

The estimated coefficient of GDP per capita is highly positively significant at a 1% level of significance in all quantiles, indicating economic growth increases CO_2 emissions in Tanzania. This is because most East African economies, including Tanzania, are based on inefficient farming practices that result in deforestation, shrub growth, and rising pollution. Similarly, OLS suggests the same results. This finding supports [2], who also found that economic growth is the key source of environmental pollution in the Middle East and North Africa (MENA) countries, including Egypt, Bahrain, Iran, Iraq, Jordan, Libya, Oman, Syria, Kuwait, Lebanon, Morocco, Yemen, and the United Arab Emirates. Electricity consumption positively affects CO_2 emission in upper and lower quantiles. Similarly, ordinary least squares suggest a positive relationship with CO_2 emissions. This result supports [1],[11],[18], who also found that electricity consumption increases CO_2 emissions.

The negative coefficient of the urban population in both ordinary least square estimates of MLR and quantile regression (Q.R.) models suggests that the urban population does not increase CO_2 emission in Tanzania. This is contrary to the study conducted by [34], who pointed out a long-term equilibrium link between CO_2 emissions and the urban population in Malaysia. In Uganda, [20] investigated the association between CO_2 emissions and the urban population, and their findings revealed that the urban population is a decreasing function of CO_2 emissions. The coefficient of the labor force has a positive sign in both Q.R. and MLR, showing that an increase in the labor force significantly impacts CO_2 emissions.

4.2 Applicability of Fitted Models for Forecasting

Every prediction method has advantages and disadvantages, and every prediction scenario is constrained by limitations such as data, time, capacity, and budget. It is a challenging and crucial responsibility of management to weigh the benefits and drawbacks of various methods in light of available resources and limits. The following significant technical selection criteria can be used to determine whether a forecasting model is appropriate for forecasting CO_2 emission:

- Data availability: By considering the volume of historical data, traditional time series models (such as ARIMA, AR, MA, and ARMA) and merged time series models (ARIMA-MLR and ARIMA-QR) consistently provide a better result (both response and predictor variables). In comparison, hybrid models demand more data than traditional time series models for better approximation, which means that when the in-sample data set is inadequate, the forecasts are less consistent and precise.
- Time series components: Every time series comprises unobserved components such as regular trends, cyclic, seasonal, and irregular variations. The classic time series models (ARIMA) are beneficial in identifying and elucidating any regular and irregular variations caused by these time series components. The hybrid/merged models (ARIMA-MLR and ARIMA-QR) can explain how some external influencing variables function in the presence of the time series components.
- Accuracy: In Forecasting CO_2 emissions, the forecasts are not directly utilized. Instead, they become part of environmental policy, i.e., the forecast is used to gauge the level of pollution. Thus, policymakers must address the forecast error, which can be abided by or acceptable to some extent. The findings in Table 5 demonstrate that the ARIMA-MLR and ARIMA-QR models are more accurate than the ARIMA model.

- Uncertainty: The ARIMA-MLR model's primary drawback is that it only offers point forecasts. The projection will only supply half the service level because this model only provides mean forecasts. The ARIMA-MLR model also considers highly volatile and skewed time series with a Gaussian distribution. The extrapolation of higher quantiles from the mean forecast will not reflect reality because the proper distribution of CO_2 emissions is not normal. Another issue is the uncertainty in the data because of expectable and unexpected effects on CO_2 emission. In these circumstances, the mean estimates from the ARIMA-MLR model may be either low or too high in terms of spike CO_2 emission, which could result in poor decision-making. Also, the ARIMA-QR model accurately predicts and identifies excessive and sparse CO_2 emissions. It is helpful to forecast the higher quantiles directly without relying on normality or quantile extrapolation assumptions.

5. Conclusion

A proper prediction is needed when modeling environmental data to help policymakers make accurate decisions. However, the time series data of CO_2 emission is usually skewed to the right with the tail towards high concentration. This is a significant constraint but is frequently not considered by most analysts or researchers in forecasting. To solve this problem, it is necessary to establish a time series forecasting model that considers forecast uncertainty and the influence of external factors like GDP per capita, electricity generation and consumption, labor force, and urban population. This study established ARIMA-MLR and ARIMA-QR models to forecast annual CO_2 emissions in Tanzania. Compared to the ARIMA model, both ARIMA-MLR and ARIMA-QR models produce more accurate training and test data set forecasts. The ARIMA-MLR model only generates the point forecast or the mean forecast. The computation of prediction intervals from the point forecast will not accurately represent reality since the proper distribution of CO_2 emissions is not normal. Furthermore, the ARIMA-QR model has the following extra advantages compared to the ARIMA-MLR model:

- It facilitates accurate, direct forecasting of the higher quantiles without extrapolation.
- The quantile regression results provide a precise and focused understanding of the impacts of the variables.
- If future CO2 emission predictions are the primary concern, the model can help to make informed judgments.

Acknowledgements

We thank the reviewers for their insightful comments, which have led to many improvements in this paper. Liu Qing's research is supported by the National Social Sciences Fund of China (No. 21BTJ035)

References

[1] A. L. I Amjad, S. Khatoon, M. Ather & N. Akhtar. Modeling energy consumption, carbon emission and economic growth: Empirical analysis for Pakistan. *International Journal of Energy Economics and Policy*, 5(2) (2015), 624-630.

[2] M. E. H Arouri, A. B. Youssef, H. M'henni, & C. Rault. Energy consumption, economic growth and CO2 emissions in Middle East and North African countries. *Energy policy*, 45 (2012), 342-349. https://doi.org/10.1016/j.enpol.2012.02.042

[3] N. S. Arunraj & D. Ahrens. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, 170, (2015), 321-335, https://doi.org/10.1016/j.ijpe.2015.09.039

[4] S. F. Azadeh, Ghaderi, & S. Sohrabkhani. A simulated-based neural network algorithm for forecasting electrical energy consumption in Iran. *Energy policy*, 36(7), (2008), 2637-2644. https://doi.org/10.1016/j.enpol.2008.02.035

[5] A. Azadeh, M. Khakestani & M. Saberi, *A flexible fuzzy regression algorithm for forecasting oil consumption estimation.* Energy Policy, 37(12), (2009),5567-5579.

[6] D.F. Benoit, & D. Van den Poel, *Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services*, Expert Systems with Applications, 36(7), (2009), 10475-10484.

[7] G. E. Box & G.M. Jenkins, *Time series analysis, forecasting and control San Francisco. Calif:* Holden-Day, *(1976).* https://doi.org/10.1177/05831024820140060

[8] R. L. Brown, J. Durbin, & J. M. Evans. Techniques for testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(2) (1975), 149-163. https://doi.org/10.1111/j.2517-6161.1975.tb01532.x

[9] C. Chatfield. *Time-series forecasting.* Chapman and Hall/CRC, (2000).

[10] D. Coondoo, & S. Dinda. Carbon dioxide emission and income: A temporal analysis of cross-country distributional patterns. *Ecological Economics*, 65(2) (2008), 375-385. https://doi.org/10.1016/j.ecolecon.2007.07.001

[11] E. Dogan & F. Seker. The influence of real output, renewable and nonrenewable energy, trade and financial development on carbon emissions in the top renewable energy countries. *Renewable and Sustainable Energy Reviews*, 60, (2016),1074-1085. https://doi.org/10.1016/j.rser.2016.02.006

[12] E. Fleisch & C. Tellkamp. Inventory inaccuracy and supply chain performance: a simulation study of a retail supply chain. *international journal of production economics*, 95(3) (2005),373-385. https://doi.org/10.1016/j.ijpe.2004.02.003

[13] M. Gilliland, *Business Forecasting Effectiveness*, Analytics, 21–25, (2011), http://www.analytics-magazine.org/july-august-2011/361-value-added-analysis business-forecasting-effectiveness

[14] M. Gilliland, & U. Sglavo. Worst practices in business forecasting. *Analytics*, (2010), 12, 17.

[15] C. W. J. Granger, H. White, H., & M. Kamstra. Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, 40(1) (1989), 87-96.

[16] R. J. Hyndman & G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, (2018).

[17] R.B. Jackson, C. Le Quéré, R. M. Andrew, J. G. Canadell, J. G. Korsbakken, J. I., Liu, & B. Zheng. Global energy growth is outpacing decarbonization. *Environmental Research Letters*, 13(12) (2018), 120401. https://doi.org/10.1088/1748-9326/aaf303

[18] M. B. Jebli, S.B.Youssef & I. Ozturk. Testing environmental Kuznets curve hypothesis: The role of renewable and nonrenewable energy consumption and trade in OECD countries. *Ecological Indicators*, 60 (2016), 824-831. https://doi.org/10.1016/j.ecolind.2015.08.031

[19] O. Kaynar, I. Yilmaz & F. Demirkoparan. Forecasting of natural gas consumption with neural network and neuro fuzzy system. *Energy Education Science and Technology Part A: Energy Science and Research*, 26(2) (2011), 221-238.

[20] S. Klasen & D. Lawson. The impact of population growth on economic growth and poverty reduction in Uganda. *Diskussionsbeiträge* (2007), No. 133,

[21] R. Koenker, & G. Bassett Jr. Regression Quantiles. *Econometrica*, 46(1) (1978), 33–50. http://doi.org/10.2307/1913643

[22] R. Koenker & K. F. Hallock. Quantile regression. *journal of econometrics perspective*, 15 (4) (2001), 143-156.

[23] C. Q. Le, J. I. Korsbakken, C. Wilson, J. Tosun, R. Andrew, R. J. Andres, & D.P Van Vuuren. Drivers of declining CO2 emissions in 18 developed economies. *Nature Climate Change*, 9(3) (2019),213-217. https://doi.org/10.1038/s41558-019-0419-7

[24] G.Q Li, S.W. Xu, Z.M. Li, Y.G. Sun & X. X. Dong. Using quantile regression approach to analyze price movements of agricultural products in China. *Journal of Integrative Agriculture*, 11(4) (2012), 674-683. https://doi.org/10.1016/s2095-3119(12)60055-0

[25] R. Mamlook, O. Badran & E. Abdulhadi. A fuzzy inference model for short-term load forecasting. *Energy Policy*, 37(4) (2009), 1239-1248. https://doi.org/10.1016/j.enpol.2008.10.051

[26] N. Meinshausen, & G. Ridgeway. Quantile regression forests. *Journal of machine learning research*, (2006),7(6).

[27] H. Nie, G. Liu, X. Liu & Y. Wang. Hybrid of ARIMA and SVMs for short-term load forecasting. *Energy Procedia*, 16 (2012), 1455-1460. https://doi.org/10.1016/j.egypro.2012.01.229

[28] H. T. Pao, H.C. Yu & Y. H. Yang. Modeling the CO2 emissions, energy use, and economic growth in Russia. *Energy*, 36(8) (2011), 5094-5100. https://doi.org/10.1016/j.energy.2011.06.004

[29] W. Ploberger & W. Krämer. The CUSUM test with OLS residuals. *Econometrica: Journal of the Econometric Society*, 60(2) (1992), 271. https://doi.org/10.2307/2951597

[30] A. Pueyo. what constrains renewable energy investment in Sub-Saharan Africa? A comparison of Kenya and Ghana. *World Development*, 109 (2018),85-100. https://doi.org/10.1016/j.worlddev.2018.04.008

[31] W. Qiao, H. Lu, G. Zhou, M. Azimi, Q. Yang & W. Tian. A hybrid algorithm for carbon dioxide emissions forecasting based on improved lion swarm optimizer. *Journal of Cleaner Production*, 244 (2020), 118612. https://doi.org/10.1016/j.jclepro.2019.118612

[32] S. A Sarkodie & P. K. Adom. Determinants of energy consumption in Kenya: a NIPALS approach. *Energy*, 159 (2018), 696-705. https://doi.org/10.1016/j.energy.2018.06.195

[33] J.W. Taylor. forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, 178(1) (2007), 154-167. https://doi.org/10.1016/j.ejor.2006.02.006

[34] W. H. Tsen & F. Furuoka. The relationship between population and economic growth in Asian economies, *ASEAN Economic Bulletin*, 22(3) (2005), 314-330. https://doi.org/10.1355/ae22-3e

[35] UNEP Annual evaluation report 2002

[36] UNEP Annual evaluation report 2003

[37] R. E. Walpole, R. H. Myers, S. L. Myers & K. Ye. *Probability and Statistics for Engineers and Scientists (9th ed.)*. Prentice Hall.