Prediction and WQI Assessment of Ganga River Sustainability by Box-Jenkins (ARIMA) model Approach

Jain Smita and Bundela Beena*
Department of Mathematics, JECRC University Jaipur, INDIA
*beena.bundela@jecrcu.edu.in; beenabundela@gmail.com

Abstract

Box-Jenkins (ARIMA) modelling approach for analysing the status of Water Quality of the Ganga River at Haridwar City of Uttarakhand, India was studied for understanding the Ganga River sustainability. The aim of the model was to predict future values of the series. Quality assessment of river water and prediction were done based on various water pollutants like pH, Dissolved oxygen (DO), Total Coliform, Chloride, Calcium, Magnesium, Hardness and Total Dissolved Solids (TDS).

Monthly data were collected for five years (2017-2021). Water Quality Index (WQI) and ARIMA model were reviewed. The WQI (Water Quality Index) of Ganga River was found more than 50 in almost all the months for five years (2017-2021). That shows that the river is polluted in most of the months. The Ljung-Box statistics gave nonsignificant p-value for 45 degrees of freedom and with 95% accuracy interval since the lag 48's p-value is higher than 0.05.

Keywords: Water quality index, Physiochemical parameters, Trend analysis, ARIMA model.

Introduction

For the human beings, river water plays very important role in the ecosystem, but due to human activities like population increase, industrial effluents, fertilizers used in the agriculture, the heavy metals are producing polluted water resources and depletion of aquatic biota[5]. Therefore, it is essential for the environment that the monitoring of water quality be investigated on a regular basis. It is problematic to understand the organic spectacles completely[1]. Many water impurities like metallic element, insecticides and persistent biological chemicals, not only disrupt the configuration and relative profusion of intestinal microbiota, but also control the metabolome, thereby disturbing the environment.

We studied the water quality in terms of physio-chemical water pollutants of the Ganga River at Haridwar City of Uttarakhand, India. Haridwar is a city situated in Uttarakhand, India. The Ganga River flows for 253 kilometres (157 miles) from its source entering the Indo-Gangetic plains of North India at Haridwar[6]. Water quality index (WQI) is the method to depict the overall water quality rank that is supportive for the assortment of suitable statistical methods to encounter the discussed issues. However, WQI calculates the composite impact of different parameters of Water and transfers water quality status to the human being and environmental decision makers[11].

When using quantitative techniques on chronological series, it is assumed that historical data can be used to predict future outcomes[4]. Conventional methods for analysing a chronological series include the moving average approach, exponential adjustment, linear and non-linear trend. In order to use these techniques, the series needs to be stationary, meaning that its mean and covariance should remain unchanged throughout time. We require comprehensive study of the interior and exterior environment of the river water. Thus, the internal behaviour of the self- factor approaches is the main subject of this study.

The ARIMA model, also known as the autoregressive moving average model, examines both the interference of random fluctuations and the water quality data in order to forecast the WQI. In this investigation into predicting short-term trends for the WQI, this model exhibits a high rate of accuracy. Environmental studies make considerable use of the ARIMA model because of a variety of factors including environment and man-made activity. We fit the ARIMA model to the WQI after it has been identified if the time series is stationary, seasonal, or stochastic[10]. The AR (Auto Regressive) and ARMA (Auto Regressive Moving Average) techniques are best suited for stationary series in this context since they yield more consistent predictions.

The term "ARIMA model" (p, d, q) refers to an ARIMA model[3].

where d is the number of differences, q is the number of moving averages and p is the number of terms in an autoregressive model.

Material and Methods

Sampling Area: This study included the water of river Ganga in India fomr Haridwar City of Uttarakhand, India as a sample site for the study.

Methodology: This study uses secondary data obtained from the Uttarakhand Government's Pollution Control Board's official website from January 2017 to December 2021. An effort was made to WQI (Water Quality Index) for the fifty-seven months. Within a year, seasonality has an impact on chronological series. Both linear stationary and linear consistent non-stationary processes can be characterised using quantitative ARIMA models. The three categories of models include: model of auto regression of order p (AR(p)),

https://doi.org/10.25303/293rjce020025

moving averages of order q (MA(q)) and auto regressive and moving average processes of order p and q [ARMA, (p q)].

To determine the conditional mean model for the underlying data, we use the sample autocorrelation functions and partial autocorrelation functions. For an autoregressive process, the sample autocorrelation functions decline regularly, but the sample partial autocorrelation functions cut off after a few lags. The sample autocorrelation functions for a moving average process, on the other hand, discontinue after a few lags, but the sample partial autocorrelation functions decrease steadily[2].

The order of homogeneity is the number of times the initial sequence must be differentiated in order to produce a stationary series[9]. Several water parameters such as pH, total coliform, DO, chloride, calcium, magnesium, total hardness and TDS were used for the study.

Statistical Analysis: To analysis the data, first WQI (Water Quality Index) was calculated. We selected ARIMA (1,1,0) model for predictions up to next 5 months for the WQI (Water Quality Index) for River Ganga by forecasting the values of WQI. The open sources Statistical Software Minitab and Excel were used for this study purpose.

Results and Discussion

Almost in all the months the river is highly polluted as the WQI is greater than 50 as in table 1.

Null hypothesis: The time series is non – stationary.

Alternative hypothesis: Time series is stationary.

We will now test the hypothesis by spreading the ADF (Augmented Dickey- Fuller) test to the time series assessing the appropriate difference of the data in dth order. We create a table (Xt = Xt – Xt-1) with the difference values between the current and the immediately prior one by differentiating in the first order (d=1).

The ADF test result showed p- value = 0.01 at 95% confidence Interval which is less than 0.05, so we reject the null hypothesis and we may conclude that the alternative hypothesis is true i.e. the series is stationary in its mean and variance. After testing the hypothesis, to get the appropriate values for p in the auto- regressive and q in the moving average for the model, we can use the ARIMA model.

To do that, we need to look at the partial and correlogram of the stationary (first order differenced) time series. Figure 1 displays the auto-correlation function (also known as the correlogram) overhead for lags 1 through 12 of the first order differentiating time series for the water quality index for the recorded fifty-seven months.

According to the Correlogram's interpretation, autocorrelations drop to zero after lag 2 and only marginally exceed significance limits at lag 1. Despite this, every coefficient between lag 5 and lag 12 fits inside the bounds [Figure 1]. Figure 2 displays the partial auto-correlation function, also known as the partial correlogram, for lags 1 through 12 of the differenced time series.

Additionally, the partial autocorrelation coefficient was found to surpass significant limits at lag 1 and to fall to zero after lag 1, according to the partial correlogram.

Table 1
Monthly Water Quality Index for January 2017 to December 2021

Month	WQI	Month	WQI	Month	WQI	Month	WQI	Month	WQI
17-Jan	50.3783	18-Jan	64.3043	19-Jan	66.0569	20-Jan	49.4953	21-Jan	66.9529
17-Feb	57.0197	18-Feb	**72.3171**	19-Feb	57.4732	20-Feb	52.4956	21-Feb	54.3959
17-Mar	59.0969	18-Mar	60.0212	19-Mar	55.6884	20-Mar	36.0103	21-Mar	**70.6621**
17-Apr	55.7401	18-Apr	55.6892	19-Apr	56.2561	20-Apr	54.5625	21-Apr	52.0912
17-May	43.749	18-May	59.4398	19-May	59.9229	20-May	67.3575	21-May	64.3933
17-Jun	49.751	18-Jun	71.1112	19-Jun	**77.2155**	20-Jun	63.3788	21-Jun	65.1371
17-Jul	55.693	18-Jul	55.6929	19-Jul	60.9841	20-Jul	60.3055	21-Jul	64.2183
17-Aug	**76.9087**	18-Aug	73.0841	19-Aug	52.3053	20-Aug	63.3306	21-Aug	62.4746
17-Sep	64.1214	18-Sep	**78.5139**	19-Sep	62.0704	20-Sep	65.4608	21-Oct	59.8673
17-Nov	60.2208	18-Oct	74.5468	19-Oct	64.8891	20-Oct	**70.1017**	21-Nov.	59.6882
17-Dec	52.8522	18-Nov	**80.5918**	19-Nov	66.3041	20-Dec	63.9468	21-Dec	54.4734
		18-Dec	68.7979	19-Dec	62.2832				

Table 2
Classification of Water Quality Index Ratings: Source[11].

WQI Ratings	0-25	26-50	51-75	76-100	>100
Cataloguing	Excellent	Slightly Polluted	Soberly Polluted	Polluted	Extremely Polluted (Unfitting)

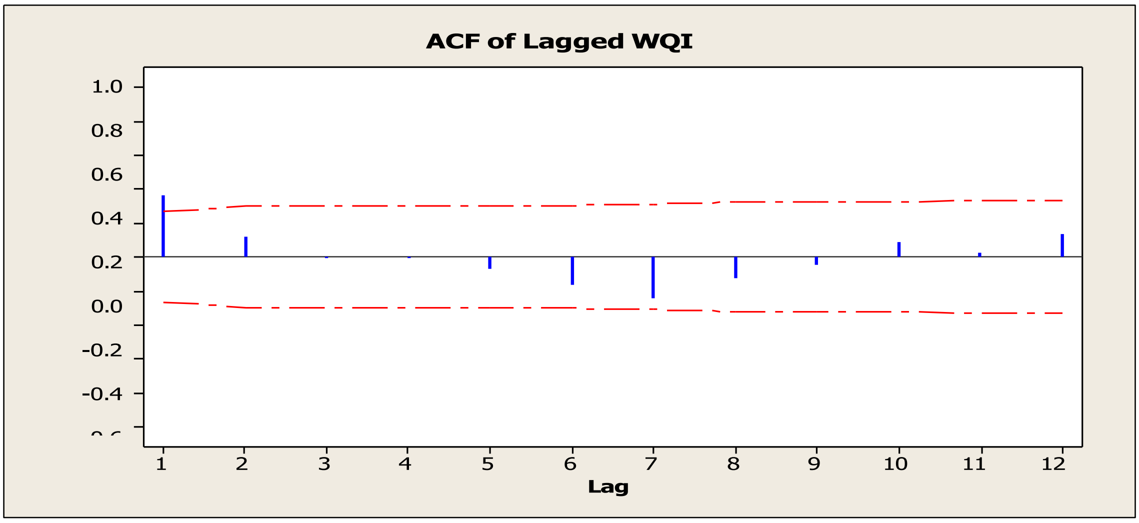

Figure 1: Correlogram (Auto correlation function) for the first order differenced time series

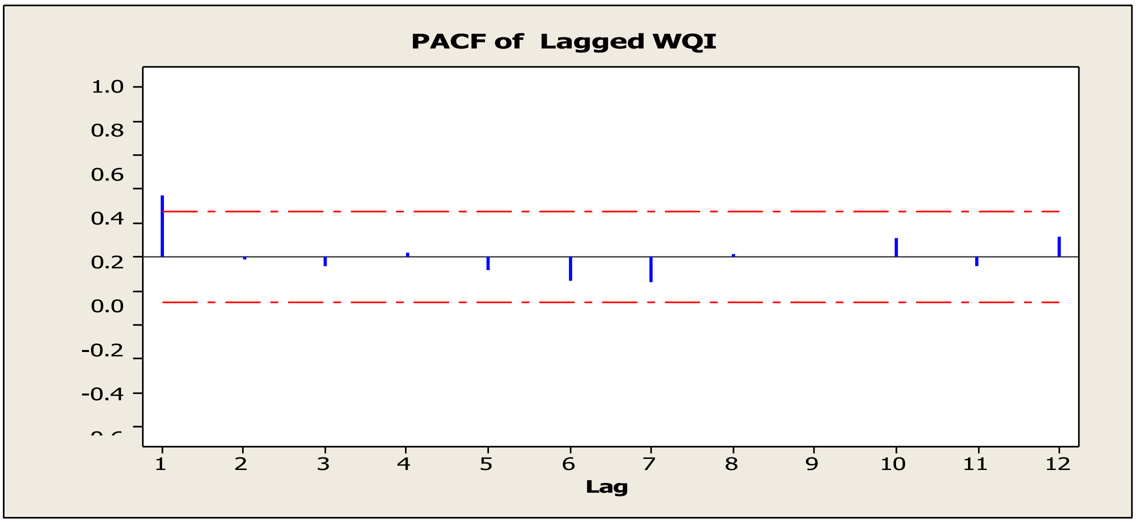

Figure 2: Partial correlogram (Partial auto-correlation function) for lags 1 to 12 of the differenced time series

From lag 2 to lag 12, other PACFs fall inside the significant bounds [Figure 2]. Since the correlogram [Figure 1] decrease to zero after lag 2 (omitting the outlier) and the partial correlogram [Figure 2] decreased to zero after lag 1 (omitting the outlier), we can define the following possible ARMA (auto regressive moving average) models for the first differenced time series data of WQI for River Ganga as ARIMA (1,1,0).

The Ljung-Box statistics gave nonsignificant p value for 45 degrees of freedom and at 5% level of significance as the p value is more than 0.05 for the lag 48 indicating that the residuals appeared to uncorrelated. Table 4 represents the prediction for the future values of the time series by chosen ARIMA (1,1,0) model for the next 5 months with 95% (low and high) prediction interval. In order to determine if the residuals with mean zero and constant variance are normally distributed or not, figure 4 displays the prediction errors of the ARIMA (1,1,0) model.

By carefully examining the numerous plots [Figures 3 and 4] of the standard residuals, we are able to conclude that standard errors are constantly distributed in the mean and variance of the fitted model. The plot displayed above appears to confirm the normalcy of errors. In order to investigate other correlations between successive prediction mistakes, now present the ACF Correlogram [Figure 5] and Partial Correlogram (PACF) [Figure 6] of the prediction errors.

We can conclude that there are no non-zero autocorrelations in the predicted residues (or standard errors) at lag 1 to 12 in the fitted ARIMA (1,1,0) model because none of the autocorrelation coefficients in the plot of the autocorrelation function above [Figure 5,6] penetrates the significant limits between lag 1 and 12.

In other words, all autocorrelation function values are fit within the significant limits. Our study reveals that in all the months of five years, the water quality index is greater than 50 and in some months, it is greater than 70 also. This is not a good indicator as the quality of water affects the human life in various aspects.

https://doi.org/10.25303/293rjce020025

The Autoregressive Integrated Moving Average (ARIMA) model for the WQI series and the Box-Jenkins approach are being used. All test parameters still had higher values. Considering the parameter values, the ARIMA (1, 1, 0) model fits the data the best. The ARIMA model has been chosen as the final model after comparison with other models.

Forecast values (calculated using the ARIMA model) for the following five months indicate a high level of WQI greater than 50 which indicates the same level of water quality.

Table 3
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	**12**	**24**	**36**	**48**
Chi-Square Statistic	18.5	34.7	55.3	58.8
Degrees of Freedom	9	21	33	45
P Value	0.030	0.031	0.009	0.082

Table 4
Forecasting by using selected Autoregressive Integrated Moving Average model

Period	**Forecast**	**Lower**	**Upper**
January 2022	52.6332	32.5968	72.6696
Feb.2022	60.4242	38.9326	81.9157
March.2022	61.9128	40.2107	83.6149
April.2022	59.262	37.5283	80.9956
May.2022	57.6521	33.411	81.8932

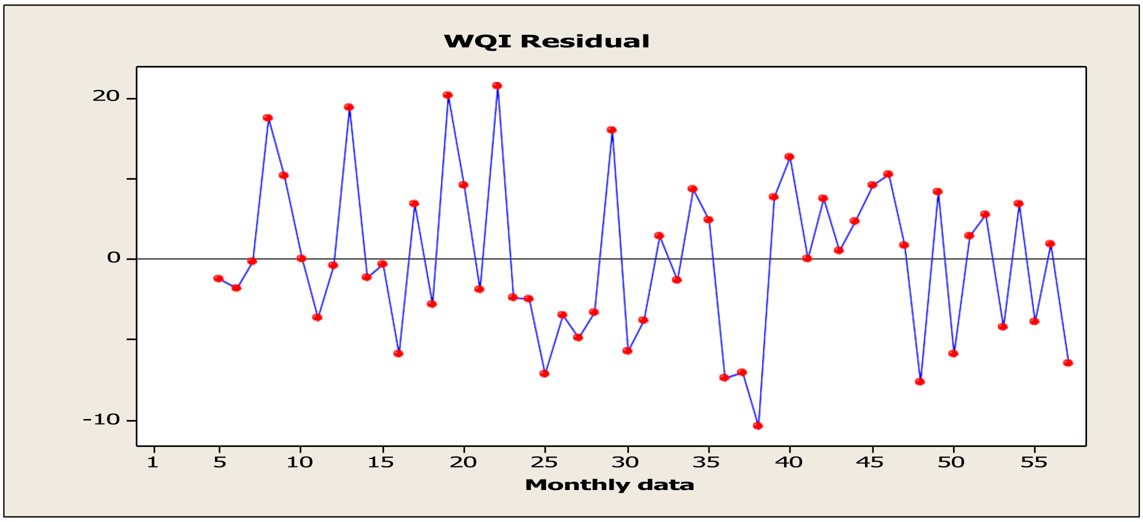

Figure 3: WQI Residual Plot

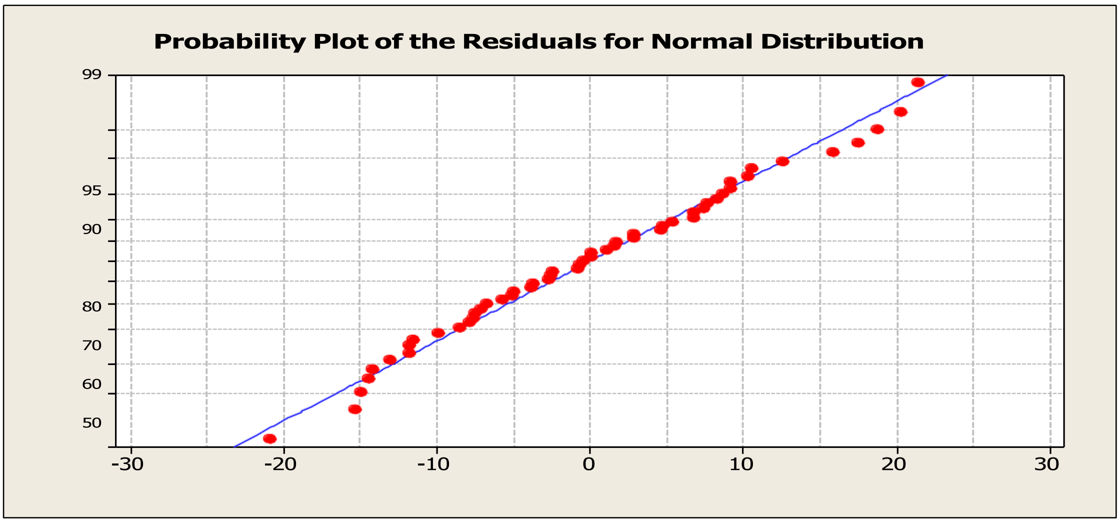

Figure 4: Normal Probability plot for Residuals

https://doi.org/10.25303/293rjce020025

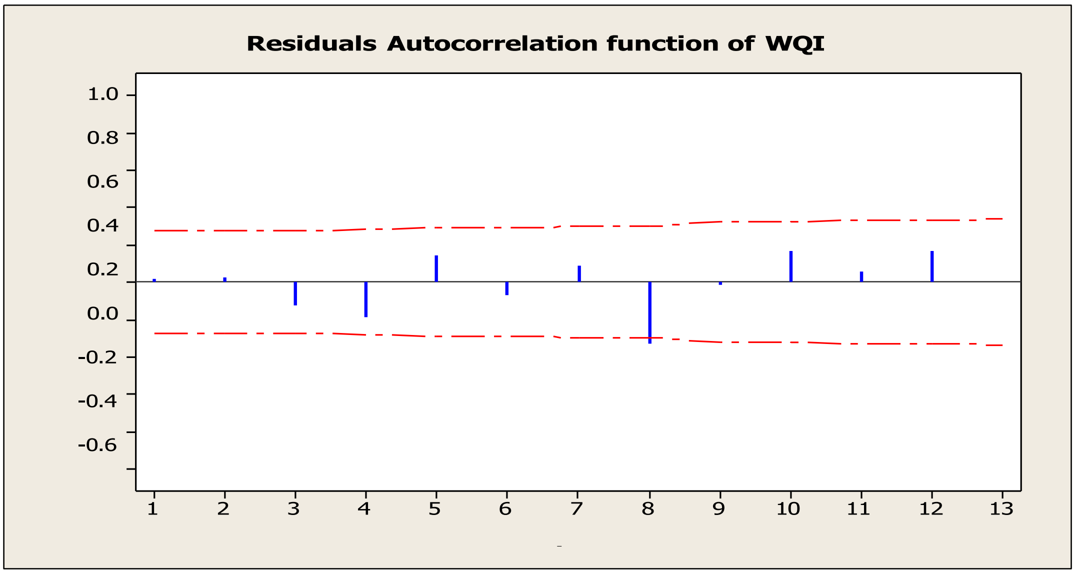

Figure 5: ACF of WQI for residuals

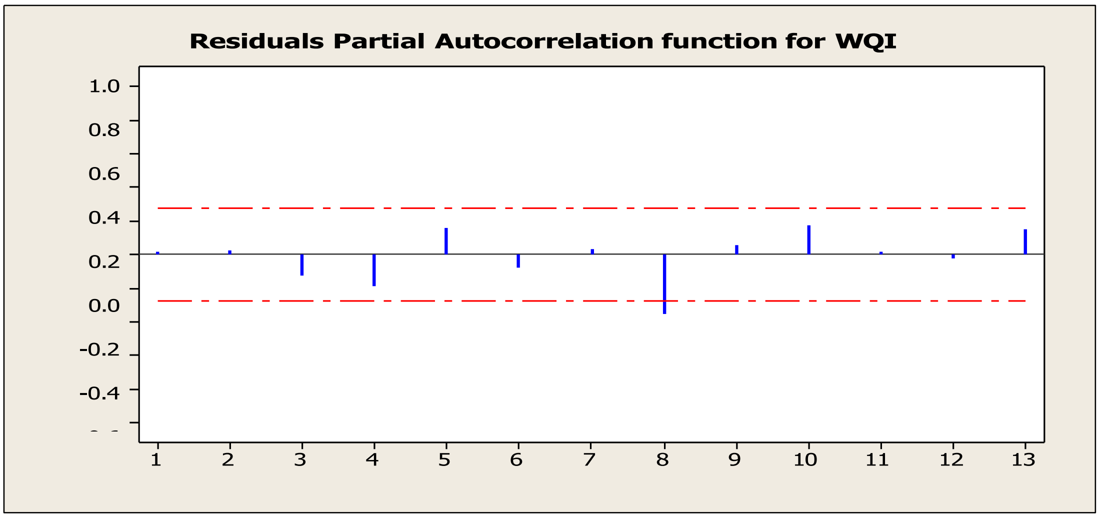

Figure 6: PACF of WQI for residuals

We recommend using this data-driven prediction and estimation strategy, which could be useful for Government and environmental protection agencies. We have applied the Modified Box-Pierce (Ljung-Box) Chi-Square statistic for the WQI residuals revealed that residuals are uncorrelated.

Conclusion

In this study first, the WQI (Water Quality Index) has been calculated for the different months for the year 2017 to 2021. In almost all the months the WQI is more than 50 that represents that the river is polluted in the complete five- year span. We after selected ARIMA (1,1,0) model for predictions up to next 5 months for the WQI (Water Quality Index) for River Ganga and by forecasting the values of WQI, we observe that the quality of the water remains same in all the leading five months. Estimates are created using ARIMA (1,1,0) and autocorrelations between the time series' consecutive values. The investigation also discovered that there was no correlation between the ARIMA time series and the subsequent residuals, or forecast mistakes in the fitted residuals, which seem to be normally distributed with mean zero and constant variance.

Hence, it has been proved that the selected Study ARIMA (1,1,0) provides an adequate predictive model for the water quality of River Ganga. The ARIMA (1,1,0) model predicted a high level of pollution in the river Ganga. So, the Government and non-government agencies should take this matter seriously and take the required decisions.

References

1. Coghlan Avril, A Little Book of R for Time Series, Readthedocs.org, Available from: http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html **(2010)**

2. Elsayir H., An Econometric Time Series GDP Model Analysis: Statistical Evidences and Investigations, *J. of Applied Mathematics and Physics*, **6**, 2635-2649 **(2018)**

3. Fattah J., Ezzine L., Aman Z., El Moussami H. and Lachhab A., Forecasting of demand using ARIMA model, *International J. of*

Engineering Business Management, https://doi.org/10.1177/1847979018808673, 10 **(2018)**

4. Gaither N. and Frazier G., Operations Management, South-Western, Ohio **(2001)**

5. Kumari Neeta and Pandey Soumya, Sustainability Assessment of Jumar River in Ranchi District of Jharkhand using River Sustainability Bayesian Network (RSBN) model Approach, Elsevier BV **(2022)**

6. Jain Smita and Thanvi Jyoti, 2018. Forecasting of Water Quality Parameters by Winter's Method with the help of Autoregressive Model, *Interdisciplinary J of Contemporary Research*, **5**, 9 **(2018)**

7. Morettin P.A. and Toloi C.M., Análise de Séries Temporais—2ª Edição Revista e Ampliada, 2nd edition, Editora Edgar Blucher, São Paulo **(2006)**

8. Pindyck R.S. and Rubinfeld D.L., Microeconomics, 7th edition, Prentice Hall, Upper Saddle River **(2008)**

9. Saha Priti and Paul Biswajit, Identification of potential strategic sites for city planning based on water quality through GIS-AHP integrated model, *Environ. Sci. Pollut. Res*, Doi: 10.1007/s11356-020-12292-9 **(2021)**

10. Tang X. and Deng G., Prediction of Civil Aviation Passenger Transportation Based on ARIMA Model, *Open J. of Statistics*, **6**, 824-834 **(2016)**

11. Tyagi Shweta et al, Water Quality Assessment in Terms of Water Quality Index, *American J. of Water Resources*, **1(3)**, 34-38 **(2013)**

12. Yujie Sha and Huan Wu, Effects of iodoacetic acid drinking water disinfection by product on the gut microbiota and its metabolism in rats, *J. Environ. Sci*., 91-104, https://doi.org/10.1016/j.jes.2022.02.048 **(2022)**

13. https://ueppcb.uk.gov.in/pages/display/96-water-quality-data.

(Received 11th August 2024, accepted 23rd September 2024)