International Journal of Education and Science Research Review
Volume-13, Issue-5 Sept-Oct-2026 E-ISSN 2348-6457
(Peer-Reviewed, Open Access, Fully Refereed International Journal) P-ISSN 2349-1817
www.ijesrr.in Impact Factor – 7.09 Email- editor@ijesrr.in

YAMUNA RIVER POLLUTION ANALYSIS

A Data-Driven Study Using Linear Regression

Subject: Environmental Data Analysis / Applied Statistics

Gaurika Bhanot, Economics Researcher and Consultant
Gaurang Bhanot, Class 12, Montfort Senior Secondary School
Amaira Madan, Class 10, Indus International School, Hyderabad

Data Sources: DPCC, CPCB, Delhi Jal Board, YAP Phase I–III Reports
Period Covered: 1993 – 2026

---

1. Abstract

The Yamuna River, one of India's most sacred and ecologically significant water bodies, has experienced severe and accelerating pollution over the past three decades, particularly along its 52-kilometre stretch passing through Delhi. Despite cumulative government expenditure exceeding ₹6,856 crore across three phases of the Yamuna Action Plan (YAP) between 1993 and 2025, water quality indicators have shown consistent deterioration.

This study employs linear regression analysis on real water quality data sourced from the Delhi Pollution Control Committee (DPCC), the Central Pollution Control Board (CPCB), and the Delhi Jal Board (DJB) to quantify pollution trends across time and space. The sequence is as follows:

(1) Year → Biochemical Oxygen Demand (BOD) (2)
Year → Dissolved Oxygen (DO)
(3) Distance from entry point → BOD

Results indicate statistically significant worsening trends ($R^2 = 0.878$, 0.725, and 0.674, respectively), with BOD rising from 14.2 mg/L in 1993 to 52 mg/L in 2025, far exceeding the safe limit of 3 mg/L. Faecal coliform levels peaked at 7.9 million MPN/100 mL in January 2025, representing 3,160 times the permissible standard. The study concludes that infrastructure investment alone is insufficient without addressing the daily sewage treatment shortfall of 641 million litres and the operational failures of 23 of 37 sewage treatment plants.

---

2. Introduction

The Yamuna River originates from the Yamunotri Glacier in Uttarakhand at an altitude of 6,387 metres and traverses approximately 1,376 kilometres before merging with the Ganga at Prayagraj. Of this length, a 52-kilometre stretch passes straight through the National Capital Territory (NCT) of Delhi; from Palla in the north to Asgarpur in the south. This section of the Yamuna, while accounts for less than 2% of the river's course, contributes an estimated 80% of its total pollution load. This makes the Delhi stretch of the Yamuna one of the most critically polluted urban river segments in Asia.

The river serves as a primary water source for Delhi, with the Delhi Jal Board drawing approximately 229 million gallons per day from it for treatment and distribution. Simultaneously, the river receives the discharge of 18 major drains carrying municipal sewage, industrial effluents, and solid waste from a metropolitan population exceeding 32 million.

The Najafgarh drain alone contributes approximately 65% of Delhi's total pollution load into the Yamuna, carrying an estimated 800 million gallons per day (MGD) of largely untreated sewage. The Government of India, in partnership with the Government of Japan through the Japan Bank for International Cooperation (JBIC) and later the Japan International Cooperation Agency (JICA), launched the Yamuna Action Plan in 1993. Three phases of this plan have been implemented over the last 30 years at a combined cost exceeding ₹6,856 crore.

Despite this significant investment, monitoring data of the Yamuna consistently shows worsening water quality parameters. This study uses a linear regression analysis to formally quantify these trends, evaluate the effectiveness of YAP interventions, and project future pollution levels under a business-as-usual scenario.

3. Objectives and Significance

The main objective of this study was -

- To analyse trends in Biochemical Oxygen Demand (BOD) and Dissolved Oxygen (DO) at the Nizamuddin Bridge monitoring station from 1993 to 2025
- To quantify the spatial progression of pollution across eight DPCC monitoring stations along the Delhi stretch
- To assess the effectiveness of YAP Phase I, II, and III interventions using data-driven evidence
- To project future BOD levels under the current trajectory and identify key infrastructure gaps

In the context of India's ongoing urban river restoration efforts and the 2025-2029 PM-led 30-point action plan for Yamuna rejuvenation, the following this study aimed to provide regression-based evidence of pollution trends. It also aimed to contribute to the discourse on whether financial investment in sewage infrastructure translates to measurable water quality improvement.

4. Literature Review

The Yamuna River sustains millions of people and forms a crucial part of the region's ecological fabric, necessitating a careful analysis of its pollution burden. A large body of work has examined the river's condition through the usual lenses of environmental science, policy, and engineering, drawing on indicators like Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), dissolved oxygen (DO), and microbial contamination. What emerges from this literature is a fair amount of heterogeneity, much of which can be attributed to differences in where and when the water samples were taken. Resolving these spatial and seasonal effects is necessary to evaluate the long-term trends or how conditions shift from year to year under different flow regimes.

The Central Pollution Control Board (CPCB) has been the backbone of this data landscape, publishing water-quality reports since the 1970s and building what amounts to a long historical baseline. Sharma and Kansal (2011) used this data to track the effects of YAP Phase I and II through BOD loads, finding that daily loads jumped from 130 tonnes in 1983 to 270 tonnes by 2008, despite the money poured into cleanup efforts under the 2 YAP phases. The present study pushes the analysis forward to 2025.

Another contribution to the literature was Yale's FORE program detecting dissolved oxygen at Nizamuddin bottoming out at 0 mg/L in 2009, a reading that essentially marked the river as biologically dead at that stretch. TERI estimates that the river needs roughly 3.46 billion litres per day of fresh flow just to support aquatic life.

A 2024 DPCC evaluation adds to this grim picture: of 37 operational sewage treatment plants in Delhi, 23 failed to meet discharge norms, and only 14 managed to hit the tightened BOD standard of under 10 mg/L. Meanwhile, JICA's mid-term review of YAP Phase III found that even with STPs removing 77-98% of BOD, the river's middle and lower reaches kept deteriorating.

More recently, researchers have started bringing computational tools into this space. Gupta et al. (2024), for example, applied time-series forecasting to pH, DO, BOD, and COD data spanning 2013-2021, showing that these methods can pick up on temporal patterns and project where water quality might be headed. That said, forecasting is fundamentally about predicting what comes next – it doesn't necessarily tell you, in plain terms, whether things have been getting better or worse over time, or how fast.

Hence, a simple linear regression, by contrast, offers something more direct: a clear read on the direction and pace of change, which makes it a genuinely useful tool for tracking long-term trends and putting a number on improvement or decline. Taken together, the literature makes a strong case for trend analysis that can feed into policy discourse. Yet for all the qualitative critique of the Yamuna Action Plan's shortcomings, there's surprisingly little quantitative regression work built on publicly available monitoring data. This study aims to fill that gap, using a transparent and reproducible approach - built in Python with scikit-learn - applied to DPCC-reported data.

5. Methodology

5.1. Data Sources

Data was compiled from the following official sources:

- Delhi Pollution Control Committee (DPCC): Monthly water quality reports (October 2024-February 2026) across 8 monitoring stations.
- Central Pollution Control Board (CPCB): Annual water quality assessments at Nizamuddin Bridge (1993-2025).
- Delhi Jal Board (DJB): Sewage generation and treatment capacity data (2000-2025).
- YAP Phase I, II, III project reports (JBIC/JICA): Phase timelines, budgets, and STP capacity data.
- National River Conservation Directorate (NRCD): Data on BOD load estimates and sewage treatment infrastructure.

5.2. Key Parameters

Three primary water quality indicators were selected for analysis:

Parameter	Unit	Safe Limit	Significance
Biochemical Oxygen Demand (BOD)	mg/L	≤ 3 mg/L	Measures organic pollution; higher = more sewage/waste
Dissolved Oxygen (DO)	mg/L	≥ 5 mg/L	Required for aquatic life survival
Faecal Coliform (FC)	MPN/100 mL	≤ 2,500	Direct indicator of sewage contamination

5.3. Datasets Constructed

Three datasets were constructed for analysis:

- Dataset 1 - Annual trend data: BOD and DO at Nizamuddin Bridge from 1993 to 2025 (33 data points), derived from CPCB annual reports and interpolated between verified anchor values where intermediate years were unavailable.
- Dataset 2 - Monthly DPCC data: Faecal coliform, BOD at Asgarpur, and DO at Palla for 17 months (October 2024-February 2026) sourced directly from DPCC monthly reports.
- Dataset 3 - Spatial station data: BOD and DO readings at 8 DPCC monitoring stations in January 2026, with station distance from Palla calculated using river course mapping.

5.4. Linear Regression Model

Linear regression is used to determine the relation between an independent variable (X) and a dependent variable (Y).

$$\mathbf{Y = \beta_0 + \beta_1 X + \varepsilon}$$

Where β_0 is the intercept, β_1 is the slope or the rate of change of Y with X and ε is the error in the model. Three different linear regression models were built. The dataset was divided into two parts, 80% of the data was used to train the models and 20% of the data was used to test the models. The performance of each model was checked using two parameters one is R Square and other is Root Mean Square Error (RMSE). R Square shows how well the model can explain the changes in the data. RMSE shows the difference between the values predicted by the model and the actual values. A lower RMSE means that the predicted values are closer to the actual values.

Models constructed:

- **Model 1:** At x axis Year (X) → BOD (y axis) at Nizamuddin. This model looks at how the BOD level at Nizamuddin has changed over the years.
- **Model 2:** At x axis Year (X) → DO at Nizamuddin (y axis). This model shows how the dissolved oxygen level at Nizamuddin has changed with time.
- **Model 3:** Distance from Palla in km (at X Axis) → BOD (at Y Axis) .This model looks at how the BOD level changes as the distance from Palla increases.

6. Results and Analysis

6.1 Exploratory Data Analysis

The annual data shows a sharp increase in BOD and a decrease in DO over the study period. At Nizamuddin Bridge, BOD increased from 14.2 mg/L in 1993 to 52 mg/L in 2022. It then decreased slightly to 41 mg/L in 2025, but this was still 13.7 times higher than the safe limit. DO decreased from 3.5 mg/L in 1993 to effectively zero in 2021-2022. It clearly indicates the deterioration of water quality.

The faecal coliform data from the DPCC monthly reports shows large changes across different months. The lowest value was recorded in October 2025 at 8,000 MPN/100 mL and in January 2025 recorded a very high value of 7,900,000 MPN/100 mL. This was 3,160 times the maximum permissible limit. And it could be linked to the lower river flow during the winter months, which causes pollutants to become more concentrated.

Variables Compared	Correlation (r)	Relationship	What it Indicates
Year vs. BOD	+0.96	Strong positive	BOD levels have generally increased as the years progressed.
Year vs. DO	−0.95	Strong negative	Dissolved oxygen levels have generally decreased over the same period.
BOD vs. DO	−0.92	Strong inverse	As BOD increases, dissolved oxygen levels tend to decrease.

6.2 Regression Model Results

Results of the three linear regression models are summarised below:

Model	Slope (β_1)	Intercept (β_0)	R^2 Score	RMSE
Model 1: Year → BOD	+0.851 mg/L/year	−1679.56	0.878	2.76 mg/L
Model 2: Year → DO	−0.091 mg/L/year	+184.14	0.725	0.32 mg/L
Model 3: Distance → BOD	+0.812 mg/L/km	+5.99	0.674	9.21 mg/L

Table 1: Linear regression model summary for Yamuna pollution indicators

6.3 Interpretation of Model 1 (Year → BOD)

Model 1 has a slope of +0.851 mg/L per year. This means that the BOD level at Nizamuddin increased by about 0.85 mg/L each year during the period studied. The R^2 value of 0.878 shows that the model explains about 87.8% of the change in BOD. This indicates a strong connection between the year and the BOD level in the dataset.

The residual results also give a reasonable picture of the model. The errors are roughly normally distributed, and their average is close to zero. This means that the main assumptions used for the regression are reasonably satisfied.

The slope remains positive during all three YAP phases, YAP-I, YAP-II, and YAP-III. BOD therefore continued to increase during each phase instead of showing a fall. The results suggest that none of the three phases was able to bring the overall pollution trend down. The regression line continues to move upward throughout the period covered by the study.

6.4 Interpretation of Model 2 (Year → DO)

Model 2 puts the slope at −0.091 mg/L per year, which works out to a drop of roughly 2.9 mg/L in dissolved oxygen across the 32-year span from 1993 to 2025. What's striking is that DO was already sitting at just 3.5 mg/L back in 1993, below the 5 mg/L minimum needed to sustain aquatic life. The R^2 of 0.725 points to a fairly strong relationship between year and DO, though it's a little weaker when compared with BOD. This is likely because DO isn't just driven by long-term trends; it also fluctuates with the seasons, typically rising during the monsoon and dipping in winter.

6.5 Interpretation of Model 3 (Distance → BOD)

Model 3 looks at how pollution changes along the Yamuna with reference to DPCC station data . Here, the slope of +0.812 mg/L per km indicates BOD rising by about 8.1 mg/L for every 10 kilometres travelled from Palla. The R Square value of 0.674 suggests a moderate relationship between distance and BOD, not a perfectly linear one or dependent. That's probably because pollution doesn't build up gradually and evenly along the entire stretch; instead, it spikes wherever major drains dump waste into the river at specific points. Palla, at the 0 km mark, recorded just 2.5 mg/L of BOD and turned out to be the only station within safe limits. By the time the river reaches Asgarpur at 52 km, BOD had shot up to 52 mg/L, a roughly 20-fold increase over that stretch alone.

6.6 Future Projections

Applying Model 1 to project future BOD at Nizamuddin under the current trend:

Year	Predicted BOD (mg/L)	Times Above Safe Limit
2027	45.0	15 x
2029	46.7	16 x
2031	48.5	16 x

Table 2: BOD projections at Nizamuddin Bridge under current trend (Model 1)

These projections assume that there will be no major changes in sewage treatment infrastructure or enforcement. The PM's 30-point action plan for 2025–2029 could change this trend if it is implemented effectively. However, the experience of YAP Phases I–III suggests that weak implementation has repeatedly reduced the impact of these initiatives and limited the expected improvement in the Yamuna's water quality.

---

7. Discussion

7.1 Why Has the Yamuna Action Plan Failed?

The regression models help us understand how the different YAP phases affected the level of pollution in the Yamuna River. One major observation is that the BOD level increased across all three YAP phases. From 20.87 mg/L during YAP-I (1993-2002), it went up to 27.08 mg/L in YAP-II (2003-2012) and reached 37.91 mg/L in YAP-III (2013-2025). This suggests that YAP has not been able to control the growing pollution in the Yamuna effectively.

The primary structural reason behind this failure is the sewage treatment gap:

- Delhi generates 3,596 MLD of sewage daily (2025)
- Installed STP capacity is 3,474 MLD (nearly sufficient on paper)
- Actual sewage treated is only 2,955 MLD
- The gap of 641 MLD enters the Yamuna untreated every day, equivalent to approximately 256 Olympic-sized swimming pools of raw sewage

The capacity of treatment plants on paper does not always show how well they are working. The DPCC's 2024 audit found that 23 of 37 STPs failed to meet discharge standards, while only 14 met the stricter revised BOD standard of less than 10 mg/Having treatment plants does not automatically mean that the sewage is being treated properly. The plants also need to work properly and treat the sewage as they are supposed to. If these plants are not working as required, their actual contribution to pollution control remains limited. In other words, there is a clear gap between the treatment capacity that is officially reported and the amount of sewage that is treated effectively.

These projections assume that there will be no major changes in sewage treatment infrastructure or enforcement. The PM's 30-point action plan for 2025-2029 could change this trend if it is implemented effectively. However, the experience of YAP Phases I-III shows that gaps in implementation have repeatedly limited the expected improvements in the Yamuna's water quality.

7.2 Limitations of the Study

- BOD and DO data for the whole 1993-2020 stretch just wasn't available. For the missing years values based on CPCB's trend descriptions are taken. Some of the individual yearly numbers probably don't line up exactly with what the official records would show.
- There's also the matter of how the data was split for training and testing. A standard random split was used, which works fine as a first pass, but isn't really the right approach if the model is meant for actual forecasting. Splitting by time instead, training on everything before 2015 and testing on what came after would give a far more honest read on how well the model generalizes, and would avoid the risk of the model "peeking" at future patterns during training.
- Linear regression assumes things move in a fairly straight line, but BOD levels dip and spike with the seasons, and monsoons. A polynomial regression fit would probably track those ups and downs more faithfully than a straight line ever could.
- Lastly, this analysis only looks at BOD and DO in isolation. It doesn't factor in things like population growth, how much rainfall varies year to year, or shifts in industrial discharge, all of which almost certainly play a role. Bringing those variables in would be a natural next step for anyone building on this.

7.3 Implications for Policy

Based on the results of this model, a few steps can be taken to improve the condition of the Yamuna.

- There should be more monitoring points along the Yamuna. Increasing the number from 8 to at least 18 would make it easier to see where the pollution is coming from, especially from the major drains.
- Sewage treatment plants should be checked regularly: The treated water should also be tested to see if it is safe and meets the required standards. The results should be published every month so that any problems can be noticed and dealt with on time.
- Industrial waste should be treated properly: The Najafgarh drain contributes about 65% of the pollution entering the Yamuna in Delhi. Industrial wastewater needs special attention. It should be properly treated before it reaches the river or gets mixed with the city's sewage.
- Untreated sewage needs to be dealt with first: Around 641 MLD of wastewater is still not treated every day. Reducing this should be one of the main goals of the 2025-2029 action plan. Building more STPs alone will not solve the problem. The existing STPs also need to be used properly and kept working.
- During winter, the Yamuna has quite less flow in some areas. Because of this, the pollution becomes more concentrated, and the condition of the river gets worse. Making sure that enough water continues to flow through the river can help reduce the concentration of pollutants and improve the river's condition.

8. Conclusion

This study used linear regression to study the changes in Yamuna river water quality over three decades. The results show a clear pattern that has also been reported by environmental studies: the condition of the river has been getting worse, and the actions taken so far have not been able to improve the situation.

The results from the three models show different changes in the Yamuna's water quality-

Model 1 (Year → BOD):
Gave an R^2 of 0.878 and a slope of +0.851 mg/L per year. This shows that BOD has increased strongly over time, with the trend in pollution explaining 87.8% of the changes in BOD. The increase has continued during all three YAP phases.

Model 2 (Year → DO):
Showed that DO has decreased by 0.091 mg/L per year. Because of this, dissolved oxygen levels have reached close to zero in the most polluted parts of the river.

Model 3 (Distance → BOD):
Showed that BOD increases by about 0.81 mg/L for every kilometre downstream through Delhi.

The findings show that increasing investment and infrastructure have not been enough to solve the Yamuna's pollution problem. There is still a major treatment gap, with 641 MLD of untreated sewage entering the river every day and 23 of 37 STPs operating below standards. If this continues, simply building more infrastructure is unlikely to change the pollution trend shown by the regression results. Better operation and regular monitoring of STPs are also needed.

The current trend suggests that BOD could reach 45-48 mg/L at Nizamuddin Bridge by 2027-2031. This shows that stronger action is needed to improve the river's condition. To change this situation, future action plans will need to focus not only on building new infrastructure. Treatment performance should be regularly measured, standards should be properly enforced, and the results should be publicly available.

References

1. Delhi Pollution Control Committee (DPCC). Monthly Water Quality Reports for Yamuna River, October 2024 - February 2026. Government of NCT of Delhi.

2. Central Pollution Control Board (CPCB). Water Quality of Rivers in India - Annual Reports. Ministry of Environment, Forest and Climate Change, Government of India.

3. Sharma, D. & Kansal, A. (2011). The Status and Effects of the Yamuna Action Plan. Yale FORE Journal of Environmental Studies.

4. National River Conservation Directorate (NRCD). Yamuna Action Plan - Phase I, II, III Project Reports. Ministry of Environment and Forests, Government of India.

5. Japan International Cooperation Agency (JICA). YAP Phase III - Impact on Water Quality of River Yamuna. JICA India Office, 2022.

6. Delhi Jal Board (DJB). Sewage Treatment Plant Performance Data 2025. Government of NCT of Delhi.

7. Energy and Resources Institute (TERI). Environmental Flow Requirements for the Yamuna River. TERI Press, New Delhi.

8. Ministry of Jal Shakti, Government of India. Action Plan for Yamuna River Rejuvenation 20252029. Press Release, May 2025.

9. Countercurrents.org. Cleaning the Yamuna River to Get Clean Water to Delhi. July 16, 2025.

10. Millennium Post. Data Shows High Faecal Coliform Levels in Yamuna. March 10, 2026.

11. Sridhar, D. (2026). Rivers under stress: A comprehensive review on pollutant sources, human and ecological impacts, analytical, statistical, and geospatial methods and restoration strategies, for evaluating river water quality in India. Environmental Geochemistry and Health, 48(4), 160. https://doi.org/10.1007/s10653-026-03053-7

12. Mandal, P., Upadhyay, R., & Hasan, A. (2010). Seasonal and spatial variation of Yamuna River water quality in Delhi, India. Environmental Monitoring and Assessment, 170, 661–670. https://doi.org/10.1007/s10661-009-1265-2

13. Gupta, N., Yadav, S., & Chaudhary, N. (2024). Time Series Analysis and Forecasting of Water Quality Parameters along Yamuna River in Delhi. Procedia Computer Science, 235, 3191–3206. DOI: 10.1016/j.procs.2024.04.302.