Original Research Article

Trend analysis and ARIMA models for water quality parameters of Brahmani River

Abstract

Statistical trend analysis and time-series prediction model are widely used in water quality regulation. Using the Mann-Kendall test, trend analysis was performed on monthly time series. The monthly findings revealed that only the potential Hydrogen (pH) and Total Califorms (TC) showed meaningful trends. Future values for the parameters which affect water quality have been predicted using the Autoregressive Integrated Moving Average (ARIMA) model. R-square, root mean square error, absolute maximum percentage error, absolute maximum error, normalised Bayesian information criteria, Ljung-Box analysis were used to validate the model. It has been found that the predictive models for pH, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), and TC are useful at 95% confidence limits. Also, the results showed that the pH values will be in the range of 7.2 to 7.5 and the predicted series were similar to the original series, providing a perfect fit. The DO (mg/l) ranges from 7.8 to 12.3 mg/l. BOD (mg/l) fluctuates continuously between 1.2 and 1.3 mg/l. The TC (MPN/100ml) values will fall, which will be beneficial to the water in the Brahmani River.

Keywords: ARIMA, Mann-Kendall, Water Pollutants, Forecasting**.**

1. Introduction

The Brahmani River, is one of the most significant peninsular river systems in India, the second-largest river in Odisha, and that is where the research is being undertaken. This mighty river during the monsoon, basically becomes stagnant pools of water held in steep gorges and potholes in the riverbed. On the banks of the river, one of India's major industrialised areas known for ore mining, steel production, power generation, cement production, and other related activities. A large number of pollutants are deposited by neighbouring companies, municipalities, and villages, therefore the river's intrinsic capability

is unable to remove them, which is the primary reason why the water quality is deteriorating. These pollutants drain into the Bay of Bengal after passing through the southern districts of Sundergarh, Deogarh, Angul, Dhenkanal, Jajpur, and Kendrapara. Water is one of the most fundamental needs of the population, hence its safety needs to be considered before use. The current study seeks to determine the physico-chemical and bacteriological characteristics of the water quality across the Brahmani river.

In hydrology, analysing trends in water quality data is crucial for understanding how water quality parameters fluctuate, planning streams, and monitoring water quality measures. Understanding long-term fluctuations in certain water quality measures is one of six requirements for water quality monitoring [1]. Additionally, a trend analysis shows whether the measured values of a water quality metric have increased or decreased over time [2]. Gocic and Trajkovic [3] examined the trends of 12 indicators of the Nisava River's water quality from 2000 to 2004. In the current study trend analysis was performed on monthly time series data using Mann-Kendall test.

A time series is a group of data points that have been arranged chronologically. It is a collection of observations taken at a succession of equally spaced points in time *ie.*, discrete time data. These techniques analyse time series data to derive important statistics and data properties. Using a model and previously observed values, time series forecasting can be executed. Kurunc *et al*. [4] examined the Yesxilirmak River's stream flow and water quality parameters using time series analysis at the Durucasu monitoring station. Taheri *et al*. [5] used time series modelling to examine the quality of the Hor Rood River at the Kakareza station. The ARIMA (Autoregressive, Integrated, and Moving Average) model was found to be suitable for generating and forecasting river water quality.

2. Materials and Method

2.1 Study area

The water quality data for the Brahmani River (pH, DO (mg/l), BOD (mg/l,) and TC (MPN/100ml)) from 2017 to 2019 were gathered for the current study from the Pollution Control Board's official website in Bhubaneswar, Odisha (https://odocmms.nic.in). Four stations data on water quality were gathered *ie.*, Rourkela (D/S), Dhenkanal (D/S), Pattamundai, and Angul.

2.2 Mann-Kendall test for trend analysis

Mann [6] described a nonparametric test for randomness against the trend. The test he described is a particular application of Kendall's test for correlation commonly known as Kendall's tau [7]. According to Mann the null hypothesis of randomness H_0 states that the data (X_1..., X_1) are a sample of n independent and identically distributed random variables. The alternative hypothesis (H_1) of a two-sided test is that the distribution of X_k and Xj are not identical for all $k, j \leq n$ with $k \neq j$. The test statistic S is defined as:

$$S = \sum_{k=1}^{n=1} \sum_{j=k+1}^{n} sgn\,(x_j - x_k) \qquad (1)$$

$$Sgn(x_j - x_k) = \begin{cases} -1 & if\ (x_j - x_k) < 0 \\ 0 & if\ (x_j - x_k) = 0 \\ +1 & if\ (x_j - x_k) > 0 \end{cases} \qquad (2)$$

Where x_j and x_k are the annual values in different years j and k, j>k, respectively. If n<10 then the value of |S| is compared directly with the theoretical distribution of S that is derived by the Mann–Kendall test. At some probability level, H_0 is rejected in favour of H_1 if the absolute value of S equals or exceeds a specified value $S_{\alpha/2}$, where $S_{\alpha/2}$ is the smallest S having the probability less than $\alpha/2$. A positive (negative) value of S indicates an upward (downward) trend [8].

For n≥10, the statistic S is approximately normally distributed with the mean and variance as follows:

$$E(S) = 0$$

$$Var\,(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p - 1)\,(2t_p + 2)\right. \qquad (3)$$

Where, q is the number of tied groups and t_p is the number of data values in the p^{th} group. The standard test statistic Z is computed as:

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(s)}} & ;if\ S > 0 \\ 0 & ;if\ S = 0 \\ \frac{S+1}{\sqrt{Var(s)}} & ;if\ S < 0 \end{cases} \quad (4)$$

The presence of a statistically significant trend is evaluated using the Z value. A positive (negative) value of Z indicates an upward (downward) trend. To test for either an upward or downward monotonic trend (a two-tailed test) at α level of significance, H_0 is rejected if $|Z| > Z_{1-\frac{\alpha}{2}}$, where $Z_{1-\frac{\alpha}{2}}$ is obtained from the standard normal cumulative distribution tables. The Kendall's τ values are calculated as Eq. 5.

$$\tau = 2\frac{S^*}{z(z-1)} \quad (5)$$

In which S* denotes Kendall's sum, computed as S* = A-B, where A represents the number of chances when the difference of x_b to x_a is greater than zero and B represents the number of chances when the difference of x_b to x_a is less than zero.

2.3 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a statistical analysis model that uses the time series data to forecast future trends. It retains a form of regression analysis seeking to predict future movements and the random walks seemingly taken by examining the differences between values in the series instead of using the actual data values. The differenced series have lags referred to as "auto-regressive" and forecasted data lags are referred to as "moving average". This model is represented as ARIMA (p, d, q), where p represents the order of auto-regression, d shows the degree of differencing, q shows the order of moving average.

Inclusion of Autoregressive and moving average processes is more favour for achieving greater flexibility of actual time series data which starts to the combination of autoregressive and moving average processes denoted as ARMA (p,q).

ARMA (p,q) is indicated by

$$\emptyset(B)y_t = \theta(B)\varepsilon_t \quad (6)$$

where

$$\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots \emptyset_p B^p \quad (7)$$

and

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \quad (8)$$

In which,

B - the backshift operator express by $B(y_t) = y_{t-1}$.

p – order of AR

q – order of MA

Box-Jenkins [9] Autoregressive Integrated Moving Average model developed by including "differencing" in the ARMA model which indicated by ARIMA (p,d,q) which is written as

$$\Delta^d Y_t = C + \phi_1 \Delta^d Y_{t-1} + \cdots + \phi_p \Delta^d Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (9)$$

In which, $\varepsilon_t \sim N(0, \sigma^2)$.

2.4 Methods of Forecast Evaluation

The modelling capacity of several models was studied in this research using two typical performance metrics. They are the Root Mean Squared Error (RMSE) and the Root Mean Square Percentage Error (RMSPE). These are measured using the following equations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (10)$$

$$RMSPE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \hat{y}_t)^2}{y_t}} * 100 \quad (11)$$

Where,

- y_t – original price at t^{th} time
- $\hat{y}_t$ – forecasted price at t^{th} time
- n - the amount of forecasts

3. Result and Discussion

The table 1 depicts the summary statistics for average monthly values for Brahmani River. The pH levels, which range from 6.5 to 8.4, are within the acceptable range. The highest pH value was 8.4 in Dhenkanal (D/S), Pattamundai, and Angul, and the lowest was 6.5 in Rourkela (D/S). From 4.1 mg/l to 10 mg/l, the DO (mg/l) value is observed. The smallest DO value was 4.1mg/l in Rourkela (D/S), and the maximum DO value was 10 mg/l in Pattamundai. The BOD value lies between 0 and 6.5, with Rourkela (D/S) recording the highest (6.5 mg/l) and Pattamundai the lowest (0 mg/l). It is necessary to have a TC level of less than 5000 MPN/100 ml but the Brahmani River contains disturbingly high levels of coliforms, especially at Rourkela (D/S), where 92000 MPN/100 ml was observed with a minimum of 45 MPN/100 ml. The water quality parameters pH, DO, and BOD are all within acceptable limits, although TC deviates marginally.

The value of skewness is range from -0.44 to 1.69 and kurtosis is range from -0.53 to 3.15. For pH, skewness was -0.44, negatively skewed and kurtosis was -0.53 indicating platykurtic shape, respectively. For DO, skewness was 0.53, rightly skewed and kurtosis was -0.43, platykurtic shape. Skewness and kurtosis for BOD was -0.12 and -0.41, negatively skewed and platykurtic shape, respectively. For TC, skewness was 1.69, rightly skewed and kurtosis was 3.15 indicating leptokurtic shape respectively, indicating data were from a population with a non-normal distribution. Similar findings were found in previous relevant studies [10].

3.1 Trend analysis using Mann-Kendall test

The dataset does not follow a normal distribution, indicating a non-parametric form of the test. For both the average monthly data of the Brahmani River from 2017 to 2019 and station-specific monthly data for the four essential monitoring stations, the non-parametric Mann-Kendall test is utilised for trend detection. Table 2 offers Mann-Kendall's test with the calculated test statistics and the corresponding p-value. P-values below 0.05 are regarded as significant; at these values, the null hypothesis would be shown to be false. The absence of a trend in the currently available data is the null hypothesis for this inquiry.

The p values for BOD and TC at station Angul were below the significant level, indicating that a trend existed in the data. The Kendall values for BOD and TC were -0.203 and -0.362, respectively, indicating a decreasing trend in the data. The p values for the

remaining variables, pH and DO, were higher than the significant level, indicating that no trend existed. For the Dhenkanal station, pH and DO's p values are below the level of significance, and their Kendall values, which are -0.369 and 0.301, respectively, show that pH and DO are trending downward and upward, respectively. BOD and TC p values for the Pattamundai station were below the significant level, and their Kendall values were, respectively, -0.314 and -0.381, indicating that there is a trend in the data. However, pH and DO p values were above the significant level, indicating that there is no trend. The p values for pH and TC in the case of Rourkela were less than significant, indicating the existence of a trend. pH and TC both have Kendall values of -2.23 and -2.80, indicating a decreasing trend in the data.

Results of the Mann-Kendall test for the Brahmani River's average monthly data from 2017 to 2019 are shown in table 3. The p values of both pH and TC were significant, *ie.*, 0.011 and 0.001 respectively, which amply demonstrated the existence of a trend. Since pH and TC have Kendall values of -0.295 and -0.474, respectively, there is a downward trend in the data. There is no trend for BOD and DO since the p values were determined to be non-significant, 0.753 and 0.474 respectively. Similar results were also obtained in [11, 12] where Mann Kendall Trend Analysis accurately identifies the trends.

3.2 ARIMA

On the average water quality characteristics of the 2017–2019 data, the ARIMA time series model is fitted. The D-F test is used to assess the data using null hypothesis H_0, which indicates that there is a unit root in time-series data, and alternative hypothesis H_1, which indicates that there is no unit root, indicating that the time-series data is stationary. The calculated p-value for the D-F test for the time-series data for each water quality metric is larger than 0.05, which allows H_0 to be accepted and confirms that stationarity is attained by differencing. The p and q values were calculated using a plot of the ACF and PACF. Following that, estimates for RMSE, MAE, BIC, and R^2 were made for various combinations of the ARIMA. Based on the least value of RMSE, MAE, BIC, and R^2, the optimal model was chosen. Table 4 showed that all the chosen ARIMA models for the water quality parameters were highly significant since their fit statistic measures were significant, indicating that the models were well-fit. Therefore, the best fit for pH time series is ARIMA (1,1,1), followed by ARIMA (2,2,5) for DO time series, ARIMA (1,1,1) for BOD time series, and ARIMA (3,1,1) for TC time series. The Ljung-Box test is applied to the residuals of the

fitted models. The results showed that all P-values exceeds 0.05 which indicates acceptance of models accuracy at 95% significant levels (Table 5). The goodness-of-fit test of the optimum ARIMA models showed non-significant autocorrelations in the residuals of the model.

Table 5 lists the various ARIMA model variables for the Brahmani River's water quality factors. The time series analysis shown here includes the Ljung-Box, RMSE, MAPE, MAE, and BIC for the best-fit ARIMA models. The observed and anticipated values, along with upper and lower limits, were graphically shown for diagnostic verification of all the chosen models (Figure 1-4). This shows that pH forecasted values follow a decreasing tendency. Forecasted values for DO indicate an upward trend. Forecasted values for BOD show an ongoing trend, while those for TC show a downward trend. With the aid of the ARIMA model, the water quality values for the year 2020 were predicted with a 95% confidence interval shown in the Table 6, and it was discovered that the pH value will be between 7.5 and 7.2. The DO mg/l ranges from 7.8 to 12.3 mg/l. BOD mg/l fluctuates continuously between 1.3 and 1.2 mg/l. The rate at which oxygen is reduced decreases with decreasing BOD. The TC (MPN/100ml) number will drop, which will be good for the water in the Brahmani River [13].

4. Conclusion

The study described in the paper gives a statistical analysis of changes in the factors relating to water quality, and it also develops a model to forecast the concentrations of various indicators in the following years. The M-K test is used to analyse historical data on water quality. The results of the M-K test reveal that some water quality parameter data from various stations have a trend. The ARIMA model was used to make predictions and by evaluating the goodness of fit statistics, the ARIMA (1, 1, 1), ARIMA (2, 2, 5), ARIMA (1, 1, 1), and ARIMA (3, 1, 1) models proved to be the most effective at forecasting future water quality parameters. The water quality parameters pH, DO, and BOD are all within acceptable limits, although TC deviates marginally.

Reference

1. Kanungo S, Kumar Bhuyan N, Hemanta Kumar P. Assessment of the Water Quality Standard of Brahmani River in terms of Physico-Chemical Parameters. Int. J. Sci. Res. Manag. 2018; 6(4):50-7.

2. Antonopoulos VZ, Papamichail DM, Mitsiou KA. Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. Hydrology and Earth System Sciences. 2001 Dec 31; 5(4):679-92.
3. Gocic M, Trajkovic S. Trend analysis of water quality parameters for the Nisava River. Facta universitatis-series: Architecture and Civil Engineering. 2013; 11(3):199-210.
4. Kurunç A, Yürekli K, Cevik O. Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilırmak River, Turkey. Environmental Modelling & Software. 2005 Sep 1; 20(9):1195-200.
5. Taheri Tizro A, Ghashghaie M, Georgiou P, Voudouris K. Time series analysis of water quality parameters. Journal of Applied Research in Water and Wastewater. 2014 Mar 23; 1(1):40-50.
6. Mann HB (1945). Nonparametric tests against trend. *Econometrica* 245-259.
7. Kendall MG (1948). Rank correlation methods.
8. Eymen A, Köylü Ü. Seasonal trend analysis and ARIMA modeling of relative humidity and wind speed time series around Yamula Dam. Meteorology and Atmospheric Physics. 2019; 131(3):601-12.
9. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015). Time series analysis: forecasting and control. John Wiley & Sons.
10. Pal AB, Mishra PK. Trend analysis of rainfall, temperature and runoff data: a case study of Rangoon watershed in Nepal. International Journal of Student's Research in Technology & Management. 2017; 2321–2543, 5(3); 21–38.
11. Mustapha A. Detecting surface water quality trends using Mann-Kendall tests and Sen's slope estimates. International Journal of Agriculture Innovations and Research. 2013; 1:108-14.
12. Da Silva RM, Santos CA, Moreira M, Corte-Real J, Silva VC, Medeiros IC. Rainfall and river flow trends using Mann–Kendall and Sen's slope estimator statistical tests in the Cobres River basin. Natural Hazards. 2015; 77(2):1205-21.
13. Chaudhuri S, Dutta D. Mann–Kendall trend of pollutants, temperature and humidity over an urban station of India with forecast verification using different ARIMA models. Environmental monitoring and assessment. 2014; 186(8):4719-42.

Table 1: Descriptive Statistics

Parameters	pH	DO mg/l	BOD mg/l	TC MPN/100ml
Minimum	7.11	6.61	0.73	955
Maximum	8.15	8.48	2.08	27233.33
Mean	7.67	7.34	1.44	7113.28
SD	0.2725	0.5004	0.3444	5846.57
Kurtosis	-0.5339	-0.4331	-0.4118	3.1509
Skewness	-0.4424	0.5343	-0.1299	1.6987
CV	3.5506	6.8137	23.7695	82.1922

Table 2: Results of the Mann-Kendall test for Station wise monthly data from 2017 to 2019

Station	Water Quality Parameters	S	Z	Kendall-tau	p-value	Results
Angul	pH	4.3	0.57	7.011	0.5655	No Trend
	DO mg/l	-57	-0.76	-0.093	0.4431	No Trend
	BOD mg/l	-124	-1.68	-0.203	0.0022	↓ Trend exists
	TC MPN/100ml	-216	-2.97	-0.362	0.0029	↓ Trend exists
Dhenkanal	pH	-228	-3.10	-0.369	0.0019	↓ Trend exists
	DO mg/l	185	2.51	0.301	0.0118	↑ Trend exists
	BOD mg/l	2.50	0.33	4.182	0.4707	No Trend
	TC MPN/100ml	-177	-1.58	-0.188	0.1132	No Trend
Pattamundai	pH	6	0.068	9.678	0.9456	No Trend
	DO mg/l	-63	-0.846	-0.101	0.3975	No Trend

	BOD mg/l	-193	-2.62	-0.314	0.0086	↓ Trend exists
	TC MPN/100ml	-233	-3.17	-0.381	0.0014	↓ Trend exists
Rourkela	pH	-165	-2.23	-0.266	0.0251	↓ Trend exists
	DO mg/l	-133	-1.52	-0.181	0.1266	No Trend
	BOD mg/l	2.7	0.35	4.330	0.7229	No Trend
	TC MPN/100ml	-206	-2.80	-0.337	0.0049	↓ Trend exists

Table 3: Results of the Mann-Kendall test for the average monthly data

Water Quality Parameters	**S**	**Z**	**Kendall-tau**	**p-value**	**Result**
pH	-186	-2.52	-0.2957	0.0117	↓ Trend exists
DO mg/l	-24	-0.31	-0.0383	0.7538	No Trend
BOD mg/l	-57	-0.76	-0.0909	0.4453	No Trend
TC MPN/100ml	-299	-4.05	-0.4749	<0.001	↓ Trend exists

Table 4: Goodness of fit statistics of different ARIMA models for water quality parameters

Parameters	**Model**	R^2	**RMSE**	**MAPE**	**MAE**	**BIC**
pH	(1,1,0)	0.092	0.267	2.714	0.208	-2.435
	(1,1,1)	**0.201**	**0.255**	**2.686**	**0.205**	**-2.430**
	(1,1,2)	0.201	0.259	2.716	0.208	-2.298
DO mg/l	(2,2,4)	0.186	0.454	4.565	0.331	-0.855
	(2,2,5)	**0.190**	**0.461**	**4.555**	**0.330**	**-0.719**
	(2,2,6)	0.169	0.476	4.601	0.333	-0.555

BOD mg/l	(1,1,0)	-0.401	0.405	23.577	0.314	-1.605
	(1,1,1)	**-0.233**	**0.386**	**20.989**	**0.283**	**-1.600**
	(1,1,2)	-0.262	0.396	21.651	0.285	-1.444
TC MPN/100ml	(3,1,0)	0.188	5156.055	75.401	3405.190	17.502
	(3,1,1)	**0.256**	**5016.875**	**75.623**	**3343.543**	**17.549**
	(3,1,2)	0.255	5106.125	74.859	3368.776	17.686

Table 5: Best fit ARIMA models of water quality parameters at a 95% confidence interval

statistics	**pH**	**DO mg/l**	**BOD mg/l**	**TC MPN/100ml**
ARIMA model	(1,1,1)	(2,2,5)	(1,1,1)	(3,1,1)
Ljung-Box Q	13.235	6.865	13.210	13.012
P-value	0.655	0.810	0.657	0.526
RMSE	0.255	0.461	0.386	5016.875
MAPE	2.686	4.555	20.989	75.623
MAE	0.205	0.330	0.283	3343.543
BIC	-2.430	-0.719	-1.600	17.549

Table 6: Monthly forecasted water quality parameters values

Month	**pH**	**DO mg/l**	**BOD mg/l**	**TC MPN/ 100ml**
Jan-20	7.52	7.82	1.30	-1118.09
Feb-20	7.48	7.83	1.33	-217.28
Mar-20	7.47	8.28	1.33	-1173.69
Apr-20	7.45	8.02	1.33	105.12

May-20	7.44	8.44	1.32	-601.85
June-20	7.43	8.54	1.32	-690.84
July-20	7.42	8.48	1.31	-1817.69
Aug-20	7.41	9.01	1.31	-2111.70
Sep-20	7.4	8.81	1.30	-2584.16
Oct-20	7.39	9.18	1.30	-2556.98
Nov-20	7.38	9.42	1.29	-2867.78
Dec-20	7.37	9.32	1.29	-3142.25

Figure 1: Observed data and ARIMA (1, 1, 1) model prediction of pH

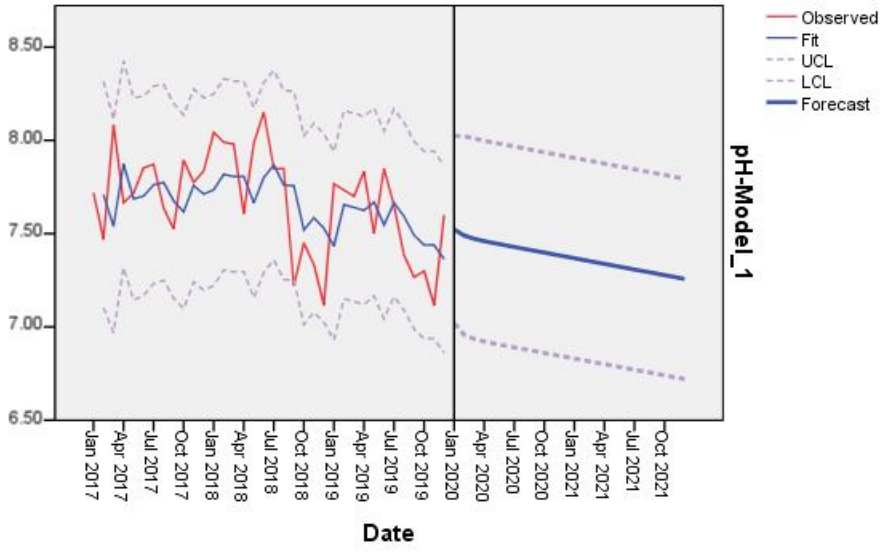

Figure 2: Observed data and ARIMA (2, 2, 5) model prediction of DO mg/l

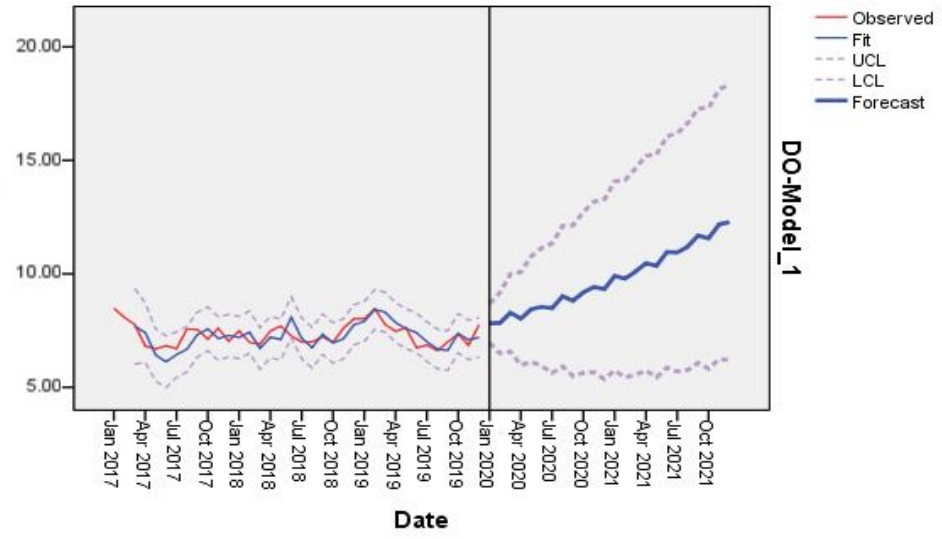

Figure 3: Observed data and ARIMA (1, 1, 1) model prediction of BOD mg/l

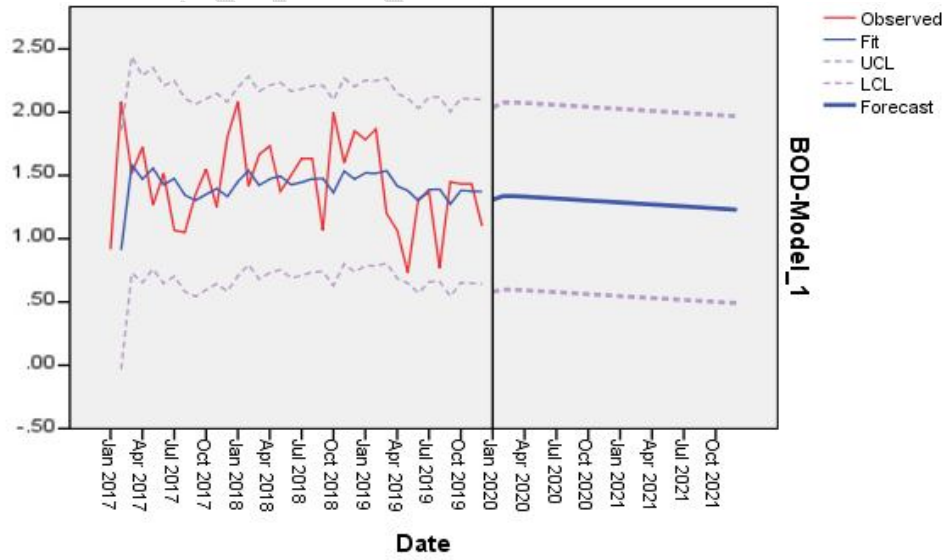

Figure 4: Observed data and ARIMA (3, 1, 1) model prediction of TC MPN/100ml