INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY: APPLIED BUSINESS AND EDUCATION RESEARCH
2023, Vol. 4, No. 6, 2094 – 2100
http://dx.doi.org/10.11594/ijmaber.04.06.32

Research Article

Forecasting Dropout Trend at King's College of The Philippines using ARIMA Modeling

Ruben M. Gambulao Jr.*

Philippines

Article history:
Submission November 2022
Revised June 2023
Accepted June 2023

**Corresponding author:*
E-mail:
rmgambulao@up.edu.ph

ABSTRACT

Student drop-outs continue to be one of the perineal problems of educational institutions. Accordingly, institution managers are trying to find ways and means to curb impending issues on drop-outs to satisfy quality education. In this paper, the researcher delved into the different time series modeling methods in order to forecast the rate of college dropouts at King's College of the Philippines-Benguet. The method considered was the Autoregressive Integrated Moving Average (ARIMA) model. The data used in this paper was the number of school dropouts from 2005 to 2018 obtained from the school registrar which shows more dropout during the first semester than the second semester. Initial result obtained from using ARIMA reveals that the best model used is the model ARIMA which is the auto regression (AR 1), then the moving average (MA 1), with first differencing on the second semester.

Keywords*: ARIMA, Differencing, Forecast*

Introduction

Education is the process of teaching, training and learning in schools and colleges for the development of knowledge and skills in building a strong foundation in one's life (Adedeji,2016). One of the most significant current discussions in every culture is education being a dynamic mechanism which is used in the contemporary world to succeed. The educational reality today is nothing like centuries ago as school dropout has become an enduring issue worldwide and often viewed as a single event in which a student simply stops attending school one day. In the Philippines, it is becoming increasingly difficult to ignore that despite the efforts to increase access to higher education dropout rate behavior incurred 6 percent for elementary level and 11 percent for secondary (Orbeta, 2010).

In the case of King's College of the Philippines (KCP), the incidence of dropout is apparent to be higher during the first semester than the second semester of each academic year as indicated in an initial insight from the 28 semestral observations obtained from the Registrar's Office. Hence, only 28 available observations were applied as data for this study. In a real application, one usually does not know

How to cite:
Gambulao Jr., R. M. (2023). Forecasting Dropout Trend at King's College of The Philippines using ARIMA Modeling. *International Journal of Multidisciplinary: Applied Business and Education Research*. *4*(6), 2094 – 2100. doi: 10.11594/ijmaber.04.06.32

which procedure fits the data best, thus a better idea is to adaptively choose a modeling procedure (Zhang,2009). An ARIMA model may not fit the data well, which may result to bigger errors. An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends (Chen, 2019). ARIMA model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities by examining the differences between values in the series instead of through actual values. This research study employed differencing to analyze the patterns in the rate of dropouts and predict the rate for the next years. ARIMA modeling is a type of univariate analysis wherein the procedures analyze and forecast equally spaced univariate time series data, transfer function data, and intervention data (Pankratz, 1983). It predicts a value in a response time series as a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series (Nau, 2014). A common obstacle in using ARIMA models for forecasting is that the order selection process is usually considered subjective and difficult to apply. Identifying an appropriate state space model for a time series is also not an easy task (Ramos, et.al, 2016). A minimum size for the training set is specified.

Methods

This research made use of the ARIMA model. In the case of the King's College of the Philippines, the data gathered were treated using said model. Construction of an adequate ARIMA model requires a minimum of about 50 observations. An ARIMA model tells how observations on a variable are statistically related to past observations on the same variable and extrapolates past patterns within a single data series into the future. One of the most versatile linear models for forecasting seasonal time series is the ARIMA. ARIMA can be understood by outlining each of its component as follows: Autoregression (AR) refers to a model that shows a changing variable that regresses on its own lagged, or prior values. Integrated (I) represents the differencing of raw observations to allow for the time series to become stationary, and the Moving Average (MA) which incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations. It contributed a lot on the academic research and industrial applications. The class of ARIMA models is broad. It can represent many different types of stochastic seasonal and non-seasonal time series such as pure autoregressive (AR), pure moving average (MA), and mixed AR and MA processes (Ramos, et.al,2016). The theory of ARIMA models has been developed by many researchers and its wide application (Box et al. 2008) who developed a systematic and practical model building method.

The initial phase in applying ARIMA approach is to check for stationarity. "Stationarity" implies that the series remains at a fairly constant level over time (Zhang, 2009). This is effectively observed with an arrangement that is intensely regular and developing at a quicker rate. Without stationarity conditions being met, huge numbers of the estimations related with the procedure cannot be processed. The time series plot provides evidence that the series has a non-stationary mean. The significant lags do not fall below the 0.3 warning level until lag 12. To eliminate the significant lags, AR1 with first differencing of seasonality 2 was applied. RMSE, AIC, and BIC were the standard measures used to check the errors.

Results and Discussions

Time Series Plot of Student's Dropout

A total of 28 semi-annual observations from 2005 - 2018 was used to investigate different ARIMA models. The line graph of data in Figure 1 shows the semestral observation of the college dropouts whose overall mean is trending upward through time. The level of the time series appears to rise and fall episodically rather than trending in one direction (Ramos et.al, 2016). The behavior of the data within the period has a nonstationary mean. To deal with this, differencing was applied to render stationarity.

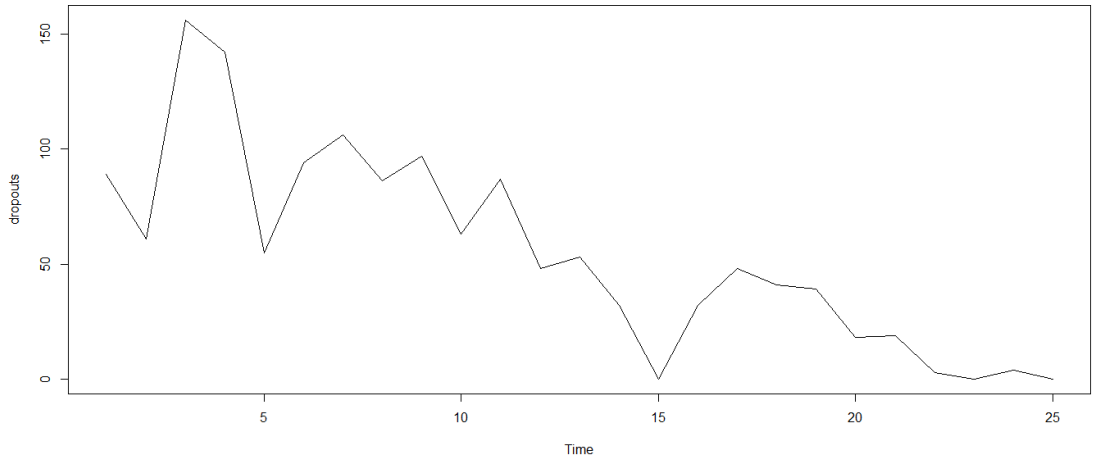


Figure 1. Time series plot of student's dropout

Based on the result of the behavior of the estimated autocorrelation function (ACF) and partial autocorrelation function (PACF), the Auto Regression (AR) was chosen over the Moving Average (MA). This is due to the presence of lots of significant spikes in the ACF at lags 1, 2, and 3. Observe that lags 1,2 and 3 exceeds the blue line as an indication of AR over MA. Based on the significant spikes in Fig. 2a, the data is non-stationary since the estimated ACF drops off slowly toward zero. Therefore, ARIMA model fits the data.

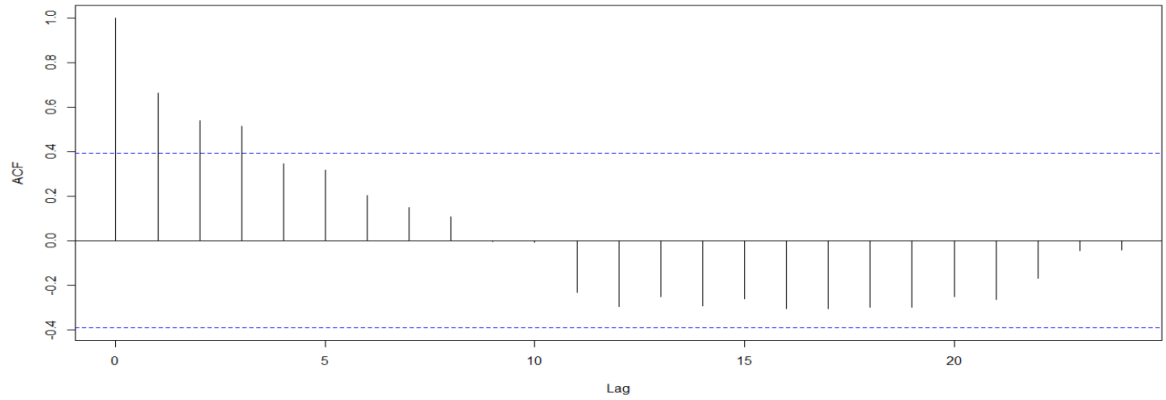


Figure 2a. Auto Correlation Function (ACF)

Figure 2a measures the self-similarity of the signal over different delay times. It defines how the data points are related on average, to the preceding data points. In relation to figure 1, it shows that the mean of the second semester is non-stationary. Meaning, there is rise and fall of the spikes as presented, indicating a non-stable trend of dropouts. Because of this non-stationary characteristic, AR (1) was chosen to represent the realization for some reasons: (1) it fits the available data (the past) well enough; (2) it has an acceptable set error measurement especially for root mean squared error (RMSE) and mean absolute error (MAE); (3) It forecasts the future satisfactorily as shown in Figure 5.

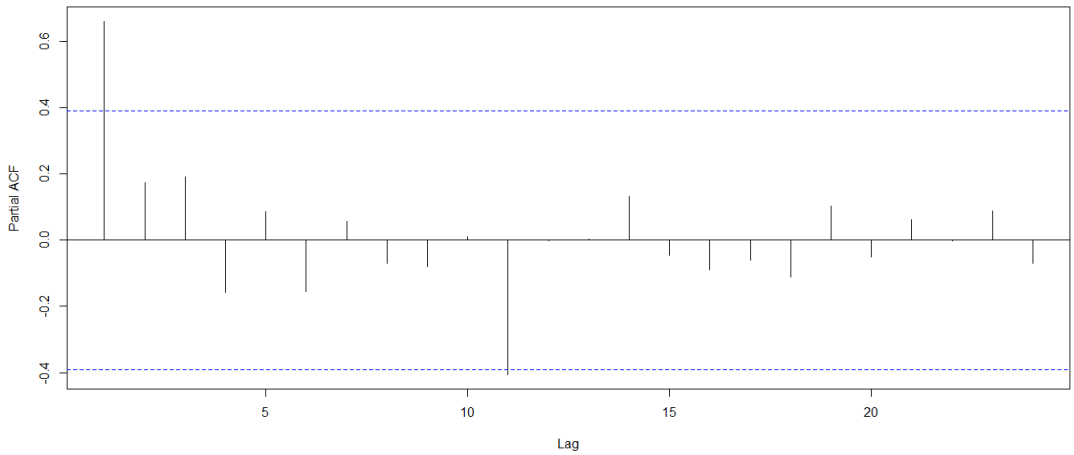


Figure 2b. Partial Auto Correlation Function (PACF)

Because of the case previously presented in figure 2a, figure 2b suggests the use of first differencing with seasonality. This first differencing with seasonality will solve the problem on lag 1, which is evidently above the blue line (warning level), indicating the lag as a critical value.

Data Related to College dropouts

Table 1 shows the data of college dropouts during the first and second semester that were considered in this study. We can easily observe that there were many dropouts recorded during the first semester with 2016 as the Academic Year registering the highest dropout of 156 students. However, it is also during this semester where there were zero dropouts specifically in the AYs 2005, 2006, and 2010. In the case of the second semester, the data revealed that there was a minimal number of dropouts. These data was simulated using R package (statistical software specifically used); thereby producing the time series plot of drop outs previously presented in figure 1.

Table 1. Data of college dropouts

Year	1st Semester	2nd Semester
2005	0	4
2006	0	4
2007	19	18
2008	39	41
2009	48	32
2010	0	33
2011	53	48
2012	87	64
2013	97	86
2014	106	94
2015	55	122
2016	156	61
2017	81	100

Table 2. Possible ARIMA models

Model	AIC	BIC	RMSE	MAE
ARIMA(1,1,0)(0,1,1)2	249.79	255.14	3.735771	2.763606
ARIMA(3,1,0)(0,0,0)2	258.19	265.42	3.748212	2.830768
ARIMA(0,0,1)(0,1,1)2	252.58	257.94	3.872928	2.755538

To arise with a good result, three possible models were computed and compared to each other. The result revealed that the first differencing with seasonality stipulated firstly in table 2 was considered. This was the case because according to (Nau, 2014) the measures with the least value of Akaike Information Criterion (AIC) is considered to be the best model. Meaning, the model ARIMA (1,1,0) (0,1,1)2 has the smallest AIC value of 249.79, qualifying it as the appropriate model to be utilized.

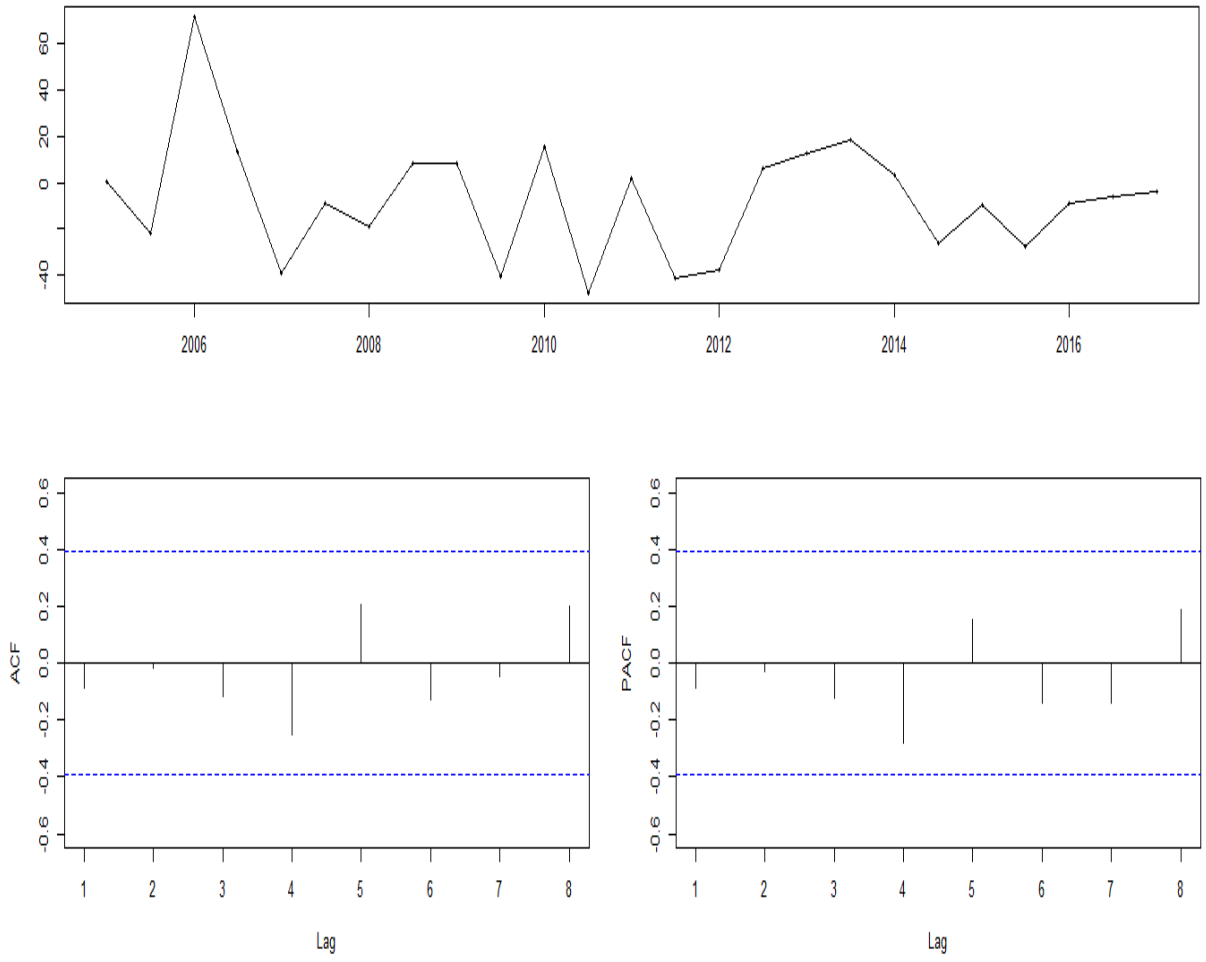


Figure 3. Estimation and diagnostic-checking results for ARIMA (1,1,0) (0,1,1) [2]

Figure 3 presents the residual of the autocorrelation function and partial autocorrelation function. It shows that the model is already Figure 4. Residuals of ACF and PACF adequate since there is no more occurrence of significant lags. Their absolute correlation is less than the practical warning level.

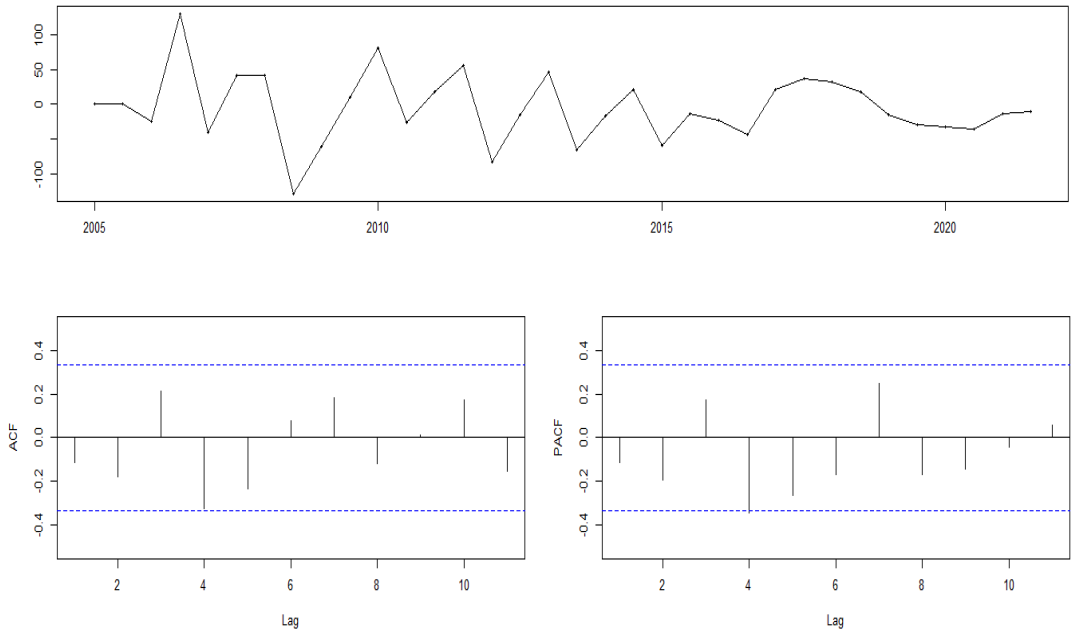


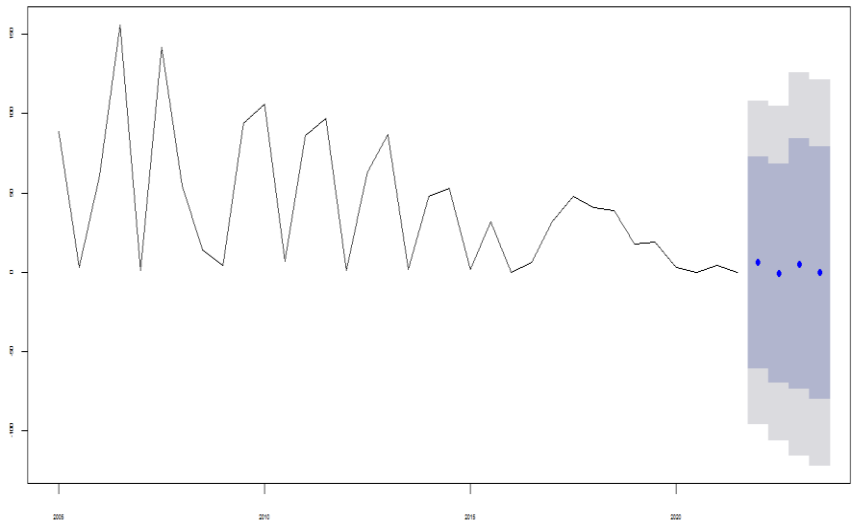
Figure 4 shows that the significant lags at 2a and 2b apparently disappeared, leaving a slight spike at lag 4 which is still manageable in selecting the best model.

Table 3. Forecasted values of AR1diff1

Year/Sem	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2022.00	6.0573959	-60.61919	72.73398	-95.91564	108.0304
2022.50	-0.6903613	-69.75780	68.37708	-106.31990	104.9392
2023.00	5.2198917	-73.78372	84.22350	-115.60571	126.0455
2023.50	-0.3971437	-80.07759	79.28331	-122.25788	121.4636

Table 3 displays the two-year forecasted value of ARIMA Model. The table shows that by school year 2023 of second semester, the drop-out will fall with a forecasted value of 5.22, or approximately 5 dropouts. Accordingly, the school year 2022 first semester will give the highest peak of dropouts with a forecasted value of 6.06, or approximately 6 dropouts.

FIGURE 5. Two-year forecast of ARIMA Model



Upon establishing from the forecasted value from ARIMA model, it can be noted that first semester of each academic year has higher dropout than on the second semester. Figure 5 shows the first blue dot as the first semester, the second blue dot represents the second semester, the third as the first semester and the last blue dot as the second semester. Based from figure 5 the first and third blue dot has increasing trend which means dropouts will likely more to occur.

Conclusion and Recommendations

ARIMA models were simulated and the most appropriate models were selected and used for the forecast of the number of dropouts in King's College of the Philippines.The number of dropouts in King's College of the Philippines was analyzed per semster. The behavior of the time series data in KCP was observed and the semester that receives the least and the greatest number of dropouts was determined. Analysis of number of dropouts in King's College of the Philippines showed that the least number of dropouts occurred during the second semester, while greatest number of dropouts occurred during the second semester. Based from the good results of the residual ACF and PACF of the models used, ARIMA (1,1,0) (0,1,1)[2]. The following conclusions are hereby preferred: ARIMA methodology also allows models to be built that incorporate both autoregressive and first differencing with seasonality together. These models are often referred to as "mixed models". Although this makes a more complicated forecasting tool, the structure may indeed simulate the series better and produce a more accurate forecast.

Based from the findings it is recommended that the forecasted value will be a basis in revisiting the student manual particularly in the dropouts. It is suggested also that the study will be a ground in encouraging the instructors and professors to strongly refer students that accumulated absences as per the student and teachers handbook. Initial interview from the Office of the Student Affairs (OSA) Director revealed that instructors and professor fails to report absences for intervention. The result will be utilized in making possible sanction/s for instructors and professors who neglect to follow the procedure. The researcher also suggests having a separate list of dropouts from the school registrar for easy access of the record. Overall to decrease students dropouts rate and encourage them to pursue their chosen course or even zero record of dropouts.

References

Adedeji,A. A.,2016. "Trend Analysis of Students Dropout Rate and the Effects on the Social and Educational Systems in Nigeria." International Journal of Latest Research in Engineering and Technology (IJLRET), ISSN: 24545031,Volume 2, Issue4.

A.C. Orbeta, A.C. Jr.,2010. Global Study on Child Poverty and Disparities: Philippines", NEDA, Makati.

Zhang,Y., 2009. "Cross-Validation for Selecting a Model Selection Procedure", Lundquist College of Business University of Oregon.

Pankratz,A., 1983 "Forecasting with Univariate Box-Jenkins Models," John Wiley & Sons, Inc.

Nau, R, 2014. "Introduction to ARIMA Models", Fuqua School of Business Duke University.

Ramos, P. et.al, 2016. "A Procedure for Identification of Appropriate State Space and ARIMA Models Based on Time-Series Cross-Validation", Algorithms, Vol.9, Issue 4.

Shao, J, 2013. "Linear Model Selection by Cross-Validation", Journal of the American Statistical Association, Vol. 88, No. 442.

Box, G., et.al, 2008.Time Series Analysis, 4th ed.; Wiley: Hoboken, NJ, USA.

Hyndman, R.J., 2006. Koehler, A.B. Another look at measures of forecast accuracy. Int. J. Forecast.

Ugiliweneza,B."Use of ARIMA Time Series and Regressors to Forecast the Sale of Electricity", University of Louisville, Louisville KY. 11. R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Chen, J., 2019. What Does Autoregressive Integrated Moving Average? Available at: https://Investopedia.com/