Nigerian Research Journal of Engineering and Environmental Sciences 8(1) 2023 pp. 206-213
p ISSN: 2635-3342; e ISSN: 2635-3350

Nigerian Research Journal of Engineering and Environmental Sciences
Journal homepage: www.rjees.com

Original Research Article

Forecasting Internet Bandwidth Demand for University of Benin, Nigeria

***[1]Dele-Ogbeide, O. and [2]Oladeinde, M.H.**

[1]Department of Computer Engineering, Faculty of Engineering, University of Benin, PMB 1154, Benin City, Nigeria.
[2]Department of Production Engineering, Faculty of Engineering, University of Benin, PMB 1154, Benin City, Nigeria.
*osaosemwen@uniben.edu; mobolaji.oladeinde@uniben.edu

http://doi.org/10.5281/zenodo.8094924

ARTICLE INFORMATION

Article history:
Received 04 May 2023
Revised 16 Jun. 2023
Accepted 17 Jun. 2023
Available online 30 Jun. 2023

Keywords:
UNIBEN
Bandwidth
Internet
Demand
ARIMA

ABSTRACT

The demand for internet service has always been on the rise especially with the advent of new technological devices and the current information age. In this study, the data showing internet bandwidth consumed daily for staff and students of the University of Benin was considered based on maximum demand. The internet bandwidth data was chronologically harvested for 370 days and used to predict internet bandwidth demand. Data was examined for stationary and model fitness using autocorrelation function (ACF) and partial autocorrelation function (PACF) tests. Several ARIMA models were considered for predicting the demand as well as an outlier detection approach and the data was split in two for training and testing the model. The training data consisted of 200 data points while the testing had 160 data points. The result obtained showed that there were 13 outliers present in the data and the seasonal ARIMA(0,0,2)(0,1,1)7 was most suited with the stationary R^2 of 0.959, R^2 value of 0.957, root mean square error (RMSE) of value of 15.296, mean absolute error(MAE) of 10.852 and the normalized Bayesian information criterion (NBIC) score of 5.731.

1. INTRODUCTION

Bandwidth is the amount of information per unit time that a transmission medium (internet connection) can handle as defined by Yildirim et al. (2023). In some parts of the world, internet is a scarce resource and needs to be managed. Fredrick and Jan (2014) reported that developing countries' internet resources are limited due to financial constraints. Despite these constraints, internet service providers and equipment manufacturers usually forecast the bandwidth subscribers need in order to match their needs with the future requirements. According to Barnett et al. (2018), equipment manufacturers like Cisco predicted internet bandwidth from 2015 to 2020. This ensures the proper planning and preparation be done on internet service in order to meet users future demand.

The University of Benin's (UNIBEN) total internet bandwidth is 150 Mbps capacity serving the campuses, "list of countries with internet speed" shows the world standard for internet usage is 3.9 Mbps but the UNIBEN uses 1 Mbps for academic staff, 600 kbps for non-academic staff and 400 kbps for students, which is below the world standard, as the growth of internet devices and subscribers on the campuses increase, there will be increased dissatisfaction in service quality.

The website and internet service are two ingredients in a university system that requires monitoring. Every minute, if either or both of the services fails, the cost of subscription paid by the university to host the website and internet bandwidth from the service provider is wasted. It also affects research by delaying sourcing for materials, collaboration with internal and external researchers amongst others.

This study is set to research the factors affecting UNIBEN website and internet service which can aid improvement, services enhancement and the University's reputation in Nigeria.

2. MATERIALS AND METHODS

2.1. Data Collection

The data for this research was collected using a Network Monitoring System (NMS) called Libre NMS. It was installed on the same server that distribute internet bandwidth to departments, faculties and halls of residence on both campuses of University of Benin. The maximum internet bandwidth usage was considered for this research because it shows usage at peak periods demands of also reflecting the genuine extent of consumption. A screenshot of the NMS is shown in Figure 1. Internet bandwidth data for the period of March 2, 2016 to March 6, 2017 was manually collected for each day for 370 days.

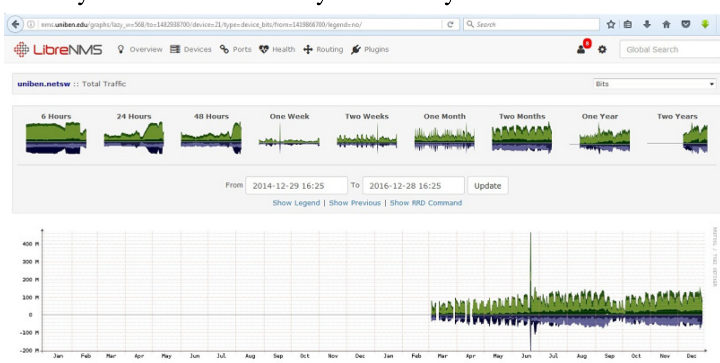


Figure 1: Graphic showing the Libre NMS

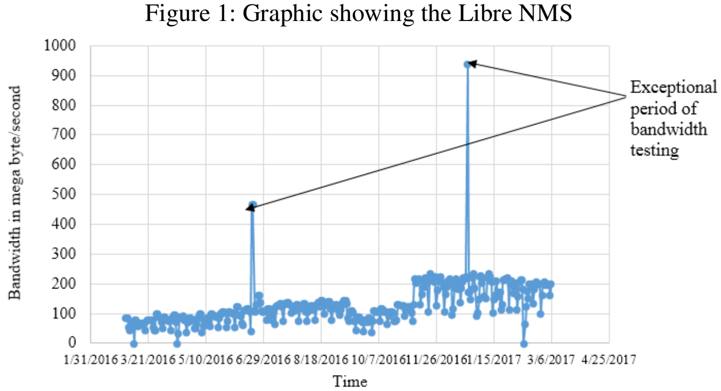


Figure 2: Plot showing the maximum internet bandwidth consumption from March 2, 2016 to March 6 2017

2.2. Methods for Forecasting Internet Bandwidth

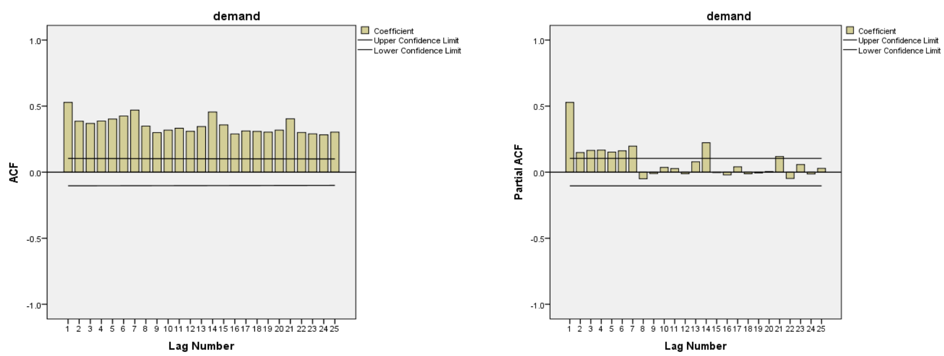
The internet bandwidth data upon collection from the NMS was examined to determine an appropriate method to forecast internet bandwidth demand. This required testing the data for Stationarity and determining the standard errors of each lag. The stationarity tests were carried out to check data patterns by employing the mean, variance and autocorrelation. These tests are of two types which are autocorrelation function (ACF) for examining the data for relationship between present data point y_t and the previous values y_{t-1}. The model for representing ACF in Equation 1.

$$ACF = r_n = \frac{\sum_{i=0}^{n}(y_n - \mu)(y_{n-1} - \mu)}{\sum_{i=0}^{n}(y_n - \mu)^2} \tag{1}$$

Where y_n = data at time n and μ = is the mean of the data

Another stationarity test is partial autocorrelation function. This is the correlation between a variable and a lag of itself with the absence of other lags. The name partial explains that it considers the correlation only between internet demand y_t and the lagged variable of interest $y_{t-1} \ldots y_{t-n}$ (Equation 2). The ACF and PACF computations are plotted to show diagrammatically the patterns of the observation in Figure 2.

$$PACF(n) = \frac{Co\,\mathrm{var}\,iance(y_n, y_m | y_{n-1} \ldots y_{n-m})}{\sqrt{\mathrm{var}\,iance(y_n | y_{n-1} \ldots y_{n-m})\,\mathrm{var}\,iance(y_m | y_{n-1} \ldots y_{n-m})}} \tag{2}$$



(a) (b)

Figure 2. An example of the ACF (a) and PACF (b) plots

These plots in addition to the Equations 1 and 2 guided in identifying the stationarity of the model and selecting the preferred ARIMA model to use. Each plot consisted of two parts which are upper part and the lower part. The upper part has a positive side of the plot where the upper confidence limit of the plot is indicated while the lower part is the negative side of the plot with the lower confidence limit. This confidence limit is determined with the Equation 3.

$$confidence\ interval = \pm 1.96\,\widetilde{\widetilde{\emptyset}} \tag{3}$$

Where $\widetilde{\widetilde{\emptyset}}$ is an estimate of standard deviation and J is a jth step forecast distribution

The standard error (SE) of the ACF and PACF was also considered. According to Ke and Zhiyong et al. (2018), the standard error is the standard deviation of the sampling distribution mean. It was used to determine the margin of errors in representing a population as well as the accuracy of a data set. The SE is given in Equation 4.

$$SE\, r_k = \sqrt{\frac{1}{n}\frac{(n-k)}{(n+2)}} \tag{4}$$

where SE r_k is standard error of the mean, *n* is the sample size and *k* is the lag

This seasonal auto regression integrated moving average (SARIMA) model is suited for observations that follow seasonal pattern was also employed in predicting UNIBEN's internet bandwidth. According to Arumugam and Saranya (2018) the SARIMA is a multiplicative model written as $ARIMA(p,d,q)(P,D,Q)m$ process. The SARIMA is displayed in Equation (5).

$$\Phi(B^m)\phi(B)\,\nabla^D_m \nabla y_t = c + \Theta(B^m)\theta(B)e_t \tag{5}$$

Where B is a backshift operator = $(1-B)$, et is the white noise process, Yt is the observed variable, C is a constant, $\nabla^D_m\ y_t = y_t - y_{t-m}$ is the seasonal difference, $\nabla y_t = y_t - y_{t-1}$ is the non-seasonal difference, $\Phi(B^m) = 1-\Phi_1 B^m - \cdots - \Phi_P B^{PM}$, $\phi(B) = 1-\phi 1 B_p - \cdots - \phi B_p$, $\Theta(B^m) = 1+\Theta_1 Bm + \cdots + \Theta_Q B^{Qm}$ and $\theta(B) = 1+\theta_1 B + \cdots + \theta_q B_q$

2.3. Outliers

Outliers are observations that differs from original data pattern. They have an ability to affect the ARIMA model which can result to an over fitted model (Arumugam and Saranya, 2018). These observed outliers in the data are first identified to determine the location before estimation using Equation (6).

$$Y(t) = \mu(t) + \sum_{k=1}^{m} \omega_k L_{0k}(B) I_{Tk}(t) + \frac{\theta(B)}{\Delta\varphi(B)}\, a(t) \tag{6}$$

Where $\mu(t)$ is the ARIMA series, y(t) is the observed series with outliers, m is the number of outliers, ω is the magnitude of the outlier, B is a backshift operator, a(t) is white noise series normally distributed, at time $\varphi(B)$ is an auto regression polynomial, $\theta(B)$ is a moving average polynomial and `I_{Tk} =is an adding function when t is 0 or 1.

Such that $L_0(B)$

in additive outlier (AO)=1,

$$\text{Innovative outlier (IO)} = \frac{1}{(\Delta\pi(B))} \quad \text{with } \boldsymbol{\pi} B = \frac{\theta(B)}{\varphi(B)} \tag{7}$$

$$\text{Level shift (LS)} = \frac{1}{(1-B)} \tag{8}$$

$$\text{Transient change (TC)} = \frac{1}{(1-\delta B))} \tag{9}$$

$$\text{Seasonal additive (SA)} = \frac{1}{(1-Bs)} \tag{10}$$

$$\text{Local trend (LT)} = \frac{1}{(1-B2)} \tag{11}$$

at (t = 1, 2, ... , n)

Outliers presence were tested using the statistic in Equation (12).

$$e(t) = \omega x(t) + a(t) \tag{12}$$

where et is residual

For j = 1 in (AO), 2 (IO), 3 (LS), 4 (TC), 5 (SA), 6 (LT) outliers the defined test statistics

$$\lambda_j(T) = \frac{\omega j(T)}{\sqrt{Var(\omega j(T))}} \tag{13}$$

Under the null hypothesis of no outlier, $\lambda_j(T)$ is distributed as N(0,1) assuming the model and model parameters are known.

2.4. Testing the Model

In every ARIMA or SARIMA process, the model was tested alongside the residuals of moving average process and evaluated for the fitness. This tells how well the residuals are predicted with the selected model. The error tests used in this study are according to Zhang et al. (2013) and are described in the preceding.

Stationary R- squared is an error test that compares a stationary part of the model to a simple mean model. When the value is negative, it means that the model under consideration is worse than the baseline model. This model is preferred to R^2 when there are seasonality and trend in data. The stationary R^2 as seen in Equation 15.

$$R_s^2 = 1 - \frac{\sum \left(\mathcal{Z}(t)\right) - \left(\hat{\mathcal{Z}}(t)\right)^2}{\sum \left(\Delta\mathcal{Z}(t)\right) - \left(\Delta\bar{\mathcal{Z}}(t)\right)^2} \tag{14}$$

Where $\bar{\mathcal{Z}}$ is the mean of the actual data, $\mathcal{Z}(t)$ is the actual data value and $\Delta\mathcal{Z}$ is the simple mean of the differenced transformed series

R- Squared. This is the goodness of fit of a linear model sometimes called coefficient of determination. It is the proportion of variation in the independent variable explained by the regression model. Small values indicate the model does not fit the data well. The R-squared model is shown in Equation 15.

$$R^2 = 1 - \frac{\sum \left(\mathcal{Z}(t)\right) - \left(\hat{\mathcal{Z}}(t)\right)^2}{\sum \left(\mathcal{Z}(t)\right) - \left(\bar{\mathcal{Z}}(t)\right)^2} \tag{15}$$

Where $\hat{\mathcal{Z}}_t$ is the predicted value at t

Mean absolute error (MAE) measures how much the series varies from its predicted level. It is determined using the Equation 16.

$$MAE = \frac{1}{n}\sum \left| y(t) - \hat{\mathcal{Z}}(t) \right| \tag{16}$$

where n = number of residuals that are not zero

Root mean square error (RMSE) is known as root mean square error. It announces how data is focused around the line of best fit and it is based on the standard deviation of prediction errors. RMSE is presented in Equation 17.

$$RMSE = \sqrt{\frac{SSE}{dfe}} \tag{17}$$

where SSE is sum of square error (residual) and dfe is the degree of freedom. *f* is described as forecast (predicted value) and O is the observed values.

Mean absolute percentage error (MAPE) is a measure of how much a dependent series varies from its model predicted level. It's independent of its unit. The smaller the RMSE and MAPE the better the model. The mean absolute percentage error is shown in Equation 18.

$$\text{MAPE} = \frac{100}{n}\sum \left| \left(\mathcal{Z}_i - \hat{\mathcal{Z}}_i\right) / \mathcal{Z}_{(t)} \right| \tag{18}$$

Normalized Bayesian information criterion (NBIC) attempts to account for model complexity by penalizing models that tend to over fitness. The NBIC is shown in Equation 19.

$$NBIC = \ln(\text{MSE}) + \text{k}\frac{\ln(n)}{n} \tag{19}$$

Ljung Box Pierce test examines randomness of the residual error in the model whether any group of autocorrelation is different from zero. The Ljung–Box test, uses a hypothesis and may be defined as:

H_0: The data are randomly distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any correlations in the data result from randomness of the sampling process).

H_a: The data are not randomly distributed; they exhibit serial correlation.

The Lung-box pierce test is written in Equation 21 as

$$Q = n(n+2)\sum_{k=1}^{h} \frac{pk}{n-k} \tag{20}$$

where n is the sample size, pk is the sample autocorrelation at lag k, and h is the number of lags being tested.

2.5. Data Splitting

The internet demand data was split into 2 parts namely A and B. (Reitermanova, 2010). The part A consisting of 200 data values from March 3, 2016 to September 17, 2016 was used to develop a prediction model while part B consisting of 140 data values from October 9, 2016 to March 6, 2017 was employed to validate the developed model. The benefit of data splitting comes to light when avoiding overconfidence of a forecast model. According to LeBaron and Andreas (1998), regardless of the model adequacy test and fitness carried out with other models, data splitting helps to affirm the selected model by pointing out the extent of discrepancies in prediction. This can be easily achieved with cross validation.

2.6. Validation of Forecast Model

After developing a prediction model on part A of the data, the predicted model was employed to the part B of the data (Bergmeir et al., 2014) in order to validate the predictability of the preferred model. Model adequacy and fitness test were used to ascertain the preferred model. The result of the model was compared with part A and part B of the data.

3. RESULTS AND DISCUSSION

3.1. Summary of Forecasting Models

Aside the test conducted on the examined models which include RMSE, MAE, Stationary r^2, and NBIC to ascertain model fitness as well as adequacy, other test like Ljung Box pierce, MAXAPE, MAPE and MAXAE tests were also carried out. Table 1 shows a summary of the models employed in predicting the internet bandwidth demand. It was found from the error and model fitness tests conducted that ARIMA (002) $(011)^7$ had the most suited scores and based on that, it was further employed in forecasting UNIBEN's internet bandwidth for the next 5 years.

Table 1: ARIMA models and the fitness values observed

Model	Stationary R^2	R^2	RMSE	MAPE	MAE	Normalized BIC	Ljung-Box Q(18) statistics	DF	Sig.
ARIMA	0.356	0.356	56.278	247.458	25.209	8.141	34.28	16	0.005
ARIMA	0.359	0.359	56.328	255.756	25.104	8.174	31.363	14	0.005
ARIMA213	0.36	0.398	56.418	320.633	25.194	8.194	28.071	13	0.009
ARIMA214	0.374	0.41	55.896	294.662	24.944	8.191	23.699	12	0.022
ARIMA312	0.361	0.361	56.314	246.868	25.085	8.19	30.872	13	0.004
ARIMA	0.36	0.36	56.433	249.873	25.18	8.21	29.307	12	0.004
ARIMA	0.385	0.385	55.4	296.635	24.133	8.189	26.276	11	0.006
ARIMA	0.362	0.362	56.342	249.47	25.028	8.207	31.155	12	0.002
ARIMA	0.36	0.36	56.526	244.348	24.926	8.229	33.107	11	0.001
ARIMA	0.376	0.376	55.892	273.011	24.446	8.223	24.162	10	0.007
ARIMA	0.407	0.407	54.8	316.038	22.715	8.247	13.118	6	0.041
ARIMA	0.462	0.439	54.473	369.567	22.35	8.109	21.187	14	0.097
ARIMA	0.414	0.414	53.652	354.697	22.281	8.029	9.828	14	0.775
ARIMA	0.65	0.445	53.977	343.93	21.34	8.091	9.246	14	0.815
ARIMA	0.959	0.957	15.296	48.029	10.852	5.731	24.679	15	0.024

3.2. Data Splitting

The split data had more values for training the model than for validating the model. This was employed following the holdout cross validation method for splitting data. The outliers identified when modeling with the first part of the split data are outlined in the Equation 21.

$$Z_t = -70.723e_t^{(9mon)} - 43.396e_t^{(11fri)} - 57.920e_t^{(14wed)} + 372.028e_t^{(23fri)} + 375.549e_t^{(23sat)} + 49.485e_t^{(24thu)} \quad (21)$$

Where Zt= residual, $e_t^{(week,\ day)}$ = error term in a specified week and day

In validating the ARIMA $(0,0,2)(0,1,1)^7$ model, the 170 data values from the second part of the split data was employed. The outliers observed are estimated as well as the standard errors. The outliers are presented in the Equation 22.

$$Z_t = -42.586e_t^{(25sat)} + 85.921e_t^{(34sun)} + 713.432e_t^{(40thu)} - 175.635e_t^{(47thu)} + 86.682e_t^{(50fri)} \quad (22)$$

Where Zt= residual, $e_t^{(week,\ day)}$ = error term in a specified week and day

3.3. Forecast Model

The developed forecast model ARIMA $(0,0,2)(0,1,1)^7$ can be written as equation 23

$$\Delta^7 yt = \varphi_{1q}\varepsilon_{t-1} + \varphi_{2q}\varepsilon_{t-2} + \theta_{1Q}{}^7\varepsilon_{t-1}{}^7 + c \quad (23)$$

Where Δ^7yt =seasonal differenced demand at 7th period, φ_{1q}= nonseasonal regression parameter for first moving average, ε_{t-1}= nonseasonal error term at t-1 period, ε_{t-2}= non seasonal error term at t-2 period, φ_{2q}= nonseasonal regression parameter for second moving average, $\theta_{1Q}{}^7$= seasonal regression parameter for first moving average following 7th seasonal period, $\varepsilon_{t-1}{}^7$ = seasonal error term at t-1 following the 7th seasonal period and c= constant

3.4. Internet Bandwidth Demand Prediction for 5 years

A 5-year forecast of UNIBEN internet bandwidth demand is performed with ARIMA (0,0,2) $(0,1,1)_7$. The forecast is plotted in Figure 3. The abscissa of the plot shows the internet bandwidth and the ordinate shows the number of days (five years) of prediction. The blue line represented as "pred" is the prediction in the plot, orange line denoted as "lcl" is the lower control limit, the grey line known as "ucl" is the upper control limit of the plot while the yellow line known as "noise" is in sample forecast error. The actual prediction as seen in the figure is the blue line in the chart which in in between the upper control limit (ucl) and the lower control limit (lcl) of the forecast and it reveals from the 358 day, there was an increase in the internet bandwidth demand for the five years.

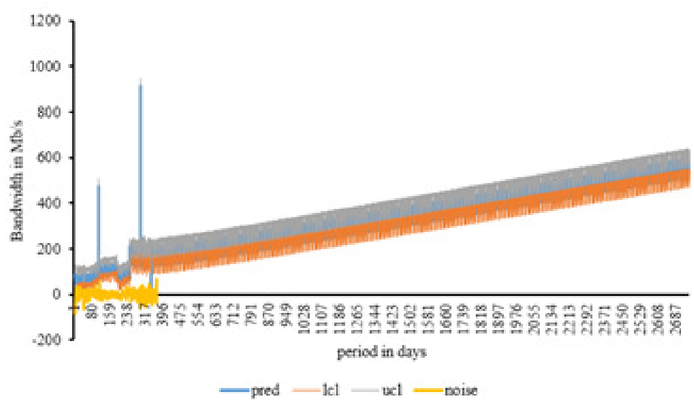


Figure 3. Internet demand prediction with ARIMA (0,0,2) $(0,1,1)^7$ model for five years

4. CONCLUSION

The forecast of UNIBEN internet demand bandwidth has shown that there will be increase in the demand for internet in the UNIBEN. The application of a seasonal ARIMA$(0,0,2)(0,1,1)^7$ out of other models tested showed the most fit based on an RMSE value of 15.296, stationary R^2 of 0.957, MAE of 10.852, NBIC of 5.731. This model predicted internet demand increase within five years between 2017 and 2022 with an average bandwidth demand from 134 Mb/s to 254Mb/s which can be translated to 89.55% increase within the stated period. In conclusion, it is not new to say that website and internet service have been sewn into the fabrics of Universities all around the world, the only astonishing factor is which University leads in the effective use of these services globally. If these models are employed on UNIBEN or any organization, there will be recorded improvements in their web ranking, website usability as well as internet service adoption.

5. ACKNOWLEDGMENT

The authors would like to recognize the support and contributions of the ICTU employees of the University of Benin, the Department of Production Engineering, and the entire University of Benin, Benin City, Nigeria.

6. CONFLICT OF INTEREST

There is no conflict of interest associated with this work.

REFERENCES

Arumugam, P. and Saranya R. (2018). Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data. *Materials Today*, 5(1), pp. 1791-1799.

Barnett, T., Jain, S., Andra, U. and Khurana, T. (2018). Cisco visual networking index (vni) complete forecast update, 2017–2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*, 1-30.

Bergmeir, C., Mauro C. and Jose M. B. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, 76(1), pp. 132-143.

Fredrick, K. S. and Jan E. (2014). Towards an instructional design model for learning environments with limited ICT resources in higher education. *African Educational Research Journal*, 2(2), pp 85-95.

Ke, Z. and Zhiyong Z. (2018). Testing autocorrelation and partial autocorrelation: Asymptotic methods versus resampling techniques. *British Journal of Mathematical and Statistical Psychology*, 71(1), pp. 96-116.

LeBaron, B. and Andreas, S. W. (1998). A bootstrap evaluation of the effect of data splitting on financial time series. *IEEE Transactions on Neural Networks*, 9(1), pp. 213-220.

Reitermanova, Z. (2010). Data splitting. *WDS'10 Proceedings of Contributed Papers*. 10(1), pp. 31-36,

Yildirim, R. and Hazer, A. (2023). A New Approach to Increasing the Bandwidth of Fiber-Optic Communication Systems. *Politeknik Dergisi*, pp. 1-1.

Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A. A. and Li, X. (2013). Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PloS one*, *8*(5), e63116.