Forecasting students' enrolment in Tanzania government primary schools from 2021 to 2035 using ARIMA model

Sophia Kornelio[a], Ramkumar. T. Balan [b], Emmanuel Deogratias [c1]

[a,b] University of Dodoma, P.O.Box 259, Dodoma, Tanzania
[c] Sokoine University of Agriculture, P.O.Box 3038, Morogoro, Tanzania

Abstract

This article focuses on Forecasting students' enrolment in Tanzania Government Primary Schools from 2021 to 2035. A school enrolment is one of the key determinants of the needed school infrastructure. In Tanzania the Fee Free Primary Education (FFPE) started in 2015 and has caused the enrolment in Tanzania government primary school to increase by 26.8% from 2015 to 2020 which resulted into shortage of schools' infrastructure. The study will provide an information on the primary schools' enrolment from 2021 to 2035 which will help in sensitized the community about the need of their contributions to school infrastructure in their localities. The study used secondary data on total enrolments in Tanzania Government primary schools collected yearly from President's Office, Regional Administration and Local Government (PO-RALG) from 1961 to 2020. Basing on the theory of Box and Jenkins method the results indicated that the ARIMA (1,1,0) is the best fit model to forecast Tanzania government primary's school enrolments because it had the lowest Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) of 1618.3417 and 1622.4626 respectively. Enrolment has increased from 10,693,349 to 11,641,341 from 2021 to 2035. The study revealed that ARIMA (1, 1, 0) is the best fitted model to forecast Tanzania government primary school enrolments and can be used to support planning process in Tanzania government primary schools.

Keywords: Forecasting; student enrolment; ARIMA model; primary schools

1. Introduction

Everyone has a right to get education (Bircan & Sunata, 2015). Education is very important for an individual's success in life and development a country's economy and is the milestone of national's development (Idris et al., 2012). For the improvement of the standard of living of Tanzanian citizens, the government recognizes that education is important to each individual in our country. According to (Rutaremwa et al., 2013).

In 2016 the 5th president of the United Republic of Tanzania (URT), the late John Pombe Magufuli, initiated Fee Free Primary Education (FFPE) in Tanzania as one of

[1] Corresponding author: Emmanuel Deogratias. ORCID ID.: https://orcid.org/0000-0000-0000-0000
E-mail address deograti@ualberta.ca

the 2015 election manifesto and the implementation of the Education and Training Policy (ETP) in which all the learning expenses are the responsibility of the government (Twaweza, 2016).

Due to FFPE caused enrolments in Tanzania government primary schools to increase by 26.8% from 8,245,382 pupils in year 2015 to 10,937,373 pupils in 2020 as shown in Table 1.1.

Table 1.1: Enrolment in Tanzania government primary schools from 2015-2020

Year	Total enrolment
2015	8,245,382
2016	8,639,202
2017	9,317,791
2018	10,111,671
2019	10,605,430
2020	10,937,373

SOURCE: PO-RALG (2015-2020)

This increase of students' enrolment every year, demands the need to realize the future trend for the upcoming years for effective future planning in Tanzanian primary school education, including having adequate school infrastructure basing on national standards.

1.1. Student enrolment and forecasting in the literature

Student enrolment is the process of arranging to attend an institution and specific classes (Caspi et al., 2020). It involves the number of students that currently attend a school or course. The enrolment process is completed after a student is granted admission to a particular school.

Kornelio (2017) conducted the research on forecasting membership enrolment of National Health Insurance fund (NHIF) (Case study Dodoma Region). The NHIF membership enrolment of the Dodoma region was studied using the Box – Jenkins (1976) Autoregressive Integrated Moving Average approach from 2002 to 2016. As a result, the NHIF membership has been anticipated for the financial years 2016/2017 to 2017/2018 (using seasonal ARIMA for twenty-four months) ARIMA (1,2,0) x $(0,0,1)_{12}$ for male, seasonal ARIMA (0,2,0) x $(1,0,0)_{12}$ for female and seasonal ARIMA $(0,2,1)_{12}$ for total combined. And the NHIF membership registration will grow up by the financial year 2016/2017 to 2017/2018 when only time is considered as a factor.

Lavilles and Arcilla (2012) conducted a study on enrolment forecasting for school Management system by using Moving Average, Single and exponential Smoothing to determine which model can represent data of the university. Over a five-year period, the study constructed a forecasting model utilizing three different methodologies. As a result, Simple Moving Average was found not a good fit for their pattern. The single exponential Smoothing approach with an alpha of 0.9 had a low MAPE and predicted 20.5 percent greater accuracy than the previous naive forecasting model. The MAPE of the double exponential Smoothing approach, which used alpha of 0.9 and beta of 0.1, was 16.4%. Based on the given data, the researchers employed 182 people to create the least error model. About 58 percent of the individuals employed double exponential smoothing, whereas the other 42 percent utilized single exponential smoothing. The

MAD and MSE values were not calculated in the investigation, and the alpha value was not optimal.

From the reviewed literature, it is clear that there are studies that have been conducted on the school enrolments at university level projecting different phenomena like application, admission through time series model like ARIMA Model and not in Tanzania government primary schools' enrolment. This study focuses on forecasting Tanzanian government primary school enrolments.

1.2. Conceptual framework

The role of a conceptual framework is to display the relationship between the variables that are used in the study. As such, it connects different facts and directs the collection of relevant facts or evidence (Msuya, 2009). The conceptual framework of the research is centered on the overall goal of the study, which is to forecast student enrolment in Tanzania government primary schools from 2021 to 2035. The conceptual framework for this study is presented in Figure 2.1 based on Box-Jenkins's methodology.

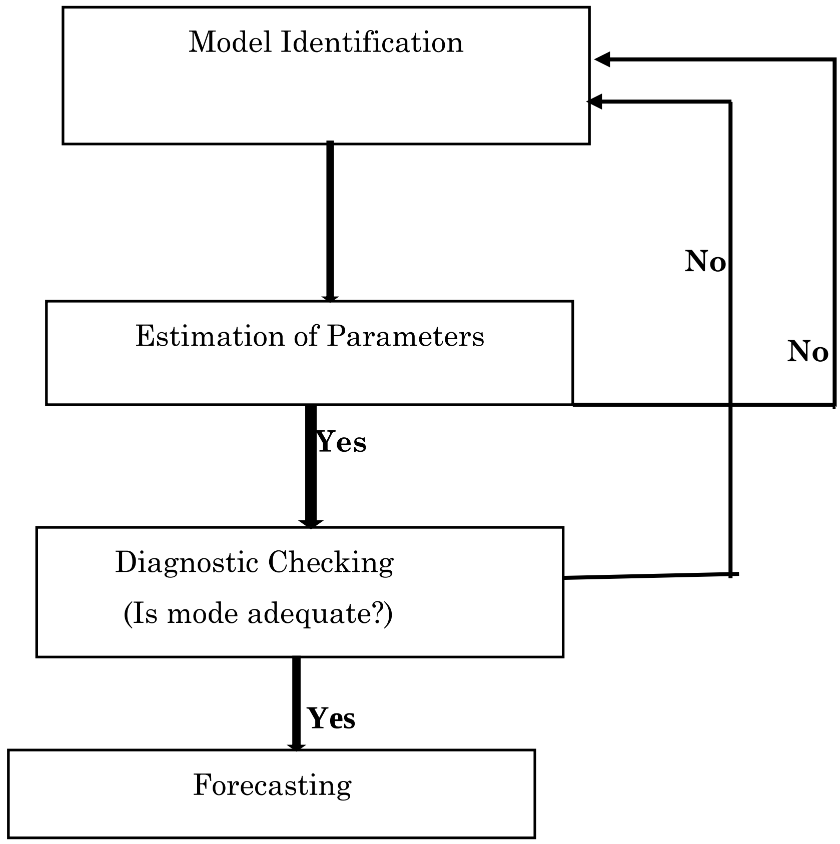

Figure 2. 1: Framework of Box-Jenkins's methodology

2. Method

The study uses secondary quantitative data on total enrolment in Tanzania government primary schools on an annual basis. The data was collected for a 60-year period from 1961 to 2020 and longitudinal research strategy was chosen for this study.

This study relied on secondary data from Tanzania's President Office, Regional Administration and Local Government (PO-RALG), which is responsible for compiling the country's basic education statistics. These data were utilized to construct an enrolment prediction model by select the best fit model for forecasting Tanzania government primary schools' enrolment.

2.1. Data analysis

Outliers were examined, missing values were tested, and stationarity was tested before data processing. Outliers were examined using a Box plot, which is suitable for expressing outliers clearly.

The Box - Jenkins method used to build an ARIMA model. This method has some steps which are: model identification, estimation of parameters, diagnosis of the fit and forecasting using the fit. The Box Jenkins method requires stationarity as one of its conditions.

After inferring if time series data are stationary or non-stationary from a visual observation of time series plots, a formal test was used to determine whether the data are stationary or not stationary. Transformation of data was conducted and the Dickey-Fuller was employed to determine whether the data were stationary or non-stationary.

Here we tested the hypothesis that:

$H_a: \emptyset = 1$ versus

$H_a: \emptyset \neq 1$

Whenever the error term is auto correlated, which is an error in this test, the significance level of the normal test is unreliable. This difficulty can be handled by generalizing the Dickey-Fuller test into the augmented-Fuller (ADF) test to aid the generic ARIMA and ARMA models (Greunen et al., n.d.).

For this study the time series data was observed to be non- stationary; transformation of data was done and Augmented-Fuller (ADF) was implemented and the data became stationary.

3. Results and discussions

3.1. Data processing

The box plot was used to test whether there was their outlier or no outlier which is crucial to be tested in times series data in order to avoid biased results.

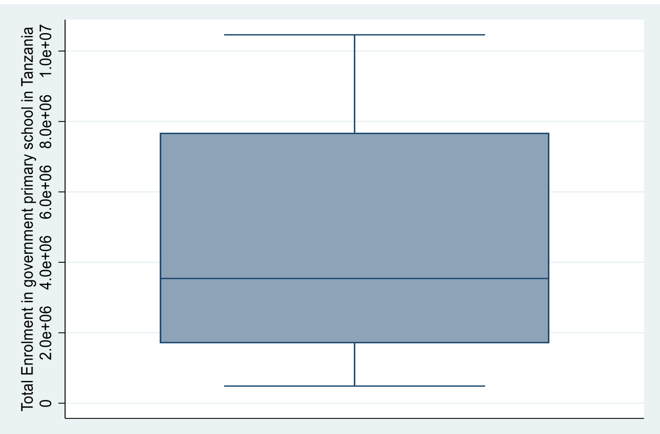

Figure 4.1 shows that there is no outlier since there is no any point which lies above or below the box plot. This implies that the findings obtained from these data are not affected by the outlier.

3.2. Model identification process

3.2.1. Time series plots for number of students enrolled from 1961 to 2020

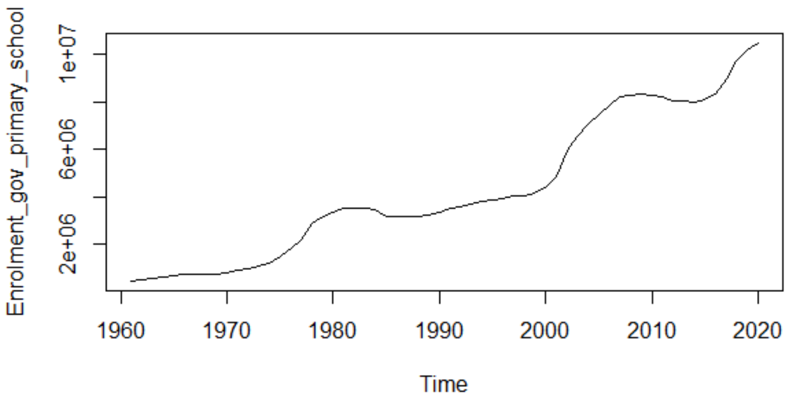

Figure 4.2: Time series plot for total enrolment of Tanzanian government primary schools

The plot depicts the presence of non-constant variance and mean. As time increases, there is increase in students' enrolment in Tanzania government primary schools ($\boldsymbol{y_t}$). Therefore, transformation should be applied to make the time series data for number of students' enrolment ($\boldsymbol{y_t}$) stationary.

3.2.2. Testing for stationarity by using Augmented Dickey-Fuller (ADF) test

Since the original data had shown the tendencies of non-stationarity in nature, therefore the transformation is useful in order to make the data stationary.The stationarity of W_t was confirmed by conducting the unit root test and in this case, the Augmented Dickey-Fuller (ADF) test was applied.

$$y_t = c + \delta t + \phi y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + \varepsilon_t$$

$H_0: \emptyset = 1$ versa $H_a: \emptyset \neq 1$

This test was done at 5% level of significance at different number of lags. the results are presented in Table 4.1

Table 4.1: Augmented Dickey-Fuller test for the first difference series of students' enrolment in Tanzania government primary schools

No	Lags	Test Statistic	Critical Value	Significance Level	Model
1	0	-41.7792	-1.9416	0.05	AR
2	5	-11.6154	-1.9416	0.05	AR
3	10	-8.9963	-1.9416	0.05	AR
4	15	-7.7234	-1.9416	0.05	AR
5	20	-6.7486	-1.9416	0.05	AR

Since all test statistic in absolute values were greater than the critical value, the null hypothesis is rejected. This implies that the first difference for number of students enrolled in Tanzania government primary schools (W_t) does not contain a unit root or is stationary at the first order of integration.

3.2.3. Model building process

After confirmation of stationarity for students' enrolment in Tanzania government primary schools time series data, W_t ARIMA models can be proposed by observing at the PACF and ACF plots as shown in Figures 4.3 and 4.4.

The PACF and ACF of stationary data for lag 1 to 20 are always observed for the model building process. Both seasonal and non-seasonal observation components were carried out to determine the order of AR and MA. The autocorrelation at lag 1, lag 2 and lag 3 were above the significance limit, according to ACF plot Figure 4.3 because these three consecutives autocorrelation are above the upper and lower limits. This implies/suggests that this (q=3) and the autocorrelation tails off to zero after lag 3. The autocorrelation at lags 6, 7, 8, and 9 exceeds the limit, but the significance limit and the rest of the data are within the limits. We can infer those lags 13, 16, and 18 as all errors that occur by chance.

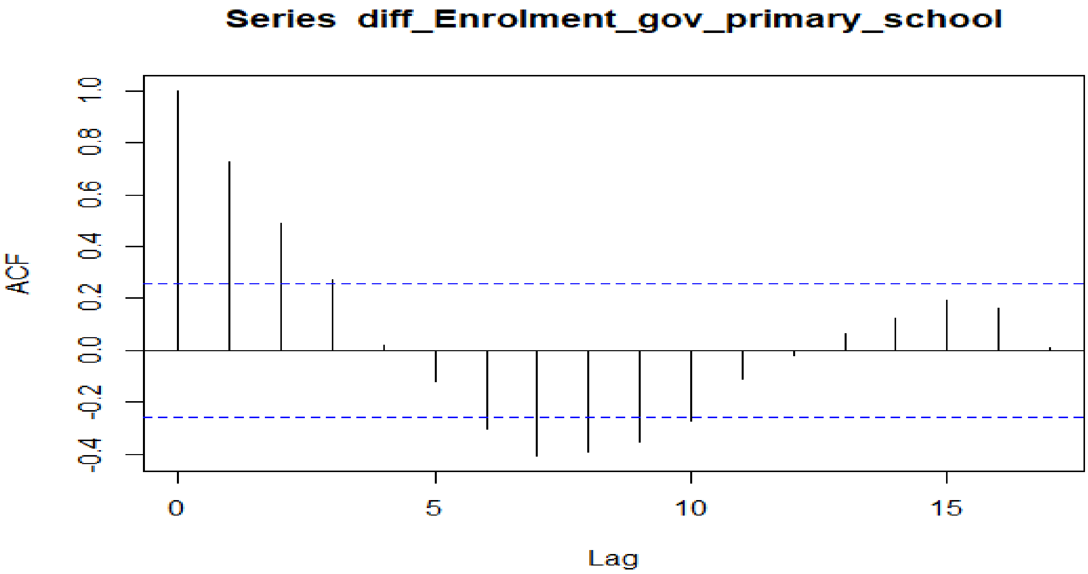

Figure 4.3: Autocorrelation Function for First Difference of Total Enrolment in Tanzania Government Primary Schools

Figure 4.3 shows that, autocorrelation coefficient at lag 0, at lag 1, at lag 3 and at lag 4 exceeds the significance limit and tail off to zero after lag 3. Although at lag 6, at lag 7, at lag 8, at lag 9 and at lag 10, it exceeds the significance limits. The rest of the coefficients between lag 0 and lag 1 are within the significance limit.

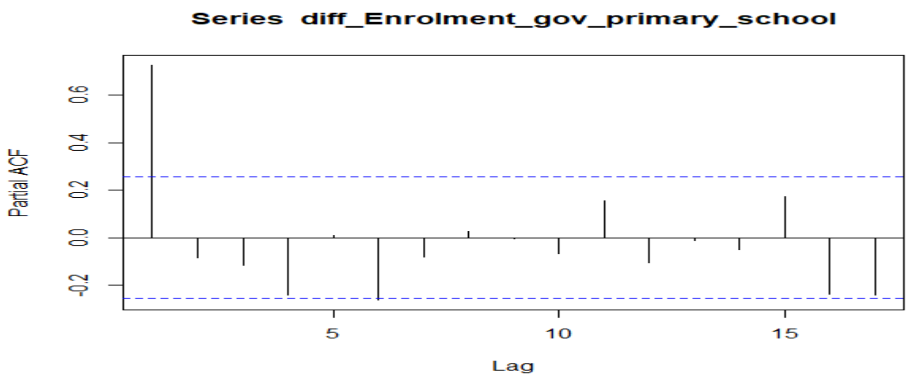

Figure 4.4: Partial Autocorrelation function for first difference of total enrolment in Tanzania government primary schools

Figure 4.4 shows that at lag 1, the partial autocorrelation coefficient at lag 1 exceeds and tail off to zero after lag 1. At lag 4, at lag 6 and at lag 16 and lag 17 it exceeds the significance limit. We have to assume these errors to have happened by chance.

3.2.4. Model suggestion

Models for the first difference for students' enrollment in Tanzanian government primary schools ($\boldsymbol{W_t}$) are ARIMA (2,1,2), ARIMA (0,1,0), ARIMA (0,1,1), ARIMA (2,1,0), ARIMA (2,1,1), ARIMA (1,1,0), ARIMA (1,1,2) and ARIMA (1,1,3).

3.2.5. Model estimation

In selecting the best ARIMA model among the suggested models, the statistics presented in Table 4.2 below were used.

3.2.6. ARIMA model selection

Table 4.1: Statistics for the suggested ARIMA Models for students' enrolment in Tanzanian government primary schools

Model	AIC	BIC
ARIMA (2,1,2) Model	1635.6277	1645.843
ARIMA (0,1,0) Model	1687.3043	1689.3818
ARIMA (0,1,1) Model	1639.7112	1645.9438
ARIMA (2,1,0) Model	1620.3172	1626.4464
ARIMA (2,1,1) Model	1633.6900	1641.8622
ARIMA (1,1,0) Model	1618.3417	1622.4626
ARIMA (1,1,2) Model	1622.3045	1630.5463
ARIMA (1,1,3) Model	1623.0186	1633.3208

Based on Table 4.2, ARIMA (1, 1, 0) model was identified and chosen as the best model as compared to other ARIMA models because has the lowest Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC).

Table 4.3: Estimated results for students' enrolment in Tanzanian government primary schools, ARIMA (1, 1, 0)

Parameter	Value	Standard Error	T-Statistic	P-Value
Constant	0	0		
AR {1}	0.82513	0.046503	17.7435	1.9347e-70
Variance	28684491149.4017	3.0712e-13	9.339968713406031e+22	0

From Table 4.3, it shows that all parameters for ARIMA (1, 1, 0) were statistically significant since the p-value was less than 0.05 for all parameters. Autoregressive integrated moving average model for ARIMA (1, 1,0) for number of student enrolments in Tanzania government primary time series data with the following equation is the best model:

$$y_t = \emptyset_1 y_{t-1} + y_{t-1} - \emptyset_1 y_{t-2} + \varepsilon_t$$

3.2.7. Model diagnostic checks

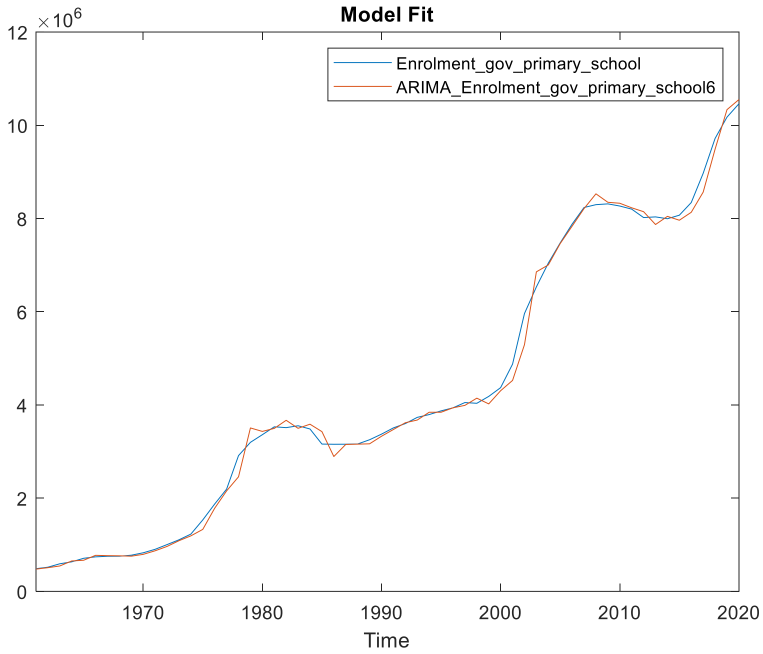

Figure 4.1: Model fit for students' enrolment in Tanzanian government primary schools

Validation was done in the context of this study for number of students' enrolment in Tanzanian government primary time series data, and the results are shown in figure 4.3. It can be seen that the fitted model for number of students' enrolment in Tanzanian government primary schools time series data closely resembles the observed data series. Also, it appears that the fitted values smooth out the observed values' highs and lows.

The model fits the observed data for students' enrolment in Tanzanian government primary schools as shown in Figure 4.3. This means that the ARIMA (1,1,0) model fits well the data for number of student enrolment in Tanzania government primary because the predicted and observed students' enrolment in Tanzanian government primary schools are similar.

3.3. Forecasting with the best fitted model for number of students enrolled in Tanzanian government primary schools

3.3.1. One-step-ahead forecast for number of students' enrolment in Tanzania government primary schools

The diagnostic checks reveal that the residuals are completely random, indicating that the fit can be used to estimate future values of the number of students enrolled in Tanzania government primary time series data. However, before we use the fit to forecast, we must assess the accuracy of its predictions. This is done by forecasting one step forward and comparing it to a test data set or observed data. Therefore, the fitted model that is, ARIMA (1,1,0) is used to forecast the number of student enrolment in Tanzania government primary schools for the period of 15 years. The predicted and observed values are depicted in Table 4.4.

Table 4. 2: Forecasting value from ARIMA (1, 1, 0) for number of student enrolments in Tanzania government primary schools

Time period	Observed value	Forecast	Forecast error	Absolute error	Squared error	Absolute (%) error
2011	8200651	8200631.56	19.44	19.44	377.9136	0.000237054
2012	8019748	8019744.08	3.92	3.92	15.3664	4.88793E-05
2013	8033926	8033924.34	1.66	1.66	2.7556	2.06624E-05
2014	7994675	7994678.67	-3.67	3.67	13.4689	4.59056E-05
2015	8070290	8070290.98	-0.98	0.98	0.960400001	1.21433E-05
2016	8341611	8341611.56	-0.56	0.56	0.3136	6.71333E-06
2017	8969110	8969110.23	-0.23	0.23	0.0529	2.56436E-06
2018	9717309	9717309.34	-0.34	0.34	0.1156	3.49891E-06
2019	10174237	10174237.2	-0.24	0.24	0.0576	2.3589E-06
2020	10460785	10460785.4	-0.35	0.35	0.1225	3.34583E-06
Total		10545083.1	18.65	31.39	411.1271	0.000383126

From table 4.4: show that from the year 2011 to 2020 the observed value and the forecast values as the same this indicate that the ARIMA (1, 1, 0) is the best fit model to forecast the students' enrolment in Tanzanian government primary schools.

3.3.2. Evaluation of forecast accuracy of ARIMA (1, 1, 0) for students' enrolment in Tanzania government primary schools

The standard measures of forecast accuracy were generated to evaluate the model's forecasting capability, as shown in Table 4.5.

Table 4. 3: Some measures of forecast accuracy for ARIMA (1, 1, 0)

Statistical Measure	**Number of Enrolled students**
Mean error (ME)	1.8650000
Mean absolute deviation (MAD)	3.1390000
Mean squared error (MSE)	41.112710
Mean absolute percentage error (MAPE)	0.0000383

3.3.2.1. Analysis of forecast errors

The forecast errors analysis is important since it is used to assess the accuracy of future forecasts of the fitted model. We examine the usual measures of prediction accuracy when evaluating the model's predicting performance in Table 4.5. These forecast accuracies of the model are tested as portion of the model fit validation process. The mean forecast error values by the ARIMA (1, 1,0) model for number of students enrolled in Tanzanian government primary schools, are near to zero mean that the forecasts produced by this ARIMA (1,1,0) model are unbiased. Also, it is detected that the fit for the number of students enrolled in Tanzania government primary schools with ME value 1.8650000 was unbiased. The positive sign proposes that the ARIMA (1, 1,0) model is forecasting too low on average (Owusu, 2010). Both MAD and MSE propose that there is low variability.

Both MAD and MSE show that the number of students enrolled in Tanzanian government primary schools forecast errors which generate low variability. The percent forecast errors or mean absolute percentage errors (MAPE) show that the value for the number of students enrolled in Tanzanian government primary schools is 0.00383%. This means that forecasting by ARIMA (1, 1, 0) model for number of students' enrolment in Tanzania government primary schools was accurate since MAPE was lower than 20%.

3.3.2.2. Forecasted number of students enrolled in Tanzanian government primary schools from 2021 to 2035

After assessing the accuracy of forecasting, the ARIMA (1, 1, 0) for number of students' enrolments in Tanzania government primary was employed to generate out-of-sample forecasts for number of students enrolled in Tanzanian government primary schools for the period starting from 2021 up to 2035.as shown in Table 4.8.

Table 4. 4: Forecasted students' enrolment in Tanzania government primary schools

Year	predicted Total Number of Enrolment	Standard error of predicted	95% lower confident limit	95%upper confident limit
2021	10693349	171295	10357617	11029082
2022	10882100	354457	10187377	11576824
2023	11035292	551995	9953403	12117182
2024	11159624	754747	9680346	12638901
2025	11260532	957406	9384050	13137014
2026	11342430	1156855	9075036	13609824
2027	11408899	1351289	8760422	14057376
2028	11462846	1539715	8445059	14480632
2029	11506629	1721659	8132240	14881018
2030	11542164	1896972	7824167	15260161
2031	11571005	2065717	7522274	15619736
2032	11594412	2228085	7227445	15961379
2033	11613409	2384346	6940178	16286641
2034	11628828	2534809	6660694	16596961
2035	11641341	2679804	6389022	16893661

From the Table 4.6 the forecasted students' enrolment in Tanzania government primary schools from the year 2021 to 2035 shows the increase in enrolments in each year which will cause the increased in classroom infrastructures from 2021- 2035.

4. Conclusions

The study concludes that ARIMA (1, 1, 0) is the best fitted model to forecast enrolment in Tanzania government primary schools. Using ARIMA (1, 1, 0), the results show that there is an increase enrolment from 10,693,349 to 11,641 from 2021 to 2035.

It is recommended that further researcher can apply the same approach at different levels of education such as pre-primary and secondary to forecast student enrolments. The government and other educational stakeholders should invest more in improving the infrastructure on government primary schools due to increase in the number of students' enrolments each academic year. Also, the members of the community should be sensitized on the need for their contributions to school infrastructure in their localities.

Acknowledgements

We acknowledge Tanzania's President Office, Regional Administration and Local Government (PO-RALG) for their permission to provide secondary data for this study.

References

Bircan, T., & Sunata, U. (2015). Educational assessment of Syrian refugees in Turkey. *Migration Letters*, *12*(3), 226–237.

Greunen, J. Van, Heymans, A., & Vuuren, G. Van. (n.d.). *The stationarity of financial time series*. 1–17.

Idris, F., Hassan, Z., Ya'acob, A., Gill, S. K., & Awal, N. A. M. (2012). The Role of Education in Shaping Youth's National Identity. *Procedia - Social and Behavioral Sciences*, *59*, 443–450. https://doi.org/10.1016/j. sbspro.2012.09.299

Lavilles, R. Q., & Arcilla, M. J. B. (2012). Enrollment forecasting for school management system. *International Journal of Modeling and Optimization*, *2*(5), 563

Rutaremwa, G., Bemanzi, J., & Juliana, G. (2013). Inequality in School Enrolment in Uganda among Children of Ages 6-17 Years: The Experience after Introduction of Universal Primary Education UPE. *Science Journal of Education*, *1*(4), 43–50.

Kornelio, J. D. (2017). Forecasting membership enrollment of National Health Insurance Fund (NHIF): a case study of Dodoma region. Dodoma: The University of Dodoma http://hdl.handle.net/20.500.12661/495

Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. *Procedia Computer Science*, *48*, 173–179.

Msuya, E. E. (2009). *Unlocking smallholders' potentials in Tanzania: value chain and other analyses: a case study of maize in Kiteto and Mbozi districts*.
