Master Thesis

River pollution & Development: The case of the Ganga

Arjun Arora

Matriculation Nr. 5244256
Master Data Science

Institut für Mathematische Stochastik

First examiner: Prof. Dr. Nicole Mücke
Second examiner: Prof. Dr. Jens-Peter Kreiß
Supervisor: Hanh My Le, M.A

March 31, 2024

Statement of Originality

This thesis has been performed independently with the support of my supervisor/s. To the best of the author's knowledge, this thesis contains no material previously published or written by another person except where due reference is made in the text.

Braunschweig, March 31, 2024

Abstract

Rising surface water pollution presents significant health and economic risks. This study focuses on the Ganga River in India, which is not only the nation's largest river but also among the most polluted globally. Millions rely on it for water, agriculture, and tourism, and its sacred status in Hinduism makes its pollution a critical issue. Despite numerous efforts to mitigate pollution, effective results have been limited. We study the long-term health implications of rising levels of pollutants like Biochemical Oxygen Demand (BOD), Nitrate, and a measure for non-drinkability of water by attributing disease-related deaths to these variables. We find that for each of these measures, there is a positive correlation with the number of deaths, notably, BOD showed a significant correlation with mortality after a 2-year lag (p-value < 0.1). We also study the long-term socio-economic implications of the same pollutants and non-drinkability status in districts along the path of Ganga and its tributaries correlation with nightlight intensity. Nightlight is a well-used proxy for economic activity in a region, highlighting regions of higher versus lower economic activity. Our results show a largely negative correlation between nightlight per capita and pollution measures. For some pollutants, effects on nightlight intensity become apparent after up to 3 years, which shows delayed effects of pollution. This study not only confirms the detrimental health effects of water pollution but also sheds light on the complex socio-economic impacts in developing nations like India, underscoring the urgent need for comprehensive environmental and public health strategies.

Contents

1 Introduction **1**
- 1.1 Research Question . . . 2
- 1.2 Structure . . . 2
- 1.3 Hypothesis . . . 4
- 1.4 Methodology . . . 6
- 1.5 Data . . . 7
- 1.6 Results . . . 8
- 1.7 Related Work . . . 10

2 Context **13**
- 2.1 Impact on World economy . . . 13
- 2.2 Impact on Indian economy . . . 14
- 2.3 River pollution - Ganga . . . 15
  - 2.3.1 The Ganga River's Ecosystem and Uses . . . 19
  - 2.3.2 Pollution Challenges . . . 20
  - 2.3.3 Impact on Society and the Environment . . . 24
  - 2.3.4 Efforts for Restoration and Improvement . . . 25
- 2.4 Nightlight as proxy for Income Inequality . . . 26
- 2.5 Economic concepts . . . 26
  - 2.5.1 Panel Data Analysis: An Overview . . . 26
  - 2.5.2 Interpreting PanelOLS Regression Results . . . 32

3 Data **35**
- 3.1 GADM . . . 35
  - 3.1.1 Data preparation and Exploratory data analysis of GADM . . . 37
- 3.2 River pollution data, CPCB . . . 38
  - 3.2.1 Exploratory data analysis of pollution data . . . 43
- 3.3 Economic activity data - VIIRS Nightlight . . . 47
  - 3.3.1 Data preparation . . . 48
  - 3.3.2 Exploratory data analysis . . . 51
- 3.4 Health and Family Welfare . . . 53

4 Empirical Framework **57**
- 4.1 River pollution's effects on Health . . . 57
  - 4.1.1 Regression Formula and Variables . . . 58
  - 4.1.2 Fixed effects model with log-log transformations . . . 61

4.2 River pollution's effects on Economic Activity through Nightlight Intensity 62
4.2.1 Regression Formula and Variables 62
4.2.2 Fixed effects model with log-log transformations 64

5 Results and Discussions 67
5.1 Regression results River pollution vs Deaths due to diseases 68
5.1.1 Biological Oxygen Demand and Mortality Rates 68
5.1.2 Nitrate Levels and Mortality Rates 70
5.1.3 Not-drinkable and Mortality Rates 71
5.2 Regression Results: River Pollution vs. Nightlight Per Capita 73
5.2.1 BOD and Nightlight per Capita 75
5.2.2 Nitrate and Nightlight per Capita 76
5.2.3 Not-drinkable and Nightlight per Capita 79

6 Conclusion 83

7 Appendix 93
7.1 Framework of Subset Regressions 93
7.1.1 Districts with More Industries than Average 93
7.1.2 Districts with More than Mean Population 95
7.1.3 Upstream vs. Downstream States (Uttarakhand vs. West Bengal) 97
7.1.4 Districts with More Literacy Rate vs Less Literacy Rate 98
7.2 Common Limitations and Considerations: 99
7.3 Technical Considerations 99

List of Figures

2.1 Ganga Basin 15
2.2 Sangam Devrayag 17
2.3 Industrial water consumption 18

3.1 India GADM map 36
3.2 Not-drinkable status 42
3.3 Total deaths per state 44
3.4 Station codes India 46
3.5 Nightlight per capita districts 49
3.6 Comparative analysis of pollution and nightlight trends. 52
3.7 Total deaths per state 55
3.8 Average deaths per year 56

List of Tables

2.1 Drainage area by states 16
2.2 Urban population growth from 2001-2011 25

3.1 Criteria for Classification of Surface Water Quality 41
3.2 Summary of Key Metrics from the Dataset 48

5.1 Impact of BOD on Death Rates 68
5.2 Impact of Nitrate on Death Rates 70
5.3 Impact of Not-drinkable water on Death Rates 72
5.4 BOD impact on Log Nightlight per Capita 74
5.5 Nitrate impact on Log Nightlight per Capita 77
5.6 Not-drinkable Impact on Log Nightlight per Capita 79

7.1 Regression Results for Districts with More Than Mean Population 94
7.2 Regression Results for Districts with More Than Mean Population 96

Abbreviations and Acronyms

CPCB Central Pollution Control Board

NWMP National Water Monitoring Programme

NMCG National Mission for Clean Ganga

CWC Central Water Commission

GADM Global Administrative Areas

SHRUG Socioeconomic High-resolution Rural-Urban Geographic Platform for India

VIIRS Visible Infrared Imaging Radiometer Suite

EOG Earth Observation Group

BOD Biological Oxygen Demand

DO Dissolved Oxygen

OLS Ordinary Least Squares

1 Introduction

Water, the most vital resource for all forms of life, plays a pivotal role in the socio-economic fabric of societies across the globe. In developing countries, where economic activities are closely intertwined with natural resources, the quality of water directly influences public health, agricultural productivity, and industrial efficiency. However, rapid industrialization, coupled with inadequate environmental regulations, has led to escalating levels of water pollution, posing severe risks to human health, biodiversity, and economic stability.

The significance of studying water pollution transcends environmental concerns, encompassing critical health and economic dimensions, particularly in developing nations. Waterborne diseases, resulting from contaminated water sources, remain one of the leading causes of mortality and morbidity in these regions. The health impacts of polluted water—ranging from acute gastrointestinal infections to long-term chronic conditions—place a substantial burden on public health systems[1], impede human capital development, and exacerbate poverty cycles.

Beyond the health implications, water pollution undermines economic activities by degrading natural ecosystems that support fisheries, agriculture, and tourism. The contamination of freshwater resources can lead to increased costs for industry and agriculture, which rely heavily on clean water for production processes. Furthermore, the economic repercussions of water pollution extend to diminished property values, reduced recreational and aesthetic benefits, and increased expenditure on water treatment and public health interventions.

Given the intertwined nature of water quality, health outcomes, and economic prosperity, this thesis aims to unravel the intricate relationships between water pollution and its health and socio-economic impacts, with a particular focus on the Ganga or Ganges river in India. By employing innovative methodologies, such as the use of nightlight intensity as a proxy for economic activity, this study endeavors to shed light on the often overlooked economic dimensions of environmental degradation. The research conducted herein is poised to contribute to the burgeoning discourse on sustainable development, offering insights that could inform policy-making, enhance environmental stewardship, and foster economic resilience in the face of escalating water pollution challenges.

As the world's developing nations strive to balance economic growth with caring for the environment, understanding water pollution's broad impacts is key. This work moves us closer to grasping the full effects of water pollution on health and economies, aiming to guide better water management and sustainable development.

1.1 Research Question

This study aims to unravel the complex relationship between the fluctuating levels of river pollution in specific stretches of the Ganga and their consequential impacts on the surrounding region's health and economic activities. Recognizing the limitations in direct economic data, this research adopts an innovative approach by using nightlight intensity as a proxy for economic activity.

This methodology is based on the hypothesis that fluctuations in river pollution levels—whether increases or decreases—exert a tangible impact on the economic vitality of adjacent areas, as evidenced by changes in nightlight intensity. Captured through satellite imagery, nightlight intensity serves as a novel, indirect measure of economic activity, offering valuable insights into the economic health and development of the region.

The core objective is to discern whether there is a significant correlation or, more ambitiously, a causal relationship between the variations in river pollution and the economic prosperity of the regions along the Ganga. This investigation is crucial, as it addresses a gap in existing research by quantitatively linking environmental degradation with socio-economic outcomes. The findings of this study are expected to provide a deeper understanding of the socio-economic implications of river pollution, which is vital for informed policy-making and sustainable environmental management in the context of the Ganga river in India.

The effects can be categorized into two levels. At the first level, the immediate health impacts of river pollution are both well-documented and evident. Rising pollution levels lead to physiological suffering among individuals reliant on the river for their essential needs. Elevated toxins and chemicals in the water escalate the risk of severe illnesses, including cholera, diarrheal diseases, hindered child growth, and pneumonia. These direct health consequences are crucial to investigate alongside the socio-economic impacts since the latter are often a direct extension or, at the very least, significantly influenced by these health issues. The ripple effects of deteriorating health include reduced workforce participation, soaring medical expenses, compromised educational opportunities for children, and a diminished pursuit of economic improvement. This study delves into the nuanced second-level impacts in depth, yet it also addresses the primary health effects to quantify the contribution of escalating river pollution to disease prevalence and mortality along the Ganga.

1.2 Structure

This thesis is organized into several chapters, each dedicated to a different aspect of the investigation into the effects of water pollution on health and economic activities, with a particular focus on the Ganga river. The structure is designed to guide the reader through the comprehensive study from theoretical foundations to empirical findings and conclusions. Here is an overview of what each chapter entails:

Chapter 1: Introduction This chapter sets the stage for the thesis, highlighting the study objective, the two hypothesis and an overview of the data and methodology used. Furthermore, this chapter presents a preliminary summary of the key findings, alongside a comprehensive review of existing literature relevant to the domain of water pollution and its effects on health and economic activities.

Chapter 2: Context and Preliminaries This chapter delves into the theoretical underpinnings and background necessary for understanding the study. It covers a detailed exposition of the Ganga river, including the types of pollution it suffers from and the specific chemicals and toxins commonly found in its waters. Furthermore, it introduces the concept of using nightlight as a proxy for economic activity, explaining the rationale behind this approach. Additionally, it outlines the economic theories and concepts that are pivotal for comprehending the potential socio-economic impacts of river pollution. This chapter aims to equip the reader with the necessary context and preliminary knowledge to grasp the complexities of the research subject.

Chapter 3: Data Description In this chapter, the thesis outlines the datasets employed throughout the study. It describes the river pollution data, mortality data related to waterborne diseases for examining health outcomes, and nightlight data utilized to assess economic effects. This section details the origins of each dataset, the manipulations and cleaning processes undertaken to prepare the data for analysis, and the rationale behind the selection of these specific datasets. The objective is to provide transparency regarding the data sources and to elucidate the steps taken to ensure the reliability and validity of the analysis.

Chapter 4: Empirical Framework Chapter 4 presents the empirical framework that underlies the thesis. It details the models and regression analyses conducted to explore the relationships between pollution levels and nightlight intensity, among various pollution measures, and between pollution and health outcomes. This chapter is the crux of the thesis, where the theoretical models are applied to the data to extract meaningful insights and patterns. The methodologies, assumptions, and statistical techniques employed in the regression analyses are thoroughly explained to underscore the robustness of the study's empirical approach.

Chapter 5: Discussion of Results This chapter discusses the results obtained from the regression analyses. It interprets the findings in light of the study's hypotheses and objectives, assessing the implications of the results for understanding the socio-economic impacts of river pollution. Additionally, it contextualizes the findings within the broader body of literature, comparing and contrasting the results with previous studies in the field. The discussion aims to shed light on the significance of the research outcomes, offering insights into the relationship between environmental degradation and socio-economic development.

Chapter 6: Conclusion The concluding chapter synthesizes the key findings of the thesis, reflecting on the study's contributions to the field of environmental economics and public health. It revisits the research questions and hypotheses, summarizing how the study addressed these inquiries and what it revealed about the socio-economic effects of river pollution. The chapter also outlines potential directions for future research and offers recommendations for policymakers, emphasizing the importance of integrated approaches to water management and economic planning.

Appendix The appendix includes supplementary materials, detailed tables, and additional analyses that support the research findings but are too voluminous to include in the main body of the thesis.

References A comprehensive list of all the sources cited throughout the thesis is included at the end, following academic referencing standards.

This structured approach ensures a logical flow of information, enabling readers to understand the complex dynamics between environmental degradation and socio-economic outcomes in developing countries.

1.3 Hypothesis

Hypothesis This study posits two primary hypotheses related to the effects of river pollution along the Ganga on both health and economic activities in adjacent regions.

Firstly, it suggests that an increase in pollution levels will negatively impact economic activities, discernible through changes in nightlight intensity, a proxy for economic activity. Nightlight intensity, as captured by satellite imagery, is believed to reflect the aggregate economic activity in an area, with brighter areas indicating more robust economic activities.

Secondly, the study hypothesizes that rising levels of river pollution are associated with deteriorating health outcomes among the human population residing along the riverbanks. Specifically, as river pollution increases, health outcomes, as measured by disease incidence and mortality rates, are expected to increase.

The hypothesis posited in this study is that the levels of river pollution in certain areas along the Ganga have a measurable impact on the economic activities in adjacent regions. Specifically, it suggests that an increase in pollution levels will negatively affect these activities, while a decrease in pollution levels will have a positive impact. This relationship is hypothesized to be observable through changes in nightlight intensity—a proxy for economic activity. Nightlight intensity, as captured by satellite imagery, reflects the aggregate economic activity in an area, with brighter areas typically indicating more robust economic activities.

Assumptions:

- Correlation Between Nightlight Intensity and Economic Activity: The hypothesis assumes that nightlight intensity reliably correlates with economic activity levels,

where increased brightness equates to enhanced economic engagement and vice versa.

- Impact of River Pollution on Economic Activity: It presupposes that river pollution has a tangible, measurable impact on economic activities, affecting health, productivity, and thereby, the economic output observable through nightlight intensity.
- Sensitivity of Nightlight to Economic Fluctuations: The assumption that changes in economic activity due to varying pollution levels are sufficiently pronounced to be detected through changes in nightlight intensity.
- Direct Health Impact of River Pollution: Increased pollution levels lead to higher rates of diseases and mortality among communities dependent on the river for their daily needs.

Reasoning behind nightlight as a proxy for economic activity: The justification for adopting nightlight as a proxy for economic activity stems from empirical findings indicating a moderate linear correlation between night light intensity and socio-economic indices like poverty rate and income inequality. Such correlation suggests that nightlight features could be instrumental in predictive models for poverty indices, encapsulating diverse socio-economic indicators.

The promising outcomes of employing nightlight data at even a provincial level underscore its potential as a proxy for both poverty and income inequality, underpinning the hypothesis that changes in river pollution levels—reflected through nightlight intensity—can serve as an indicator of economic health and activity adjustments in regions adjacent to the Ganga.

Alternate hypothesis Contrary to the commonly observed negative impact of river pollution on economic growth, this study proposes an alternate hypothesis: regions experiencing higher levels of river pollution may concurrently exhibit signs of increased economic activity. This paradoxical situation is identifiable through enhanced nightlight intensity, serving as an unconventional yet telling indicator of economic vibrancy. Two primary assumptions underlie this hypothesis:

- **Industrial Agglomeration:** The areas suffering from significant river pollution may simultaneously function as industrial agglomerations. These zones, characterized by dense concentrations of manufacturing and processing facilities, could emit more light during the night due to extended operational hours and the presence of large industrial complexes. Despite the environmental degradation, these areas might report economic growth as industries capitalize on lower land prices and more lenient regulatory environments.
- **Urban Economic Centers:** Rivers, especially those that traverse through or near urban centres, can be subject to pollution due to the high density of economic

activities. The economic pulse of urban areas, regardless of the environmental conditions, tends to be stronger. Nightlight data might therefore reveal increased brightness correlating with urban economic hubs, despite the presence of river pollution. This brightness is a reflection of the urban sprawl and the economic activities that thrive within city bounds, including commerce, services, and even informal economies that burgeon in less regulated environments.

This alternate hypothesis challenges traditional perceptions by suggesting that environmental degradation, specifically through river pollution, does not uniformly lead to economic downturns but may be associated with an increase in observable economic activity. The thesis will critically examine this phenomenon, exploring the complex inter-dependencies between environmental health and economic dynamics.

1.4 Methodology

Methodology for pollution and economic activity The study utilizes a comprehensive dataset, combining yearly district-level river pollution measurements from the Central Pollution Control Board (CPCB) with socio-economic data from the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). The CPCB data provides a detailed account of the water quality in the Ganga River across various districts, while SHRUG offers granular socio-economic and infrastructural information. Upon acquiring the SHRUG dataset, significant manipulation and cleaning were performed to align it with the district-level granularity of the CPCB pollution data. This involved extracting and reconciling district names and other relevant geographic identifiers, from Database of Global Administrative Areas (GADM), to ensure consistency across both datasets. The meticulous preparation of the SHRUG data ensured that the socio-economic variables corresponded accurately to their respective districts. The cleaned and manipulated datasets were then merged to form a panel data structure. This panel data encompasses multiple years and districts, enabling the study to investigate changes over time and across different geographic locations. The data were then filtered to focus specifically on 121 districts situated along the banks of the Ganga River, representing the areas most directly impacted by its pollution levels. Prior to regression analysis, an exploratory data analysis (EDA) was conducted to uncover patterns, detect outliers, and understand the underlying structures of the data. This critical step provided initial insights and informed the subsequent modeling approach.

Using the filtered panel data, the study employs Panel Ordinary Least Squares (PanelOLS) regression models to investigate the impact of various pollution measures on the mean values of district-level nightlight intensity per capita. The regression models are extended to include lagged effects, assessing the impact of pollution on economic activity over several years following the measurement of pollution levels.

To account for potential heteroskedasticity and autocorrelation, the regressions are conducted with standard errors clustered at the district level. This adjustment ensures that the estimated coefficients are robust and the inference is reliable.

The study further refines the analysis by creating logical subsets of data, such as comparisons between districts in the upstream versus downstream sections of the river, as well as contrasting districts with higher industrial activity against those with less. This stratification allows for a nuanced understanding of how pollution's impacts may vary in different contexts and under varying economic conditions.

Methodology for pollution and health data This segment of the study extends the analysis to explore the relationship between river pollution and health outcomes, specifically focusing on mortality due to diseases. To this end, the methodology integrates yearly district-level river pollution data from the CPCB with state-level health data from the Statistical Book of India. The latter provides comprehensive records of deaths attributed to various diseases in each state for a given year.

The primary challenge involved aligning the CPCB's district-level pollution data, which includes specific measures such as nitrate levels, Biochemical Oxygen Demand (BOD), and drinkability indicators, with the state-level health outcomes. To bridge this gap, the pollution data were meticulously mapped to correspond with the state names and years provided in the health outcomes dataset. This mapping facilitated a multi-level analysis, allowing the study to assess the impacts of river pollution on health outcomes at a more granular level, despite the initial scale difference between the health and pollution datasets.

1.5 Data

River pollution data The dataset employed in this study is a comprehensive collection of annual pollution metrics provided by the CPCB [2], covering the years 2012 to 2021. The CPCB's mandate includes regulating and monitoring pollution in the country's water bodies, and the data reflects this through a detailed, district-wise breakdown of pollution levels in the Ganga River. The measurements encompass a range of indicators that collectively offer insights into the river's quality and ecological health. These may include, but are not limited to, BOD, chemical oxygen demand (COD), total dissolved solids (TDS), pH levels, coliform counts, and the presence of heavy metals and other hazardous substances.

The district-wise granularity of the data allows for a localized examination of pollution, facilitating a more precise analysis of the socio-economic impacts at a micro level. This specificity is crucial for understanding how pollution affects different sections of the river differently, which, in turn, has varied implications for the economic activities and social conditions of the adjacent districts.

Given the longitudinal nature of the dataset, the study can also track changes over time, identifying trends, spikes, or decreases in pollution levels. This temporal aspect enables the examination of the effects of any regulatory changes, industrial activities, or conservation efforts that may have occurred during this period.

Nightlight data The SHRUG dataset provides gridded night lights data that are widely used as a proxy for economic activity. The night lights data are aggregated and calibrated for consistent time series estimation and are available annually from 1992 to 2021 across U.S. Air Force Defense Meteorological Satellite Program (DMSP) and Visible Infrared Imaging Radiometer Suite (VIIRS). The Colorado School of Mines, Earth Observatory Group (EOG), produces the data, and they are available at the national level and can be aggregated at the Shrid, District, Subdistrict, and Village/Town levels. The night lights data are used to estimate economic activity, electrification, and other socio-economic indicators when time series data on economic activity are otherwise unavailable[3].

While the night lights data are not a direct measure of economic activity, they have been used as a proxy for economic activity in several studies[4]. For example, a study used night lights data to estimate the economic impact of floods in different regions worldwide[5]. Another study used night lights data to analyze the relationship between firms, poverty, and night lights in India[6]. The study found that night lights data can be used to estimate the number of firms in a region and can be used as a proxy for economic activity.

Health Data The health data utilized in this study are derived from the Central Statistics Office (CSO)[7], which compiles the Statistical Year Book of India. This comprehensive directory encompasses reports across multiple domains, including Education, Agriculture, Public Finance, and Health and Family Welfare. Specifically, this research employs data from the report numbered 30.15, titled "NUMBER OF CASES AND DEATHS DUE TO DISEASES," which provides state-level data on deaths due to diseases for the years 2008-2015. This dataset was selected to examine the direct physiological impacts of river pollution on populations in nearby states. The underlying assumption is that in areas where river pollution is relatively low, such as Uttarakhand, the incidence of disease-related deaths would be lower. Conversely, in regions where the river becomes more polluted, as seen in stretches through Uttar Pradesh, it is hypothesized that the number of disease cases and deaths would increase proportionally. Therefore, a panel data structure combining river pollution data with mortality figures, stratified by state and year, has been created for analysis.

1.6 Results

The ultimate goal of this study is to measure and understand the health and socio-economic impacts of river pollution, specifically within the context of the Ganga River in India. This will be achieved by analyzing the relationship between river pollution and economic activity, with the latter proxied by nightlight intensity. The study sets out to test two primary hypotheses: the traditional hypothesis that increased pollution leads to decreased economic activity, and the alternative hypothesis that suggests a potential increase in economic activity in polluted areas due to factors such as industrial agglomeration and urbanization.

Operationalization of Variables

To facilitate this analysis, the study operationalizes variables as follows:

Dependent Variables:

- For the analysis of economic activity, the dependent variable is Mean Nightlight Intensity per capita at a District Level, measured using satellite imagery to capture the luminosity levels across districts along the Ganga River, reflecting the economic vibrancy of each region.
- For examining health impacts, the dependent variable is Deaths Due to Diseases in a Year at State Level, aggregated from the "Statistical Year Book India," indicating the health consequences of river pollution.

Independent Variables: The independent variables for both analyses include measures of river pollution derived from the CPCB dataset: BOD, Nitrate, and a binary variable not-drinkable, indicating whether a district's river stretch is considered polluted or not. These variables act as indicators of the pollution level in the Ganga River.

Control Variables: A set of control variables is utilized to isolate the effect of river pollution on the dependent variables. For the economic analysis, controls include District Population from the 2011 census and Temperature, along with Fixed Effects for District and Years to account for unobservable heterogeneity across districts and temporal trends. In the health impact analysis, Temperature is excluded as a control variable due to its lesser relevance to the direct health outcomes being studied.

Analytical Setup

The study employs two distinct analytical setups to explore the impacts of river pollution:

Economic Activity Analysis: This analysis investigates the relationship between river pollution and nightlight intensity as a proxy for economic activity. It utilizes panel data regression models to estimate the effect of pollution measures (BOD, Nitrate, and not-drinkable) on mean nightlight intensity at the district level, incorporating the specified control variables and fixed effects. This setup aims to test the hypotheses regarding the socio-economic repercussions of river pollution.

Health Impacts Analysis: Similarly, the relationship between river pollution and health outcomes is analyzed using panel data regression models, with the state-level annual deaths due to diseases as the dependent variable. The same pollution measures serve as independent variables, albeit without Temperature as a control in this model. This analytical setup seeks to quantify the direct health impacts of river pollution, examining how varying levels of pollutants correlate with mortality rates due to diseases.

1.7 Related Work

Health, economic growth, and water pollution are closely linked. Nearly all industries, are dependant on clean water and create pollution as a by-product. It is important that this pollution is properly treated before dumping it back to the rivers. According to the estimates from Ministry of Water Resources (MoWR), around 7-8 % of the total freshwater is used by industries in India[8]. The industrial sector is second highest consumer of water after agriculture. Majorly this water is used for process and cooling requirements.

Studies have shown that an increase in river pollution can lead to a decrease in economic activity and socio-economic growth in the affected regions. When rivers become heavily polluted, regions downstream can experience reductions in economic growth, losing between 1.4 and 2.5 percent of economic growth[9]. A 2023 study found that extreme weather conditions due to climate change deteriorates river water quality. The reasons include hydrological alterations, rises in water and soil temperatures, and interactions among hydroclimatic, land use and human drivers[10].

Moreover, heavily polluted water can reduce economic growth by up to a third in some countries, worsening health conditions, reducing food production, and exacerbating poverty in many countries[11]. Children exposed to high nitrate levels can experience hindered growth and brain development, affecting their health and diminishing their earning potential as adults. Specifically, for each additional kilogram of nitrogen fertilizer used per hectare, childhood stunting can rise up to 19 percent and potential adult earnings may decrease by up to 2 percent[11].

The tourism industry can also lose close to $ 1 billion each year, mostly through losses in fishing and boating activities, as a result of water bodies that have been affected by nutrient pollution and harmful algal blooms[12].

Understanding the relation between water pollution, specifically river water and income inequality is important in regions where the major source of water is the said river. Traditionally we have lived close to rivers or lakes for easy access to freshwater. Today, however, more than 50% population lives in urban areas, and water can be directed to tens of kilometers via pipelines[13]. Over time, the relationship between human populations and fresh water bodies has changed due to the pollution of water bodies, socioeconomic factors like urbanisation, income inequality, and cultural reasons[14].

A major study published in June 1999 [15] summarizes the socio-economic effects of water pollution in the Danube river basin, Europe. They look at multiple industries affected by changing pollution levels like fishing, tourism, shipping, hydro-power etc. This report also evaluates the current water demand and supply and projects the planning 10-20 years in the future.

The work done by world bank in 2019[16] is phenomenal in finding the economic impacts and sources of water pollution. Some of the major outcomes of their analysis include:

- When BOD – a measure of how much organic pollution is in water and a proxy measure of overall water quality – passes a certain threshold, GDP growth in

downstream regions is lowered by a third.

- In middle-income countries – where BOD is a growing problem because of increased industrial activity - GDP growth downstream of highly polluted areas drops by half.
- Nitrate exposure in infancy: wipes out much of the gain in height seen over the past half-century in some regions and harms children even in areas where nitrate levels are deemed safe.
- This report also reveals that enough food is lost due to saline waters each year to feed 170 million people every day – that's equivalent to a country the size of Bangladesh. Such a sizable loss of food production to saline waters means food security will continue to be jeopardized unless action is taken.

2 Context

The water quality crisis in developing countries, such as those in the Ganga river basin in India, has profound effects on public health, the economy, and overall societal well-being. Poor water quality can lead to serious health issues, disproportionately affecting the population in these nations. The lack of proper treatment systems and the consumption of contaminated water can result in various health conditions, including skin discoloration, nervous system and organ damage, developmental effects, and kidney failure. These health issues can lead to increased healthcare spending, reduced workforce productivity, and economic damage, as citizens are unable to attend school or work due to their health conditions [17]. In developing countries, up to 80% of illnesses are linked to inadequate water and sanitation, highlighting the significant impact of poor water quality on public health[18]. Moreover, the lack of access to clean water and sanitation disproportionately affects women and girls, further exacerbating gender inequality and limiting opportunities for their physical safety and security[19]. The pollution of water sources is a significant challenge in developing countries, with factors such as poor wastewater management, lack of political will, under-investment, and industrial and agricultural pollution contributing to the water quality crisis. India's most sacred river faces significant challenges, primarily due to its diminishing water levels. This reduction is a result of water being extracted for irrigation at a pace faster than what the rainy season can replenish. The situation is further exacerbated by the construction of over 300 dams and diversions along the river and its tributaries, which disrupt the natural flow of the river[20]. Additionally, studies have shown that there is a positive effect of water pollution on inequality, indicating that water pollution can further perpetuate existing socioeconomic disparities in these nations[21].

2.1 Impact on World economy

River pollution's global ramifications extend beyond immediate health concerns, intertwining with climate change to exacerbate water scarcity and quality issues. As climate patterns shift, increased incidences of floods and droughts further contaminate water sources, spreading pollutants over wider areas and affecting larger populations. This dynamic intensifies the urgency for comprehensive environmental policies that address both pollution control and climate adaptation. The interconnection between river pollution and climate change underscores the need for global cooperation in tackling these environmental challenges, ensuring sustainable water management and protection for future generations[22]. A modelling study has found that up to 5.5 billion people worldwide could be exposed to polluted water by 2100, highlighting the urgent need for action to ad-

dress the global water quality crisis[23]. Environmental pollution significantly increases the concentration of poor self-rated health, physical discomfort, and chronic disease among populations, further perpetuating existing socio-economic health inequalities[24]. Water degradation threatens human health, reduces ecosystem functioning, and hinders socio-economic growth.

A World Bank report found that heavily polluted water is reducing economic growth by up to a third in some countries, worsening health conditions, reducing food production, and exacerbating poverty in many countries[16]. The health costs relating to water pollution cost the country's economy upto £7.5 billion per year. The 2019 study found that a lack of clean water limits economic growth by one-third and that when BOD crosses a certain threshold, GDP growth in downstream regions drops by as much as a third because of impacts on health, agriculture, and ecosystems. The report also highlighted the impact of nitrogen pollution on human capital, stating that early exposure of children to nitrates affects their growth and brain development, impacting their health and adult earning potential. The study emphasizes the need for immediate global, national, and local-level attention to water pollution to enable countries to grow faster in equitable and environmentally sustainable ways.

2.2 Impact on Indian economy

The pollution of rivers in India has had significant effects on the economy, public health, socio-economics, and inequality. The contamination of water bodies has resulted in the spread of various infectious diseases, such as typhoid, cholera, hepatitis, E. coli, dysentery, and salmonella, as well as malnourishment in children[25].

According to Niti Aayog, an Indian government public policy think tank more Indians die prematurely due to the impacts of environmental pollution, and roughly 200,000 Indians lose their lives every year due to health problems caused by drinking contaminated water[26]. The effects of river pollution are also linked to socio-economic inequality, as vulnerable populations, particularly those in rural areas, are disproportionately affected by the lack of access to clean water and the prevalence of waterborne diseases[21].

More than 10.5 million gallons of wastewater flow into rivers and other watercourses in India[21], posing a serious threat to the country's economy and society. The health costs of water pollution in India are estimated to be about $ 6.7–8.7 billion per year, with more than 400,000 Indians dying from diarrheal illness due to inadequate sanitation and hygiene[27]. Additionally, roughly 200,000 Indians lose their lives every year due to health problems caused by drinking contaminated water[28]. Water pollution in India is associated with a 9% drop in agricultural revenues and can account for the loss of up to half of GDP growth in middle-income countries like India[28]. Furthermore, pollution upstream could cut economic growth in downstream regions by nearly a half percentage point[29].

The situation calls for urgent and comprehensive measures to combat water pollution and ensure access to clean and safe water for all. The implementation of stricter regulations, wastewater treatment plants, awareness campaigns, and community-based

Figure 2.1: Map of the Ganga basin. Source: wiki[32].

initiatives is crucial to addressing this multifaceted issue[30][31].

2.3 River pollution - Ganga

The Ganga is a sacred river in India, but it faces severe water quality degradation due to various factors such as urbanization, industrialization, and wastewater discharge. In this section we will introduce the river, its geography, significance over the years and what kind of pollution challenges it faces today.

Geographical overview The Ganga Basin shown in Figure 2.1 is India's most extensive river basin, encompasses 26% of the nation's territory, spanning 861,404 square kilometers, and sustains roughly 43% of its populace, amounting to 448.3 million individuals according to the 2001 census. Positioned between 73°02' to 89°05' East longitudes and 21°06' to 31°21' North latitudes, it covers a total area of 1,086,000 sq km across India, Nepal, and Bangladesh, with India housing about 79% of this basin. It stretches across 11 Indian states, including Uttarakhand, Uttar Pradesh, Madhya Pradesh, Rajasthan, Haryana, Himachal Pradesh, Chhattisgarh, Jharkhand, Bihar, West Bengal, and Delhi. The National Ganga River Basin Projects (NGRBP), supported by World Bank funding under the aegis of the National Mission for Clean Ganga (NMCG), primarily concentrates on five key states along the main course of the Ganga River: Uttarakhand, Uttar Pradesh, Jharkhand, Bihar, and West Bengal[33]. Detailed information on the drainage area allocated to each state is presented in the accompanying Table 2.1.

Table 2.1: Drainage Area by State in India. Source: NMCG [33]

States	**Drainage Area (km^2)**
Uttarakhand and Uttar Pradesh	294,364
Madhya Pradesh and Chhattisgarh	198,962
Bihar and Jharkhand	143,961
Rajasthan	112,490
West Bengal	71,485
Haryana	34,341
Himachal Pradesh	4317
Delhi	1484
Total	861,404

The Ganga Basin, distinguished by its expansive alluvial trough[34], is noteworthy for several reasons:

- It is home to one of the most significant aquifers in quantitative terms.
- The water quality is relatively high, although it decreases as one moves downstream towards the river's outfall.
- Near the Himalayan foothills, the water quality is excellent, benefiting from continuous recharge from Himalayan streams.

Demography has an important bearing on the state of the river as it is significantly affected by the population living within the basin. Average population density in the Ganga basin is 520 persons per square km as against 312 for the entire country (2001 census). Major cities of Delhi, Kolkata, Kanpur, Lucknow, Patna, Agra, Meerut, Varanasi and Allahabad are situated in the basin. The cities in the basin have large and growing populations and a rapidly expanding industrial base. The summary of urban population in the states covering Ganga basin is given in Table 2.2. It can be seen that between 2001 and 2011, urban population increased by 30% approximately. This trend is likely to continue. The pollution load is also expected to increase correspondingly[33].

The Ganga River is a trans-boundary river that flows through India and Bangladesh, stretching approximately 2,525 km (1,569 mi) in length[32]. It originates in the western Himalayas and flows through one of the most fertile and densely populated regions in the world[**?**]. The Ganga River basin is home to more than four hundred million people[**?**]. Some key geographical features of the Ganga River include:

- Tributaries: The Ganga has numerous tributaries, such as the Gomti River, Ghaghara River, Gandaki River, Kosi River, Yamuna River, Son River, Punpun, and Damodar[32].
- Hydrology: The hydrology of the Ganga River is complex, especially in the Ganga Delta region, which results in different ways to determine the river's length, discharge, and drainage basin size[32].

Figure 2.2: The confluence of Alknanda river and Bhagirathi river, which thereafter flow on as the Ganga river.

Type of Industry	*Total Units*	*Water Consumption (MLD)*	*Wastewater Generation (MLD)*
Chemical	27	210.9	97.8
Distillery	23	78.8	37.0
Food, Dairy & Beverages	22	11.2	6.5
Pulp & Paper	67	306.3	201.4
Sugar	67	304.8	96.0
Textile, Bleaching & Dyeing	63	14.1	11.4
Tannery	444	28.7	22.1
Others	41	168.3	28.6
Total	764	1123	501

Source: CPCB (2013a)

Figure 2.3: Industrial water consumption and wastewater generation in Ganga basin. Source: wiki [32]

- Course: The Ganga River flows through the Gangetic Plain, receiving the waters of many tributaries along the way. The first point from where it actually starts being called Ganga from the confluence of two rivers, as can be seen in Figure 2.2, is Devprayag in Uttrakhand. Midway in its course, near Allahabad, it is joined by one of its chief tributaries, the Yamuna (Jumna) River[33].
- Delta: The Ganga River empties into the Bay of Bengal, where its mouths form a vast delta. At the delta, it is joined by the southward-flowing Brahmaputra River[32].
- Ecology: The Ganga River is home to approximately 140 species of fish, 90 species of amphibians, and various reptiles and mammals[32].

Cultural and Historical Significance The Ganga River is considered one of the most sacred rivers in the world, with great religious and historical significance[35]. In Hinduism, the river is personified as the goddess Ganga, and it is believed that bathing in the river can cleanse a person's soul of all past sins and cure illnesses[32]. The river is also considered a "life-giving river" and is called Mother Ganga[32]. Many Hindus believe that life is incomplete without bathing in the Ganga at least once in their lifetime, and people travel from distant places to immerse the ashes of relatives in the waters of the Ganga[32]. The river is also important for irrigation, transportation, and agriculture, with its fertile plains feeding hundreds of millions of people in India and Bangladesh. However, the river is now heavily polluted with human and industrial waste, which poses a threat not only to health but also to faith. Despite this, the Ganga River remains an important symbol of India's cultural and religious heritage.

The Ganga River has been important historically, with many former provincial or imperial capitals located on its banks or the banks of tributaries and connected waterways, such as Pataliputra, Kannauj, Sonargaon, Dhaka, Bikrampur, Kara, Munger, Kashi, Patna, Hajipur, Delhi, Bhagalpur, Murshidabad, Baharampur, Kampilya, and Kolkata[32].

2.3.1 The Ganga River's Ecosystem and Uses

Natural Ecosystem The Ganga River, renowned for its self-purification ability, benefits from high levels of dissolved oxygen (DO) and the presence of radioactive radon. This ecosystem supports a significant population of macrophages-parasites, believed to combat bacteria effectively[36]. However, the diminished flow of the Ganga and Yamuna Rivers, combined with pollution from untreated industrial and municipal waste, agricultural runoff, and decreased oxygen levels, has severely impacted the river's innate capacity to cleanse itself, leading to a decline in water quality[37].

Major Uses of the River The Ganga River has played a significant role in the lives of people living nearby, providing various benefits and services. Some of the major uses[32] of the Ganga River include:

- Irrigation: The Ganga River has been used for irrigation since ancient times, both during floods and through gravity canals. This has increased the production of crops such as wheat, sugarcane, cotton, and oilseeds[38].
- Fishing: The river is home to approximately 140 species of fish, providing a source of food and livelihood for local communities[39].
- Transportation: The Ganga River has been an important transportation route since ancient times, facilitating the movement of goods and people[40].
- Religious and cultural significance: The Ganga River is considered sacred in Hinduism, and many people travel from distant places to immerse the ashes of relatives in its waters. The river is also worshiped as the Mother Ganga, and several sites along its banks are considered especially sacred.
- Agriculture: The fertile plains of the Ganga River support one of the most fertile and densely populated regions in the world, feeding hundreds of millions of people in India and Bangladesh.
- Tourism: The Ganga River attracts millions of pilgrims and tourists every year, contributing to the local economy.
- Hydropower: The river's flow can be harnessed for generating electricity, providing a source of renewable energy.
- Flood control: The Ganga River's floodwaters can be controlled and managed to prevent damage to human settlements and infrastructure.

These uses have contributed to the development and growth of the communities living along the Ganga River, making it an essential resource for their daily lives and well-being.

2.3.2 Pollution Challenges

Sources of Pollution India's rapid economic progress and burgeoning population have taken a heavy toll on the Ganga and its tributaries, with urbanization, industrialization, and extraction for irrigation seriously degrading and depleting the water[41].

The river is polluted in several segments, with the worst affected stretch being between Kannauj and Allahabad. The main water quality issues are organic pollution indicated by BOD and pathogens indicated by coliform count[42]. The pollution is attributed to the disposal of human sewage, industrial waste, and increasing population density. The Ganga provides water to about 40% of India's population across 11 states, serving an estimated population of 500 million people[43]. The ongoing pollution poses a significant threat to human health and the environment. To address this, the Indian government has launched the Namami Gange program, which includes investments to prevent sewage and industrial effluent from entering the river untreated and aims to improve water quality in major cities along the river. The program has shown some early successes, and the government is also promoting sustainable farming and engaging communities to reduce pollution and overextraction of river water[44][45].

Impact on Water Quality The impact of pollution on the Ganga' water quality is profound and multifaceted. Elevated levels of BOD and coliform bacteria not only indicate severe organic pollution but also point to the presence of harmful pathogens, making the water unsafe for human consumption and bathing. This deterioration affects the river's aquatic life, leading to diminished biodiversity and the collapse of local fisheries, which many communities depend on for their livelihood. Furthermore, the decrease in water quality exacerbates the challenge of securing clean drinking water, with significant implications for public health. Pollutants such as heavy metals and chemicals from industrial waste contribute to long-term health issues, including cancer, liver and kidney damage, and neurological disorders among the population relying on the river for their water needs.

The decline in water quality also has a direct impact on agriculture, as polluted river water used for irrigation can contaminate soil and crops, leading to decreased agricultural productivity and food safety concerns. Moreover, the spiritual and cultural significance of the Ganga is tarnished by pollution, affecting millions who partake in religious rituals and ceremonies along its banks. The visible pollution and the resulting public health alerts often lead to a reduction in religious tourism, affecting the local economy.

Water Quality and Pollution Measures in the Ganga Basin

The Ganga Basin's water quality is influenced by various factors, including geographical location, human activities, and environmental policies. Understanding the health of water bodies within the basin requires comprehensive monitoring of water quality

parameters, categorized into physical, chemical, and biological aspects. Each category plays a pivotal role in assessing the overall ecological balance and identifying pollution sources.

Key Water Quality Parameters and Their Significance Water quality parameters are indispensable for evaluating the ecological health of water bodies. They encompass:

Physical Parameters

- **Temperature:** Crucial for the solubility of oxygen and metabolic rates of aquatic life. Fluctuations can significantly impact ecosystem stability.
- **Turbidity:** Indicates water cloudiness due to suspended particles. High turbidity levels can obstruct light penetration, affecting aquatic plants and organisms.
- **Conductivity:** Reflects the water's ion concentration, serving as an indirect measure of its salinity and overall ionic makeup.

Chemical Parameters

- **pH:** A measure of water's acidity or alkalinity, affecting mineral solubility and biological activity. It's a fundamental parameter for maintaining aquatic life.
- **DO:** Essential for aquatic organisms; low DO levels are indicative of pollution and can lead to eutrophication.
- **Nitrate:** High nitrate levels can cause eutrophication, posing risks to aquatic ecosystems and potentially leading to oxygen depletion.

Biological Parameters

- **Bacteria:** The presence of fecal coliforms is a common indicator of water contamination, signaling potential health hazards.
- **Algae:** Algal blooms can deplete oxygen levels, disrupting the aquatic environment and harming species dependent on these waters.

These parameters are crucial for monitoring and maintaining the health of aquatic ecosystems and ensuring the safety of water for various uses, including drinking, industrial processes, and supporting aquatic life.

Specific Pollution Concerns in the Ganga Basin

The Ganga River faces significant pollution challenges, notably from nitrate and BOD, which stem from various sources and exert considerable pressure on the aquatic ecosystem and public health.

Nitrate Pollution

Nitrate levels in the Ganga River raise considerable environmental and health concerns[46], primarily sourced from:

- **Agriculture:** Fertilizer use significantly contributes to nitrate levels.
- **Industrial waste:** Discharges from tanneries, sugar and distillery, and pulp and paper mills enrich the river with nitrates.
- **Urban wastewater:** Household sewage adds to the river's nitrate pollution.
- **Animal husbandry:** Livestock and poultry waste further exacerbates nitrate pollution.

The adverse effects[47] of nitrate pollution include:

- **Eutrophication:** Elevated nitrate levels can trigger algal blooms, depleting oxygen and disrupting aquatic life.
- **Human health risks:** Nitrates pose risks like methemoglobinemia, especially in infants.
- **Fish kills:** Disrupted aquatic ecosystems due to high nitrates can result in fish deaths.

Mitigating nitrate pollution requires reducing agricultural, industrial, and urban wastewater inputs through best management practices and regulatory measures.

BOD

Biochemical oxygen demand is an analytical parameter representing the amount of DO consumed by aerobic bacteria growing on the organic material present in a water sample at a specific temperature over a specific time period. The BOD value is most commonly expressed in milligrams of oxygen consumed per liter of sample during 5 days of incubation at 20 °C and is often used as a surrogate of the degree of organic water pollution[48]. BOD is an indicator of the organic matter pollution, with high levels signifying reduced water quality[49]. Key contributors include:

- **Domestic waste:** Waste generated from populations connected to a sewer system. Considered a point source directly discharged into the stream network.
- **Industrial emissions:** Emissions from industries, including those connected to UWWTPs and large facilities that treat and discharge waste directly. Data from the European Pollutant Release and Transfer Register database (E-PRTR; EEA, 2018).

- **Livestock waste:** A major source of BOD pollution, estimated from global livestock distributions and accounting for different types of livestock and their waste[50][51].
- **Urban wash off:** BOD from urban runoff, treated by UWWTPs but can overflow during intense storms. Urban BOD was estimated as proportional to annual urban runoff volume.
- **Natural areas emissions:** Organic matter from natural areas contributes to BOD through litter and organic matter washed off into freshwater systems.

The consequences of high BOD levels encompass:

- **Eutrophication:** High BOD levels, indicating the presence of organic pollutants, can lead to the excessive growth of algae and aquatic plants. This process depletes the oxygen in the water, harming aquatic life[52].
- **Fish Kills:** The depletion of oxygen in water due to eutrophication can result in fish kills, as aquatic organisms suffocate from lack of dissolved oxygen[53].
- **Human Health Risks:** Polluted water with high BOD can contain pathogens and contaminants that pose risks to human health if used for drinking, bathing, or other domestic purposes, leading to waterborne diseases[54].
- **Ecosystem Damage:** Eutrophication and oxygen depletion from high BOD can disrupt aquatic food webs and damage sensitive ecosystems like lakes, rivers, and coastal areas.

Addressing BOD pollution entails minimizing inputs from key sources through improved waste management and stricter industrial regulations.

Measurement Techniques for Water Quality Parameters in the Ganga Basin

To accurately monitor and assess the water quality of the Ganga River CPCB[55] uses a standard guidelines for water quality monitoring. These tools enable the precise measurement of key water quality parameters, contributing to effective management and remediation efforts.

1. **Temperature:** Measured with thermometers or temperature sensors, providing data critical for assessing the river's thermal conditions and its effects on aquatic life.
2. **Turbidity:** Determined using turbidimeters, these devices quantify the light scattering by suspended particles, offering insights into the water's clarity and potential contaminants.
3. **Conductivity:** Conductivity meters are used to measure the water's electrical conductivity, which is indicative of its ion concentration and overall salinity.

4. **pH:** The acidity or alkalinity of the water is measured with pH meters, essential for understanding the chemical balance of the river and its suitability for various biological processes.

5. **DO:** DO meters gauge the oxygen levels dissolved in the water, crucial for maintaining healthy aquatic ecosystems and indicating potential pollution.

6. **Nitrate:** Nitrate concentrations are measured using nitrate meters or spectrophotometers, essential for identifying eutrophication risks and pollution sources.

7. **BOD:** BOD meters measure the oxygen demand by microorganisms for organic matter breakdown, reflecting the river's organic pollution levels.

These measurement techniques provide invaluable data for identifying pollution sources and devising strategies to enhance the Ganga River's water quality, ensuring its health and sustainability for future generations.

2.3.3 Impact on Society and the Environment

Health and Livelihood The pollution of the Ganga River has dire consequences for both health and livelihoods across vast regions of India. Communities that depend on the Ganga for drinking water, bathing, and irrigation face increased risks of waterborne diseases such as cholera, hepatitis, and gastrointestinal infections. The ingestion and use of contaminated water not only lead to acute health crises but also contribute to long-term health problems, including chronic illnesses and decreased life expectancy. For millions, the river is also a source of livelihood through fishing, agriculture, and tourism. Pollution adversely affects fish populations and agricultural yields, undermining food security and economic stability for countless families. The decrease in water quality discourages tourism, especially religious and cultural tourism, which is a significant source of income for many communities along the river. The combined impact of these health and economic challenges places a considerable strain on the societal fabric, exacerbating poverty and limiting opportunities for development and improvement of living conditions.

Demographic Pressures The demographic pressures on the Ganga are immense and growing, with population density in the Ganga basin significantly higher than the national average. This dense population not only contributes to the pollution load through domestic waste and sewage but also demands more water for drinking, sanitation, and agriculture, leading to overextraction and further degradation of water quality. Urbanization and industrialization in the basin have accelerated these pressures, with cities expanding and more industries discharging their effluents into the river. The rapid growth of the urban population, which increased by approximately 30% from 2001 to 2011 as seen in Table 2.2, is expected to continue, further exacerbating pollution and resource depletion. This demographic trend poses a significant challenge to managing the river's resources sustainably. It calls for urgent and comprehensive measures to balance

Table 2.2: Urban Population and Number of Statutory Towns in Various States (2001-2011). Source: NMCG[33]

States	No of Towns	Population (2001)	Population (2011)
Bihar	143	8,681,880	11,758,016
Jharkhand	41	5,993,741	7,933,061
Haryana	91	6,115,304	8,842,103
Himachal Pradesh	58	595,581	688,552
Madhya Pradesh	394	15,967,145	20,069,405
Chhattisgarh	188	4,185,747	5,937,237
Rajasthan	205	13,200,000	17,048,085
Uttar Pradesh	670	34,539,582	44,495,063
Uttarakhand	80	2,179,074	3,049,338
West Bengal	138	22,427,251	29,093,002
Delhi	6	12,905,780	16,368,899

the needs of the population with the imperative of preserving the river's health. Efforts to improve water quality, promote sustainable use, and reduce pollution are critical to ensuring that the Ganga can continue to support the millions of lives that depend on it, now and in the future.

2.3.4 Efforts for Restoration and Improvement

Government Initiatives Efforts to clean up the Ganga including the Namami Gange program, which has been implementing various measures to improve water quality and reduce pollution. However, the problem of river pollution in India is still a significant and ongoing challenge, with 6614 million litres per day (MLD) wastewater with an organic pollution load of 426 tonnes per day (TPD) is discharged into the Ganga river[56].

Impact of COVID-19 Lockdown The water quality of the Ganga has witnessed some improvement during the COVID-19 lockdown period, which led to a reduction in the discharge of industrial effluents into the river[57]. However, untreated sewage from urban locations continues to contribute to poor water quality downstream, affecting fish stocks and the fishing-dependent economy. Specific measures such as preventing direct disposal of sewage into the river, promoting sustainable farming, and engaging communities are being implemented to address these issues[58].

The pollution of rivers in India threatens not only the aquatic ecosystems but also the health, livelihood, and well-being of millions of people who depend on these rivers for their sustenance. Some of the main sources of water pollution in India include untreated sewage, agricultural runoff, and unregulated small-scale industry[59].

2.4 Nightlight as proxy for Income Inequality

Now that we know that river pollution causes socio-economic impacts on nearby communities, especially in terms of income inequality, we would like to measure the said impact. The data for river pollution measures are readily available through the CPCB data for the Ganga river. However, income and census data are not available at a district or smaller level for our time period. To study granular effects, we need data at least at the district level. To counter this, we employ a widely used technique of substituting economic activities with nighttime light intensity.

Studies have successfully identified significant correlations between nightlight intensity and socioeconomic variables. Henderson et al. [60], first in 2012, measured economic activity and income growth over time using satellite data on nightlights. Researchers have used nightlight intensity as a proxy for approximating poverty and income inequality by demonstrating linear correlations between them. They show that nightlight features in developing countries can provide a reliable decision boundary between the extremely rich and poor[61]. Anthony demonstrated that in 42 countries in Africa, regional income inequalities can be illustrated, indicating its effectiveness as a proxy for income per capita and wealth. Meanwhile, Henderson[60] constructed a Gini index, and Elvidge[62] developed a Nightlight Development Index (NLDI) to measure income inequality, whereas Anthony constructed a Mean-Log Deviation (MLD), which is quite similar to NLDI [63]. To find relationship between nightlight and economic development in India Singhal et al.[64] shows that regional inequality measured by nightlight follows the Kuznets curve pattern. This implies that starting from low levels of socio-economic development or quality of institutions, inequality rises as regional socio-economic factors or quality of institutions improve, and with subsequent progress in socio-economic factors or quality of institutions, regional inequality declines. In our study, we did not build a national-level index for income inequality but directly worked with nightlight intensities (per-capita) and river pollution measures.

2.5 Economic concepts

2.5.1 Panel Data Analysis: An Overview

Importance and Application: Panel data, encompassing observations of entities across various time points, offers a robust foundation for economic research by accommodating the analysis of time-invariant characteristics. This dataset structure is instrumental in distinguishing between effects that vary within groups and those that persist across groups, thereby enhancing the accuracy of econometric evaluations.

Methodological Approach: Our research constructs a panel dataset capturing time-series data on nocturnal luminosity and pollution indices for districts along the Ganga River from 2012 to 2021. Utilizing this dataset, we apply econometric techniques, such as fixed-effects and random-effects models, to assess the influence of river pollution on economic vibrancy while addressing unobserved heterogeneity.

Limitations: Despite its merits, panel data analysis is not devoid of challenges. It may confront multicollinearity issues when incorporating numerous time-invariant predictors. Additionally, inaccuracies in data measurement or incomplete datasets can skew results. The need for extensive datasets also presents potential logistical and computational hurdles.

Ordinary Least Squares (OLS) & Heteroskedasticity

OLS regression stands as a foundational econometric tool designed to discern the relationships between a dependent variable and one or several independent variables. It aims to identify the line, or in more complex scenarios, a hyperplane that minimizes the sum of squared residuals, which are the differences between observed and predicted values. Some of the key characteristics are:

Simplicity and Interpretability: The straightforward nature of OLS models promotes their widespread application in preliminary economic analyses, offering intuitive interpretations of data relationships.

Assumptions: The OLS estimators are recognized as the best linear unbiased estimators (BLUE), given that certain core assumptions are satisfied. These include linearity of the relationship, absence of perfect multicollinearity, homoscedasticity (or uniform variance of error terms), and the normal distribution of error terms.

Applications: The utility of OLS regression spans various domains within economics, such as policy assessment, economic forecasting, and the investigation of diverse economic phenomena.

Limitations and Heteroskedasticity: While OLS regression is widely employed, it operates under stringent assumptions that, if violated, can lead to inefficient estimates and biased hypothesis tests. Notably, heteroskedasticity, which is characterized by non-uniform variance in the error terms, can undermine the estimator's efficiency and lead to fallacious standard error calculations, affecting the trustworthiness of confidence intervals and significance tests.

Addressing Heteroskedasticity: The existence of heteroskedasticity within a dataset calls for specialized solutions to preserve the integrity of statistical inferences. One such remedy is the use of robust standard errors, also known as White's standard errors. By adjusting the standard errors to account for the irregularity of variance among error terms, robust standard errors enhance the dependability of statistical testing in the face of heteroskedasticity. This approach does not rectify the heteroskedasticity itself but adjusts the estimation procedure to yield valid conclusions, ensuring that the OLS model remains a valuable instrument in econometric analysis.

Overview of Econometric Models

Econometric analysis requires careful model selection to examine the relationships between variables within panel data effectively. Distinguishing between fixed-effects, random-effects, and 2-Way Fixed Effects models is essential for accurately interpreting the dynamic interplay of these variables.

Choosing Between Models The decision between fixed and random effects models hinges on the specific research question and the nature of the unobserved heterogeneity within the data. The Hausman test plays a pivotal role in this decision, determining whether the unique errors are correlated with the regressors, which would suggest a preference for fixed-effects models.

Fixed-Effects Model Fixed-effects models excel in analyzing time-varying changes within entities, controlling for all time-invariant characteristics unique to each entity. They are optimal for within-entity analysis, providing a granular view of the impact of predictors over time.

Generalization: The model's ability to control for time-invariant unobservables within an entity facilitates the examination of multiple time-varying determinants of the outcome, enhancing causal inference capabilities.

Random-Effects Model Random-effects models treat variation across entities as random and are efficient under the assumption that this variation is uncorrelated with the predictors. They are suitable when the study aims to understand the impact of heterogeneity across entities.

Model Parameters: Random-effects models treat these entity-specific effects as random variables, allowing for the incorporation of time-invariant predictors in the analysis.

2-Way Fixed Effects Model The 2-Way Fixed Effects Model expands upon the fixed-effects approach by controlling for unobserved heterogeneity both across entities and over time. This model is highly advantageous for panel data analysis where controlling for both entity-specific and time-specific unobserved factors is necessary.

Application and Advantages:

Implementing a 2-Way Fixed Effects Model includes incorporating dummy variables for both entities and time periods to control for unobserved heterogeneity. Its advantages include comprehensive control over unobservables, enhanced causal inference, and flexibility in empirical research.

Considerations:

This model comes with considerations, such as increased dimensionality, potential challenges in capturing dynamic relationships, and concerns regarding external validity due to the removal of unobserved heterogeneity.

Model Selection

Model selection is informed by both theoretical underpinnings and empirical testing. The Hausman test is instrumental in guiding the choice, as it assesses the correlation between unique errors and regressors. A significant test result suggests that a fixed-effects model is more appropriate due to the correlation between entity-specific heterogeneity and the independent variables, ensuring estimator consistency.

The selection of an appropriate econometric model is a crucial step in the analysis of complex economic relationships. Fixed effects models offer a detailed analysis within entities, accounting for time-invariant unobservables, which provides a nuanced understanding of individual entity dynamics. Random effects models, on the other hand, allow for broader analyses that incorporate variations between entities. The 2-Way Fixed Effects Model synthesizes the strengths of both approaches, enabling a comprehensive analysis that controls for unobserved heterogeneity across both entities and time. Such an integrated methodology is vital for capturing the intricate relationships inherent in panel data.

Expanding upon this, our methodological approach is designed to rigorously examine the interplay between river pollution and economic vitality, using nighttime light intensities as a proxy for economic activity. By employing panel data analysis techniques and robustly addressing econometric challenges like heteroskedasticity, we aim to generate estimates that are not only statistically reliable but also hold significant economic implications. The findings of our study are intended to inform policymakers about the intricate ties between environmental health and economic growth, particularly within the context of India's development trajectory. The goal is to provide a solid empirical foundation for crafting policies that align environmental sustainability with economic objectives, ultimately guiding informed and effective decision-making for long-term development and prosperity.

Standard Error Clustering in Econometric Analysis

The Role of Clustering Standard Errors In the realm of econometric analysis, clustering standard errors is a pivotal technique for achieving robust standard errors, especially when observations within clusters may exhibit correlation. This adjustment is crucial for overcoming the limitations of the assumption that error terms are independently and identically distributed (i.i.d.) across observations, a common premise in OLS regression models.

Significance and Application Clustering standard errors is essential for yielding consistent standard errors in the presence of intra-cluster correlation. Such correlation often arises when observations within a specific group (e.g., geographical area or time frame) are more similar to each other than to those in different groups. Ignoring intra-cluster correlation can lead to underestimation of standard errors, resulting in overly optimistic

t-statistics and a heightened risk of Type I errors—falsely rejecting the null hypothesis when it is true.

Application of Clustering Standard Errors In our investigation into the impact of river pollution on nighttime light intensities, the application of clustering standard errors at the district level is paramount. This adjustment is crucial for acknowledging that observations within the same district are likely interrelated, potentially due to district-specific events or policy shifts that simultaneously influence river pollution levels and economic conditions, as inferred from nighttime lights data. By clustering standard errors by district, we can more accurately account for these localized correlations, enhancing the reliability of our regression analysis.

Implementation of Clustering Standard Errors The practical application of clustering standard errors occurs post-regression coefficient estimation. This involves recalibrating the covariance matrix of the regression coefficients to reflect the within-cluster correlation of error terms. Statistical software packages, including R and Python, provide dedicated functions for this purpose, facilitating the clustering of standard errors across various dimensions such as time or geographic location. This methodological step is critical for ensuring that our econometric analysis accurately reflects the underlying data structure and relationships, thereby providing more trustworthy insights into the effects of river pollution on economic activity as measured by nighttime light intensities.

Example: Clustering Standard Errors in Python Using the `statsmodels` library in Python, clustering standard errors after fitting a regression model can be implemented with the following code snippet:

```
import statsmodels.api as sm
from statsmodels.regression.linear_model import OLS

# Assuming 'X' is the matrix of independent variables and 'y' is the
    dependent variable
model = OLS(y, X).fit(cov_type='cluster', cov_kwds={'groups': data['
    District']})

# 'data['District']' is the variable indicating the cluster groups (
    districts in this case)
```

This example illustrates the process of fitting a regression model with the OLS method and adjusting for clustered standard errors based on district groupings. The key parameter here is `cov_type='cluster'`, which specifies the type of covariance estimator to use, and `cov_kwds={'groups': data['District']}`, which defines the clusters.

Advantages and Limitations of Clustering Standard Errors Clustering standard errors is indispensable for achieving more precise inferences in the presence of cluster-specific data patterns. This technique is particularly beneficial in panel data analysis, accom-

modating the multi-dimensional structure of such data and enhancing the robustness of statistical findings.

Limitations:

Notwithstanding its benefits, clustering standard errors may result in larger standard errors, potentially diminishing the statistical power of tests. Moreover, the reliability of clustered standard error estimates hinges on a sufficient number of clusters. Insufficient clusters could undermine the accuracy of these estimates, detracting from the reliability of the econometric analysis.

Employing clustering standard errors is a fundamental aspect of econometric analysis in panel data contexts, vital for substantiating the statistical inferences derived from regression models. This approach is crucial for addressing correlated observations within clusters, thereby refining the accuracy of research outcomes and reflecting more faithfully the economic realities under study.

Subsetting Data for Regression Analysis

Rationale: Subsetting data serves as a methodological strategy to dissect datasets into smaller, more homogenous groups based on specific criteria, such as demographic characteristics, geographic regions, or time periods. This segmentation aims to isolate subsets where the interplay between variables may exhibit distinct patterns or hold particular relevance.

Significance: Through subset regression analysis, variations in variable relationships that might be masked in aggregate data can be illuminated, uncovering subtle insights and facilitating a deeper comprehension of the variable dynamics within particular data segments.

Creating Subsets: The initial step in subset regression involves segmenting the dataset according to predetermined criteria relevant to the research question. For example, in examining the effects of an educational program, data might be stratified by age groups or types of educational institutions, enabling targeted analysis of the intervention's impact within specific populations.

Performing Regression Once the data is subdivided, regression analysis is conducted on each subset as though it were an independent dataset. This allows for the derivation of regression coefficients that are specific to each group, facilitating a comparison across subsets or against the coefficients from a regression analysis on the full dataset.

Example: Utilizing Python for subsetting data and performing regression analysis can be efficiently executed with the pandas library for data manipulation and the statsmodels library for statistical modeling:

```
1 import pandas as pd
2 import statsmodels.api as sm
```

```
3
4 # Assuming 'data' is the DataFrame containing the dataset
5 # Subset the data for a specific category
6 subset_data = data[data['Category'] == 'specific_category']
7
8 # Prepare the data for regression
9 X = subset_data[['independent_variables']]
10 y = subset_data['dependent_variable']
11 X = sm.add_constant(X) # Adds a constant term to the predictor
12
13 # Fit a regression model to the subset
14 model = sm.OLS(y, X).fit()
15
16 # Output the model's summary
17 print(model.summary())
```

Considerations:

- It is essential to ensure that each subset contains a sufficient number of observations to yield reliable estimates.
- The risk of multiple comparisons must be managed carefully. Performing numerous regressions increases the likelihood of encountering statistically significant findings merely by chance.

Limitations:

- Subsets with a limited number of observations might lead to imprecise regression estimates, hampering statistical power.
- Variations in estimates across subsets may stem from sampling variability rather than genuine differences.

Subsetting data for regression analysis offers a nuanced approach to exploring variable relationships, potentially unveiling specific patterns within subgroups that may be obscured in a comprehensive analysis. Nonetheless, it demands meticulous statistical consideration and thoughtful interpretation of the results, always aligning findings with the overarching goals of the research.

2.5.2 Interpreting PanelOLS Regression Results

For hypotheses predicting a relationship between variables (e.g., pollution and economic outcomes), coefficients are central. A negative coefficient suggests a negative correlation, particularly important for environmental studies where pollution may inversely affect economic factors. Coefficients, significant at a p-value below 0.05, indicate reliable relationships. Their magnitude reveals the responsiveness of the dependent variable to changes in the independent variable(s), offering insights into elasticity in economic contexts.

Understanding Statistical Indicators

- **Standard Error (SE)**: Indicates the estimate's precision, with lower values signifying greater confidence in the coefficient's accuracy.
- **P-value**: Assesses the hypothesis that the coefficient is significantly different from zero, with values less than 0.05 generally considered evidence of a statistically significant effect.
- **F-value**: Evaluates the model's overall fit, with higher values indicating that the model explains a significant portion of the variance in the dependent variable.
- **Parameter Estimates**: Reflect the effect size of predictors on the outcome, crucial for understanding the direction and magnitude of relationships.

Determining Correlation and Causality Initial analysis may establish correlation if the variables move together in a statistically significant manner. However, asserting causality requires further steps, such as using Instrumental Variable (IV) techniques or fixed effects models to control for unobserved variables, thereby strengthening causal inferences.

Coefficient Interpretation Understand each variable's coefficient as the expected change in the dependent variable for a one-unit increase in the independent variable, holding all other variables constant. This interpretation is essential for quantifying the impact of predictor variables on the outcome.

Significance Levels Recognize significance levels (e.g., * for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$) as indicators of how likely it is that the relationship observed in the sample can be generalized to the population. The stars or asterisks next to the coefficients in regression output tables serve as a visual cue for their statistical significance, guiding researchers in assessing the reliability of their findings.

Standard Errors View standard errors in parentheses next to coefficients as a measure of the variability of the coefficient estimate. Smaller standard errors suggest more precise estimates of the coefficient, thereby increasing confidence in the robustness of the findings.

Fixed Effects Acknowledge fixed effects (district and year) as controls for unobserved heterogeneity within those dimensions. By controlling for all unchanging characteristics within each district or year, the analysis focuses on within-group changes over time, which is crucial for identifying causal relationships in panel data.

Model Fit and Diagnostics Examine R-squared (R^2) and F-statistics to evaluate the overall fit of the model. R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 value suggests a better fit of the model to the data. The F-statistic assesses whether the overall regression model is statistically significant, comparing the model with no predictors (only an intercept) to the specified model.

Elasticity and Dynamic Relationships When interpreting regression results with both dependent and independent variables in logs, a coefficient represents the elasticity between these variables, measuring the percentage change in the dependent variable expected from a one percent change in the independent variable. Utilizing lags (past values) as independent variables helps capture delayed effects, indicating how past values of a variable influence the current value of another variable. This approach is crucial for understanding dynamic relationships over time, especially in economic and environmental studies where effects may not be immediate.

In the next chapter, we would look at the data sources used in this study and the processing done to get it in shape for our regressions.

3 Data

This study utilizes four major data sources: river pollution data, a dataset for measuring economic activity (through nightlight intensity), the GADM database for detailed geographic information of districts, and health and family welfare data to assess the impact of pollution on health to establish correlations and potential causation between environmental factors and socio-economic indicators.

In this chapter, we explain the data sources, how we use them in our study, processing and cleaning, and compiling them together to successfully run regression models on them to find relationships between variables. We start with the GADM data, then explain the pollution data, the nightlight data, and at the end the health and family welfare data.

3.1 GADM

The GADM[65] database is a high-resolution database of country administrative areas, with current and comprehensive data on administrative boundaries. It includes boundaries for multiple levels of subnational divisions for every country. The GADM project aims to produce a detailed and comprehensive global database of administrative areas, making it an invaluable resource for researchers, policymakers, and Geographic Information System (GIS) professionals.

Source The GADM database is compiled from various sources, including national and sub-national government databases, non-governmental organizations, and research institutions. The data are continually updated and refined to reflect changes in administrative boundaries, new subdivisions, and corrections. The GADM project is not affiliated with any government entity but relies on publicly available data and contributions from a global community of users.

Uses GADM data has diverse applications across disciplines such as geography, environmental science, public health, political science, and economics. Some of the primary applications include:

- Geospatial Analysis: GADM provides a foundational layer for mapping and spatial analysis in GIS, enabling researchers to overlay other geospatial data on administrative boundaries for detailed spatial analyses.
- Epidemiological Studies: Public health researchers use GADM data to study the spread of diseases within and across administrative boundaries, facilitating targeted interventions and resource allocation.

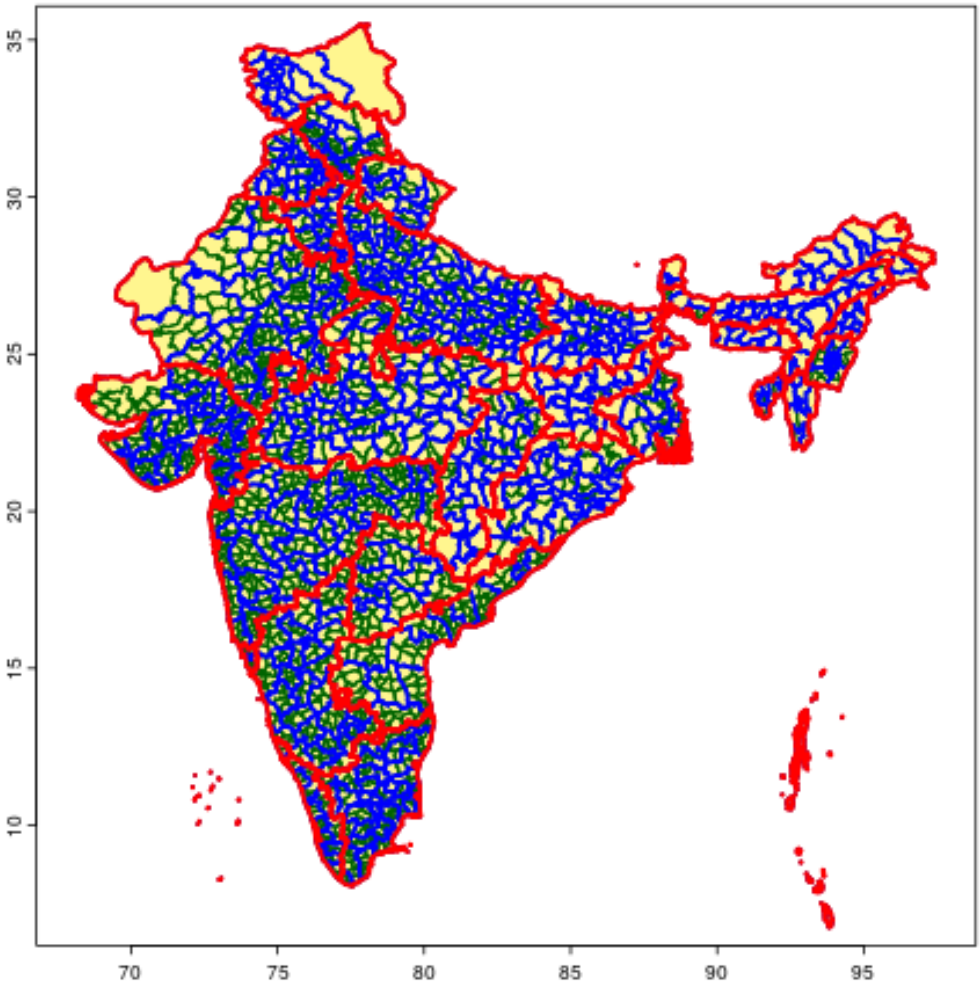


Figure 3.1: India GADM map. Source: GADM website[65]

- Environmental Conservation: Environmental scientists and conservationists utilize GADM boundaries to assess ecological patterns, biodiversity, land use changes, and conservation areas within specific administrative units.
- Socio-economic Research: Economists and social scientists employ GADM data to analyze socio-economic trends, policy impacts, and development indicators at national and sub-national levels.
- Emergency Response and Disaster Management: GADM data are critical for planning and executing emergency response strategies, allowing responders to understand the administrative context of affected areas.

Scope in this study In this study, we use the GADM database to provide detailed geographic information for districts represented in the pollution dataset. Utilizing the Python GADM library (GADMDownloader) and version 4 of the database, we download district-level (level 2) geo-information for India. This geographic information is crucial for subsequent analyses, such as identifying the nearest stations to the districts in our pollution dataset and exhaustive merging with nightlight data using standardized district names from the database.

3.1.1 Data preparation and Exploratory data analysis of GADM

This particular version of dataframe we use in our study starts with 666 rows and 16 columns, with geometrical information for India's level 2 (district) information. The important columns being NAME_1 (represents state names), NAME_2 (district names), geometry column (a multi-polygon datatype structure with longitude and latitude information).

The main preparation steps for this dataset include:

- Extract district centroid using the geometry column. The district centroid contains the combined Point data structure combining longitude, latitude. We use this information to calculate nearest pollution station codes later.
- Extract longitude and latitude for each district from district centroid.
- Filter the data for our relevant states by a filter on NAME_1 column.
- We remove columns such as ID_0 (country ID), ID_2 (district IDs), and ENGTYPE_2 (granularity of the data) that are not pertinent to our analysis.

To refine the integration of the GADM data with the pollution data at the district level, we utilized a fuzzy matching algorithm. This choice was driven by the recognition that district names could vary significantly across the two datasets due to differences in spelling, abbreviation, or translation. Fuzzy matching algorithms are designed to find matches that are "close" in a non-exact sense, which is ideal for overcoming these discrepancies.

The algorithm assesses the similarity between district names in the GADM dataset and those in the CPCB pollution dataset, assigning a score based on the closeness of a potential match. We then set a threshold score (0.9) to determine a match, ensuring that only districts with a high degree of name similarity are considered aligned. This method allows for a flexible yet accurate integration of datasets, significantly reducing the potential for mismatches or exclusions due to strict name discrepancies.

This fuzzy matching process was crucial for the integrity of our study, as it ensured that each district's pollution data from the CPCB could be accurately linked with its corresponding geographical information in the GADM database. By facilitating this precise integration, we were able to maintain a comprehensive and accurate dataset that includes only the relevant states for our analysis, enhancing the reliability and validity of our findings.

After processing, the final dataset comprises 170 rows and 6 columns, detailing district-level geographical information for 170 unique districts in proximity to the Ganga River and its tributaries.

3.2 River pollution data, CPCB

The CPCB is a government-run organization in India responsible for monitoring and controlling pollution in the country. The CPCB is responsible for enforcing environmental laws and regulations, conducting research and development, and providing technical assistance to industries and local governments[2]. The CPCB has been involved in various initiatives to clean up the Ganges River, including the National Mission for Clean Ganga (NMCG), which was launched in 2014 to clean up the river and its tributaries. NMCG also launched 'Namami Gange Programme' in 2014, a flagship programme by the union government to effectively abate the pollution, conservation and rejuvenation of the National River Ganga[33].

The water quality assessment under National Water Quality monitoring Programme (NWMP) is carried out with the objective of assessing the impact of possible sources of pollution on water quality of recipient water bodies thus most of the locations under NWMP are impact locations as outlined in the Guidelines for Water Quality Monitoring, 2017[55]. Water Quality is carried out by respective SPCBs/PCCs and sites are selected based on the criteria for identification of monitoring locations under NWMP. Water, being State subject , implementation of action plan for restoration or rejuvenation of water bodies is carried out by respective State Govt/UT Administration in their jurisdiction[66].

CPCB Observations

- The CPCB monitors water quality at 233 locations, focusing on parameters such as pH, Conductivity, Dissolved Oxygen (DO), BOD, Total Coliform, and Fecal Coliform.

- Organic pollution, evidenced by BOD and Coliform counts, is identified as the predominant concern, affecting the aquatic ecosystem.
- While the Himalayan Segment and the Diluted Segment exhibit comparatively good water quality, the lower segments face significant pollution challenges due to excessive pollutants and water extraction.

Scope in this study We are interested in studying the adverse health and socioeconomic effects of river pollution, specifically the largest river in India, the Ganga river. As CPCB is a central body, it publishes data for all major rivers of India, and groundwater observations too. In light of this, we chose only the observations from NWMP data from Water quality data, which contains yearly reports from 2012-2021. We selected our study timeframe as per the temporal scope of this data. For each year, we have extracted the data for the Ganga river and its major tributaries like Yamuna, Tons, Betwa, Chambal, etc.

River pollution data preparation

Now that we have the raw pollution data from CPCB for Ganga and its tributaries for the years 2012-2021, we want to clean, process, and concatenate this data so we can work with it combined with the nightlight data later on. We perform a series of data manipulation to make that happen, including but not limited to, filtering, grouping, deleting, removing duplicates, etc.

Data Preparation and Extraction The initial phase of our data analysis involved processing raw data obtained from the CPCB, which included annual datasets for all river systems in India. The process, documented in the notebook river_pollution_analysis.ipynb, entailed several key steps:

1. **Districts list:** We first use the list of districts on river Ganga and Tributaries. This list is prepared by NMCG [67], and it provides a comprehensive list of all the districts on Ganga and its tributaries. It has two relevant columns for us, the district name, and the river on which it lies. This has in total 139 districts, and provides as the basis of all our further analysis. This had some districts repeated multiple times because they lie on multiple rivers, so we groupby the list and get a list of unique districts counting 122 distinct ones.
2. **GADM data:** We then use the processed GADM data from above and merge the geographical information from GADM to our list of 122 districts, giving us a dataset with shape (122,8).
3. **Water quality station codes:** We then import the complete database of water quality station codes from CPCB, which provides geographical information of all the station codes working and collecting water quality measures in the country. It is an exhaustive list with: Water quality station code, Name of the monitoring

station, State/UT, Type of water body, Frequency of Monitoring, Latitude and Longitude information. The dataset has dimension of 4111 rows and 9 columns. We use this dataset to find the nearest stations for each district.

4. **Closest stations calculation:** We wrote a function to calculate the 5 nearest station codes and their distance in KMs for each of our unique 122 districts from their district centroids from GADM and the station codes location from the above dataset. This populated our dataframe with 10 new columns. Water quality monitoring station codes from the CPCB were associated with each district, identifying the five nearest stations per district and mapping them alongside their distances in kilometers. We mapped each district with 5 nearest station codes by calculating the distance between the district centroid calculated above and the geolocation of each station from CPCB directory of station codes. This approach has a few problems that sometimes the nearest station code according to the geolocation might not be the best station for gathering data as it might not be measured regularly, and aanother station might be more popular and measured more regularly or located at a better location. We can deal with this situation using the average measurements from all 5 stations to get a better placed values in future. For now, we proceed with the nearest station's measurements.

5. **Pollution data assembly:** We write another function to load the pollution data from multiple excel files and sheets within them. The data was first converted from pdf to excel, filtered for Ganga and its tributaries, and saved as excels. This function goes thorugh all the sheets, for all the years, and does quality checks to see if they have the same number of columns and data types and then concatenates them. The final data has a shape of (2630, 20). We merge this data with our original list of Ganga districts to get the pollution measurements, from 2012-2021 for these 122 districts basis the logic of closest station code first. So from the 5 closest stations for our districts, the loop runs and tries to find the polution measures, and fills data for whichever station code is available first. So priority is given the the closest station, then second closest and so on till we find the data. For some districts, in these years, there was no data recorded in the 5 closest stations, so we don't get any pollution data for them. These districts were: Begusarai, Gaya, Kaimur, and Purnea. The final dataset is then of size (767,39) as for each district and its matched station code, we have some years of data available, the maximum being 10. So for each unique district, we can have 10 rows maximum.

6. **Data conversions:** We convert each column to its proper data type. So pollution measures are conberted to numerics, year to datetime, etc.

7. **Not-drinkable:** The CPCB provides standards for water quality criteria, which classifies a body of water in the classes from A to E, with explanations for criteria in the Table 3.1. We classified a particular district in a particular year if in that year the water quality standards did not meet the criteria mentioned in class A in the table.

Table 3.1: Criteria for Classification of Surface Water Quality

Designated-Best-Use	Class of water	Criteria
Drinking Water Source without conventional treatment but after disinfection	A	Total Coliforms Organism MPN/100ml shall be 50 or less, pH between 6.5 and 8.5, Dissolved Oxygen 6mg/l or more, Biochemical Oxygen Demand 5 days 20°C 2mg/l or less
Outdoor bathing (Organised)	B	Total Coliforms Organism MPN/100ml shall be 500 or less, pH between 6.5 and 8.5, Dissolved Oxygen 5mg/l or more, Biochemical Oxygen Demand 5 days 20°C 3mg/l or less
Drinking water source after conventional treatment and disinfection	C	Total Coliforms Organism MPN/100ml shall be 5000 or less, pH between 6 to 9, Dissolved Oxygen 4mg/l or more, Biochemical Oxygen Demand 5 days 20°C 3mg/l or less
Propagation of Wild life and Fisheries	D	pH between 6.5 to 8.5, Dissolved Oxygen 4mg/l or more, Free Ammonia (as N) 1.2 mg/l or less
Irrigation, Industrial Cooling, Controlled Waste disposal	E	pH between 6.0 to 8.5, Electrical Conductivity at 25°C micro mhos/cm Max.2250, Sodium absorption Ratio Max. 26, Boron Max. 2mg/l

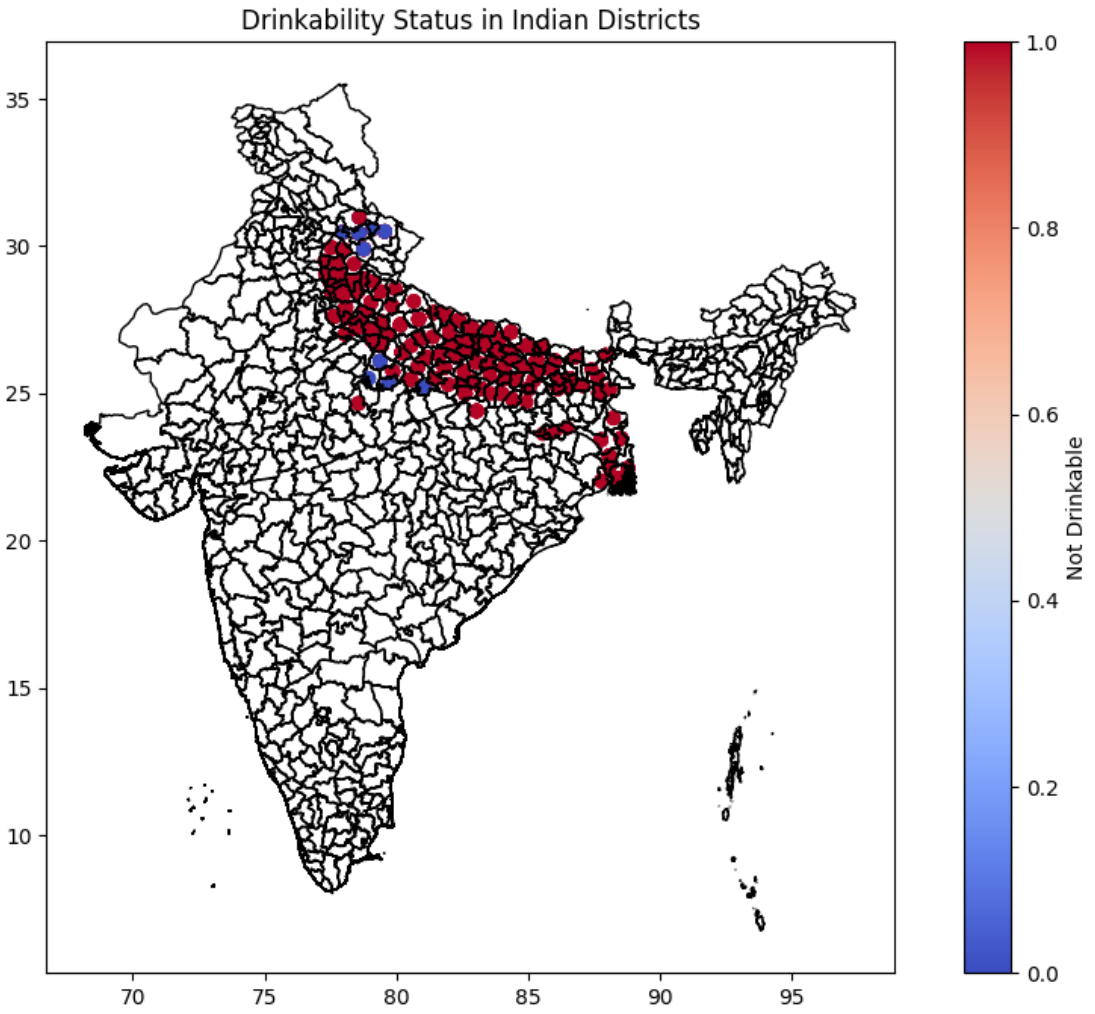


Figure 3.2: Not-drinkable status mapped on Ganga districts as per A class of water.

That completes our preparation of the CPCB and pollution data. Next we try to find some interesting insights from the data while performing exploratory data analysis on it.

3.2.1 Exploratory data analysis of pollution data

This section presents a detailed exploratory data analysis (EDA) of the dataset compiled to study the effects of river pollution on health and nightlight. The dataset encompasses a comprehensive range of variables, including pollution measurements, geographical locations, and administrative data from various monitoring stations. This analysis aims to understand the dataset's structure, coverage, distribution of pollution measures, geographical reach, and the distribution of categorical variables. Additionally, we present visualizations, such as box plots, to examine the distribution of key pollution indicators.

Dataset Overview The dataset consists of 767 entries, each corresponding to a monitoring station's observations, across 40 columns. It includes both numerical and categorical data types, covering a wide range of information:

- **Geographical Information:** Latitude and longitude coordinates, district names, and matched district names, providing a granular view of the dataset's geographical distribution.
- **Monitoring Station Data:** Closest station identifiers and distances, station codes, and names of monitoring locations, indicating the extensive network of pollution monitoring.
- **Pollution Measurements:** Various pollution indicators such as temperature (min and max), dissolved oxygen (min and max), pH levels (min and max), conductivity, BOD, nitrate levels, fecal coliform, and total coliform counts, alongside measurement units (°C, mg/L, µmhos/cm, MPN/100mL).
- **Temporal Data:** Year of observation, allowing for temporal analysis of pollution trends.
- **Health Indicator:** A binary variable indicating whether the water quality is not-drinkable.

Pollution measures The dataset offers a comprehensive view of various pollution measures, including temperature, dissolved oxygen, pH levels, conductivity, BOD, nitrate levels, fecal coliform, and total coliform counts. The measurement units are consistent across observations, allowing for a standardized analysis of water quality indicators. For all our purposes, we use the max values for each district.

- **Temperature:** Ranges from 0°C to 52°C for maximum temperatures, highlighting extreme variations in environmental conditions across different monitoring locations. The average minimum and maximum temperatures are 16.85°C and 28.91°C,

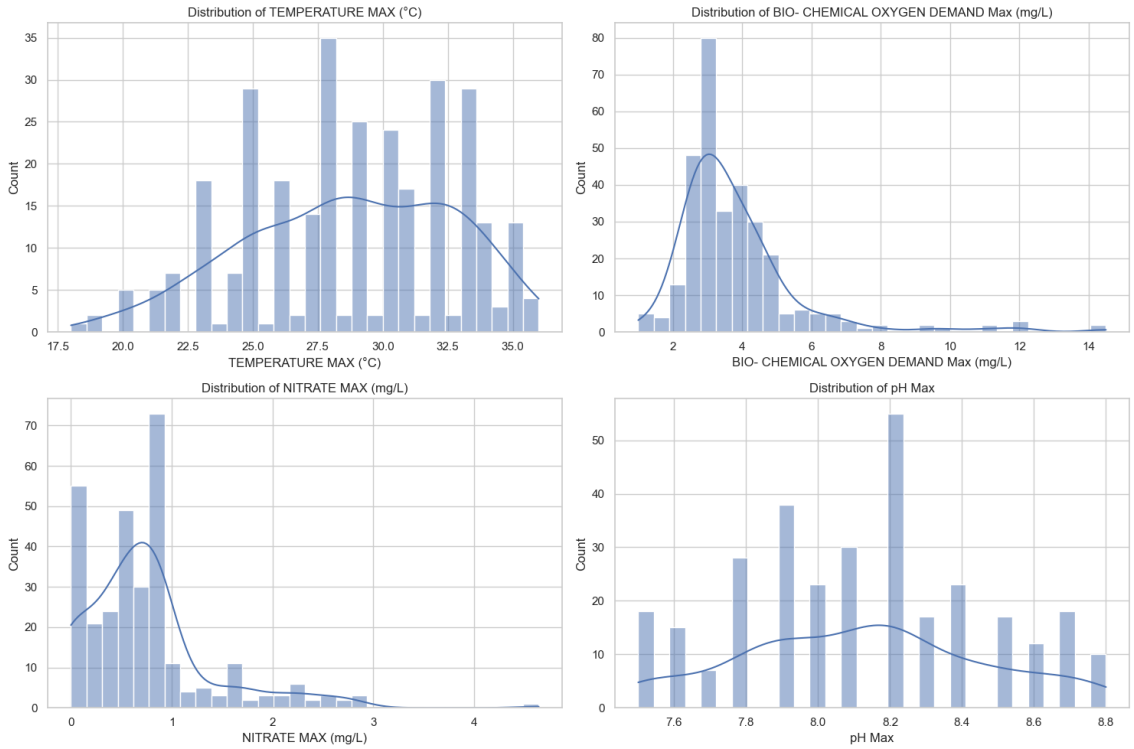


Figure 3.3: Temperature, BOD, Nitrate, and pH distributions for CPCB pollution data.

respectively, indicating a wide range of thermal pollution impacts as seen in Figure 3.3.

- **Dissolved Oxygen:** The minimum levels range from 0 mg/L to 10.1 mg/L, and maximum levels up to 54 mg/L, underscoring the variability in water oxygenation, which is crucial for aquatic life. The average minimum and maximum dissolved oxygen levels are 6.03 mg/L and 9.10 mg/L, respectively.

- **pH Levels:** The dataset records pH values from as low as 1.5 to as high as 811, with the latter likely indicating data recording errors or outliers. The average pH levels range from 7.38 to 9.23 for minimum and maximum values, respectively, pointing towards alkaline conditions in many locations.

- **Conductivity:** Indicates a wide range of 1 µmhos/cm to 32,900 µmhos/cm, with an average minimum and maximum conductivity of 295.94 µmhos/cm and 1,030.79 µmhos/cm, respectively, suggesting varied levels of water electrolyte concentration due to pollution.

- **BOD:** A key indicator of organic pollution, BOD values range from 0 mg/L to 440 mg/L, with average minimum and maximum levels of 4.16 mg/L and 8.55 mg/L, respectively. Higher BOD values in some areas indicate significant organic pollution.

- **Nitrate Levels:** Range from 0 mg/L to 44.4 mg/L, with average minimum and maximum levels of 0.56 mg/L and 1.48 mg/L, respectively, highlighting nutrient pollution issues in certain locations.

- **Fecal Coliform:** The counts range significantly, from 1 MPN/100mL to 35,000,000 MPN/100mL for maximum levels, emphasizing severe contamination in some areas. The average fecal coliform counts further affirm the presence of waterborne pathogens in many monitoring locations.

- **Total Coliform:** Similar to fecal coliform, total coliform counts exhibit wide ranges, up to 54,000,000 MPN/100mL for maximum levels, indicating widespread microbial contamination.

Geographical distribution The dataset's geographical information reveals a wide distribution of monitoring stations across various districts and states, ensuring a diverse representation of river pollution across different ecological and administrative regions. The latitude and longitude coordinates span from 22.026699° to 30.978351° and from 77.276470° to 88.778291°, respectively, covering a substantial geographical area.

- **Ganga Basin:** The Ganga Basin extends across 139 districts and encompasses five states: Uttarakhand, Uttar Pradesh, Bihar, Jharkhand, and West Bengal.

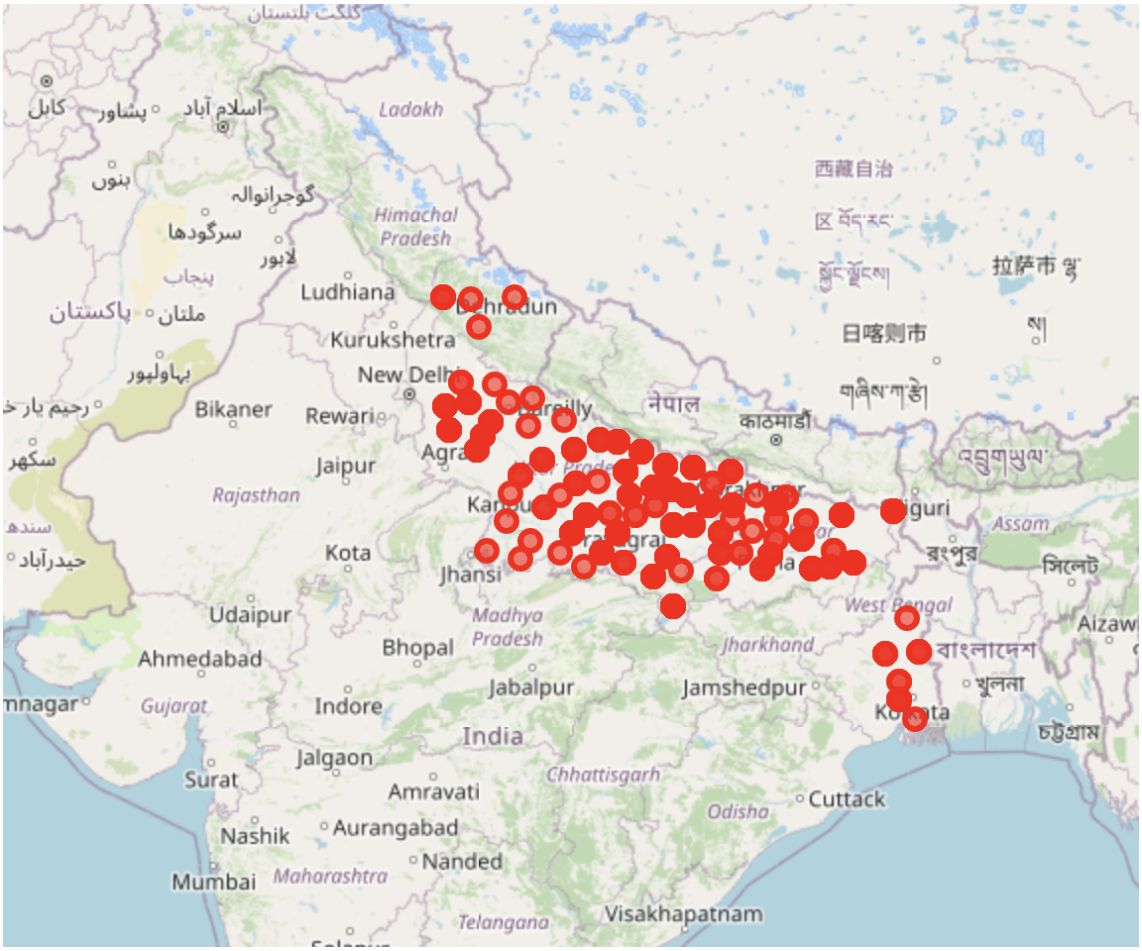


Figure 3.4: River pollution measuring stations on Ganga and its tributaries.

- **Closest stations:** We have a total of 100 unique matched stations for our 122 districts. The distances to the closest monitoring stations provide insights into the network density and the spatial distribution of pollution monitoring efforts. On average, the distance to the first, second, third, fourth, and fifth closest stations are 23.34 km, 30.31 km, 34.51 km, 40.33 km, and 44.65 km, respectively. This shows that all the readings we use to assign pollution measures to a district are within a 50 km radius, and thus represent the ground truth.

- In total, there are 110 water quality monitoring stations, with 39 along the main stream of the Ganga and 71 along its tributaries and sub-tributaries.

- Central Water Commission (CWC) stations examine surface water for sixty-eight 'Standard Hydrology Project Water Quality Parameters'. These parameters span across:
  - Physical, chemical, and biological aspects.
  - Sub-categories such as field determinations, nutrients, organic matter, and others.

- The frequency of monitoring varies, with surface water bodies being checked monthly or quarterly, while groundwater stations are monitored bi-annually.

- Notable monitoring sites include Devprayag, Rishikesh, Rudraprayag, and Varanasi, among others.

3.3 Economic activity data - VIIRS Nightlight

The nightlight data, derived from the VIIRS nighttime data, plays a pivotal role in our analysis. This dataset encompasses remote-sensing nighttime light emissions collected between 2012 and 2021, offering a unique lens through which nocturnal activity can be quantified. The VIIRS data, characterized by its global daily measurements of nocturnal visible and near-infrared light, is crucial for understanding various phenomena that occur under the cover of darkness, including socioeconomic developments and environmental changes.

Source This invaluable dataset is made freely available as open-source data, thanks to the EOG[68] pioneering the collection of nighttime satellite imagery. The history of Nighttime Light map production by EOG dates back to 1994, utilizing the Operational Linescan Sensor (OLS) onboard DMSP satellites. The advent of the Joint Polar-orbiting Satellite System (JPSS) marked a significant advancement in low-light imaging capabilities with the VIIRS Day Night Band (DNB) on board, offering superior quality in global Nighttime Light products. For the context of India, the SHRUG[6] has been leveraged, facilitating the granular level data sharing necessary for our study.

Uses Nightlight data serves as a proxy for a myriad of socioeconomic indicators, including but not limited to GDP growth, urban expansion, public expenditure, and electrification. Its high geographic resolution makes it particularly valuable for studies in low-income countries where alternative development data may be scarce or unreliable. This dataset's versatility has been demonstrated in various research efforts, ranging from examining the distribution of manufacturing employment across rural India to poverty mapping exercises aimed at improving the targeting of development programs.

Scope in this study In our study, we focus specifically on the VIIRS data spanning from 2012 to 2021 to align with the temporal availability of river pollution data from the CPCB. By selecting data at the district level, we maintain consistency with the granularity of the CPCB data, allowing for a nuanced analysis of the interplay between nightlight intensity and river pollution. This choice is further supported by the temporal alignment of the datasets, both presented on a yearly basis, facilitating a comprehensive examination of trends and patterns over time. The utilization of nightlight data, thus, significantly enhances our capacity to explore and understand the socioeconomic dimensions intertwined with environmental quality across India.

Summary Table

A summary table is presented below to encapsulate the key metrics derived from the dataset:

Metric	Mean	Std Dev	Min	Max
Latitude	26.66	1.81	22.03	30.98
Longitude	82.19	3.21	77.28	88.78
1st Closest Station Distance (km)	23.38	19.18	0.85	106.04
VIIRS Annual Mean	1.09	1.04	0.02	10.17
Total Population (2011 Census)	2.78M	1.48M	242,285	7.72M
Total Households (2011 Census)	496,281	297,168	53,542	1.73M

Table 3.2: Summary of Key Metrics from the Dataset

3.3.1 Data preparation

This section of the analysis integrates nightlight intensity data to investigate its association with pollution levels across districts in India. Utilizing the SHRUG dataset, nightlight intensity at the district level from 2012 to 2021 is examined.

Data Preparation and Extraction The initial phase of our data analysis involved processing raw data obtained from the CPCB, which included annual datasets for all river systems in India. The process, documented in the notebook river_pollution_analysis.ipynb, entailed several key steps:

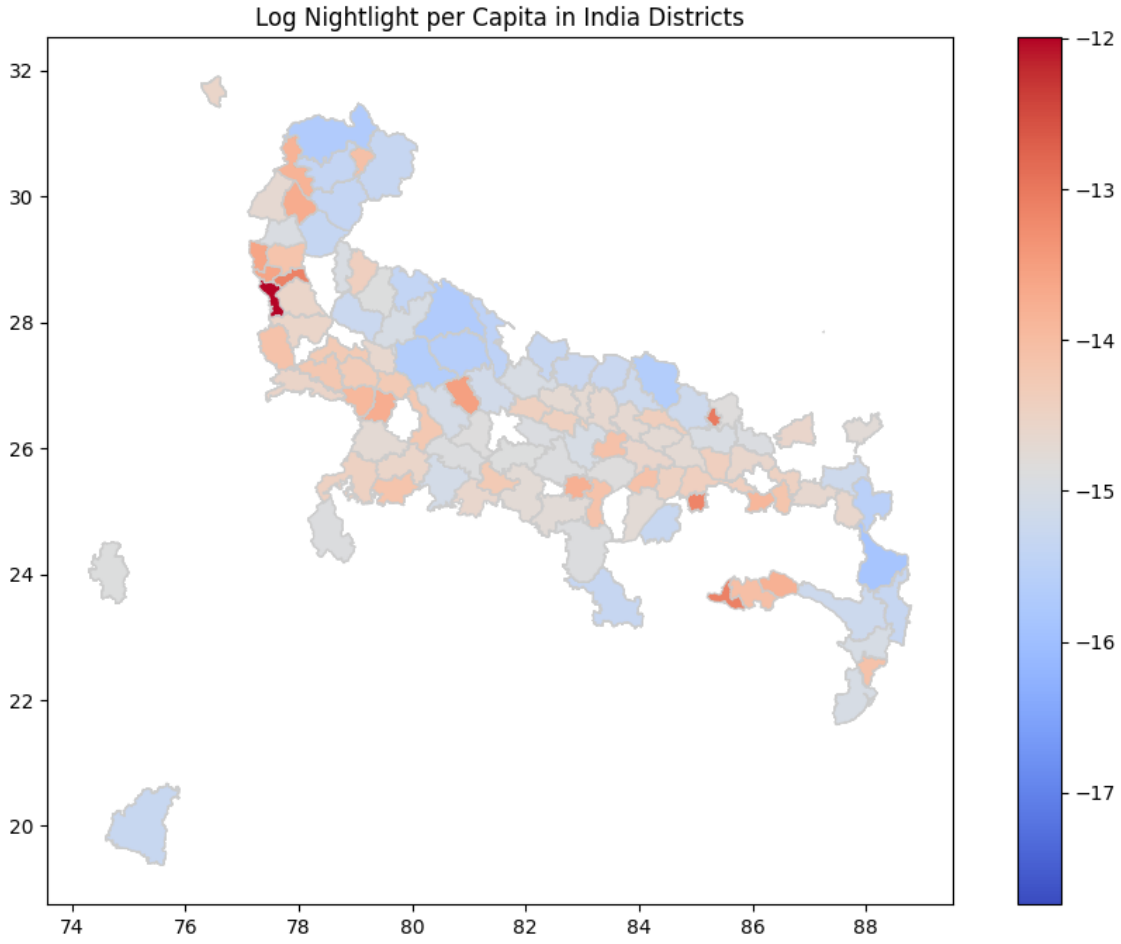


Figure 3.5: Log of nightlight per capita of districts near Ganga. We can see low nightlights at upstream districts, getting higher in plains after Haridwar.

1. **Shurg viirs annual district:** We first load the shrug viirs nightlight data a district level. This data contains the nightlight minimum, maximum, sum, and mean values for 10 years, between 2012-2021 for all districts of India. The data is at a district level and contains 640 unique districts. The total dimensions are (12800, 8) as for each district, the data has 20 data points (10 years of data for both mean and median nightlight values). A consolidation step merges these values by state, district, and year, yielding a singular entry per district annually. We assume that the mean and median values both represent the intensities and not having sufficient evidence, decide to merge them, as the values move in the same directions, both in intensity and direction (6400,8).

2. **Population data and nightlight per capita:** The nightlight data is merged with the 2011 Census population data to compute per capita nightlight intensity, shedding light on nightlight distribution relative to population density. 2011 population data was used, as later census data is not conducted yet. This gives us a control in our further regressions to see that the nightlight in a place is not attributed due to population.

3. **District Identifier Addition:** A unique district identifier (shrid2) is appended to each district, enabling district name retrieval from the SHRUG database. The refined dataset comprises 6930 entries, detailing district and year-level nightlight data. Shrid2 is a granular identifier specific to SHRUG database that can help to find our level of keys and values. Here we use shrid2 to map it to district names. One can use it to map subdistricts, election constituencies, or even granular cells. This gives us a total of 640 districts as before, but only 625 unique district names, with some districts associated with multiple IDs, leading to an excess of entries for certain districts. A detailed review identified 56 instances requiring data correction to maintain accuracy. We handle this by using only the district names in our states and handling the duplicates.

4. **Fuzzy Matching Process:** A critical step involved fuzzy matching, with a 90% threshold, to accurately align district names from the nightlight dataset with the pollution data. This ensured precise correlation of nightlight intensity with pollution levels. District names from SHRUG were mapped with district names taken from CPCB pollution data. These are two different sources altogether, hence fuzzy matching was applied for a certain leeway. Even then, we had to manually map some of the data points. Example: South 24 Parganas is a district in West Bengal state of India. It is named so in SHRUG nightlight dataset, but in CPCB, it is named Dakshin(South in Hindi) 24 Parganas, which was mapped manually.

5. **Creation of Matched Dataset:** The integration introduced a column named `nightlight_district_matched`, combining pollution data with nightlight intensity information into a unified dataset consisting of 829 rows, representing 108 distinct districts.

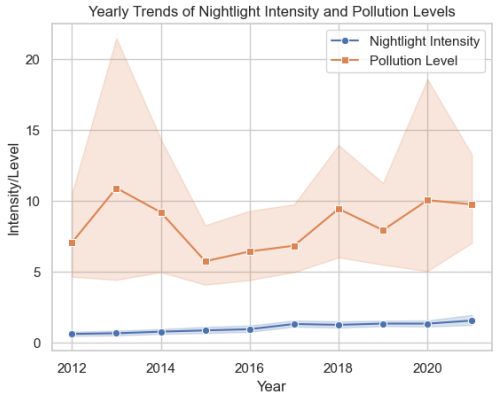
6. **Identifying Exogenous Pollution Variables:** Identifying an exogenous variable related to pollution, potentially influenced by factors like economic development or policy changes, was vital. Such variables were examined for their correlation with variations in nightlight intensity, aiming to discern their impact on the economic conditions of the districts.

3.3.2 Exploratory data analysis

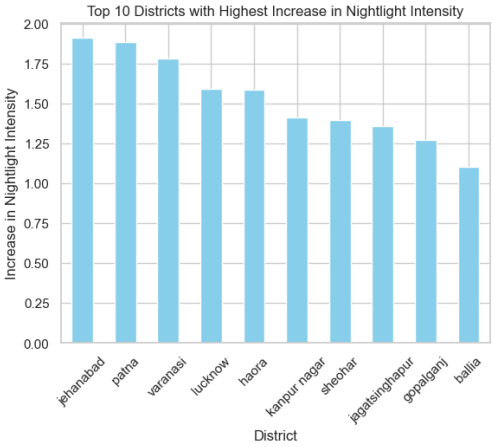
1. **Shrug nightlight data:** We see some of the examples of the nightlight plots in Figure 3.6. The minimum intensity values stretch from slightly below zero (-0.064925) to 19.967077, suggesting nuances in data calibration or processing and delineating zones from absolute darkness to those with minimal yet noticeable light emissions. On the other end of the spectrum, maximum intensity values soar from nil to a striking 4243.375488, accentuating the profound disparities in nightlight across locales, potentially mirroring the contrast between urban and rural areas or marking territories with significant light sources like metropolitan cities or industrial complexes. The dataset's average annual intensity hovers at a modest 1.511998, pointing to generally low to moderate light emission levels, albeit with a standard deviation of 5.200562 indicating a substantial spread in average luminosity among different sites. The sum of intensity figures further elaborates on this narrative, ranging widely to reflect the total annual light output from unlit locations to those radiating extensive cumulative emissions, thus showcasing the aggregated nocturnal activities captured per cell. Furthermore, the observed cell count, varying from 75 to an expansive 451404, provides a glimpse into the spatial reach of the dataset, hinting at a comprehensive coverage that spans a diverse array of geographical extents.

2. **Nightlight per capita along Ganga:** We see in Figure 3.5 that the nightlight at the beginning of the river is quite low, and as it reaches more populous areas, it starts going up. It makes sense as per geography as well, as the upper regions where Ganga is originated are mountainous glaciers which are not inhabitided a lot, and not a lot of industrious activities. But in states like Uttar Pradesh, Bihar, we have an exponential increase, as these are the most populated states in India and we have large tanning, leather factories, and agriculture activities in these states. Hence, the nightlight distribution is indeed showing economic ground truth in the study region.

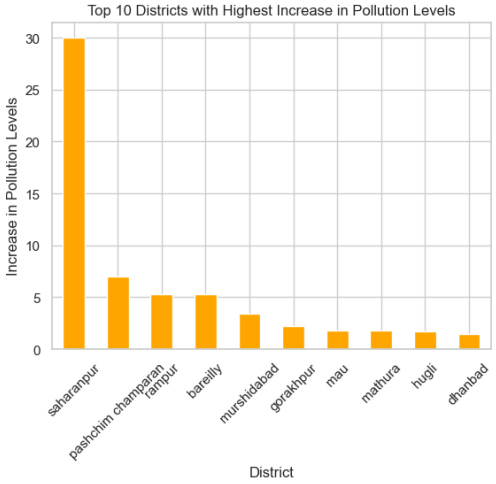
3. **Missing values:** In the combined pollution nightlight data, we had some missing values like the district names were not populated for 66 rows, which constituted for more than 1 row for a single district in multiple years. And hence we did not have the nightlight data for those rows from the SHRUG dataset.



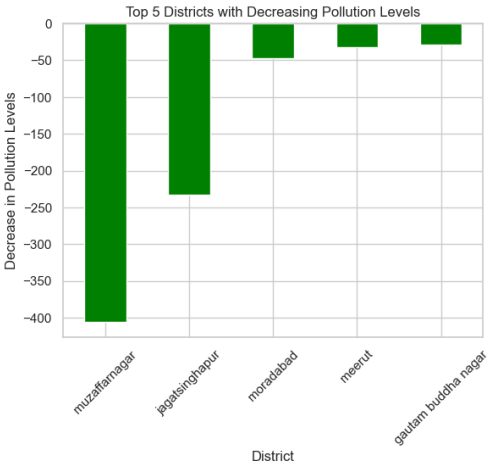
(a) Yearly trend of pollution (BOD) and mean nightlight trend over the years.



(b) Top 10 districts with the highest increase in mean nightlight intensity over the years.



(c) Top 10 districts with the highest increase in pollution levels (BOD) over the years.



(d) Top 5 districts with the highest decrease in pollution levels (BOD) over the years.

Figure 3.6: Comparative analysis of pollution and nightlight trends.

3.4 Health and Family Welfare

This section explores the correlation between water-borne diseases and their impact on health and the economy, utilizing data from the Statistical Year Book India, compiled by the Central Statistics Office (CSO)[7]. This resource, in its 44th edition, aggregates socio-economic indicators at national and state levels, enabling a comprehensive analysis across various sectors.

Source The statistical year book India provides various figures in multiple domains through years 2011-2018. Each year it publishes multiple reports containing data from Population, finance, agriculture, livestock, roads, shipping, education, Health and family welfare etc. We are going to use the dataset Health and Family welfare. The dataset then contains multiple reports like Post-graduate admissions, national family survey and number of cases and deaths due to diseases. We want to use the last report to find a concrete relationship between river pollution and how many people suffer physically and what part of these diseases were caused due to river pollution.

Scope in this study The latest available data in this report Number of cases and deaths due to diseases is available from 2008 till 2015. For each year, we are provided with state level figures on cases and deaths in each year for a variety of diseases like Malaria, Japanese Encephalatis, Acute Diarrhoel diseases (directly correlates to river pollution), Acute respiratory, and Viral hepatitis. The data is presented for 36 state and union territories of India. And is sourced by Central Bureau of Health Intelligence, Ministry of Health & Family Welfare, making it reliable and trusted.

Data Processing and Cleaning

We first download the latest Number of cases and deaths due to diseases report. It contains the data for all previous years. We manipulate the data as follows:

1. We aggregate all the diseases for each state and remove the cases column, to get total deaths due to all diseases for a single state in an year.
2. Then we reshape the dataset so we have for each state, from 2008-2015, total deaths due to diseases.
3. After loading the data in Python, we can use this data to analyse relationships between river pollution in Ganga and deaths.
4. We also load our Ganga pollution data with centroid and geographical information here.
5. We build a dataframe "Ganga_pollution_disease_death" by merging these two datasets using a left join on pollution data. We pull state level deaths to our pollution data using "NAME_1" from pollution and "State" from deaths due to diseases.

6. As the pollution data spans from 2012-2021, and health data from 2008-2015, our combined data has temporal dimensions from 2012-2015.

7. We use this subset, with 276 not-null death observations as our base for manipulation and analysis.

8. We replace some column names for better readability, replace non-numeric strings with Nan. Convert measurement columns to float. State columns to string. Fill deaths where not available to 0.

9. Generate lag variables for Nitrate Max, BOD max, and not-drinkable for 3 years to study delayed effects.

EDA of pollution and deaths data

Deaths due to diseases The dataset consists of 240 entries and three columns. The columns are as follows:

State: The name of the state, which is a categorical variable. Year: The year the data pertains to, which is a numerical variable (integer). Deaths: The number of deaths, which is also a numerical variable (integer).

The data types for Year and Deaths are integers, which is appropriate for numerical analysis. The State column is an object, which typically means it's a string or a mix of different types, suitable for categorical data.

Summary Statistics for the numerical columns are as follows:

- There are a total of 240 records spanning from the year 2008 to 2015.

- The average (mean) number of deaths is approximately 213, with a standard deviation of around 315, indicating a wide variance in the number of deaths.

- The median (50th percentile) number of deaths is 94.5, which is significantly lower than the average, suggesting a right-skewed distribution with some years or states having very high numbers of deaths.

- The maximum number of deaths in a single year for a state is 2,387.

- The bar chart 3.7 shows that West Bengal, Uttar Pradesh, Assam, Orissa, and Andhra Pradesh have the highest total number of deaths over the observed years. On the other end, Jammu & Kashmir, Goa, Nagaland, and Sikkim have the lowest totals.

- The line plot for Average Deaths per Year indicates a general decreasing trend in average deaths from 2008 to 2015, with some fluctuations. The highest average occurred in 2008, and the lowest in 2015.

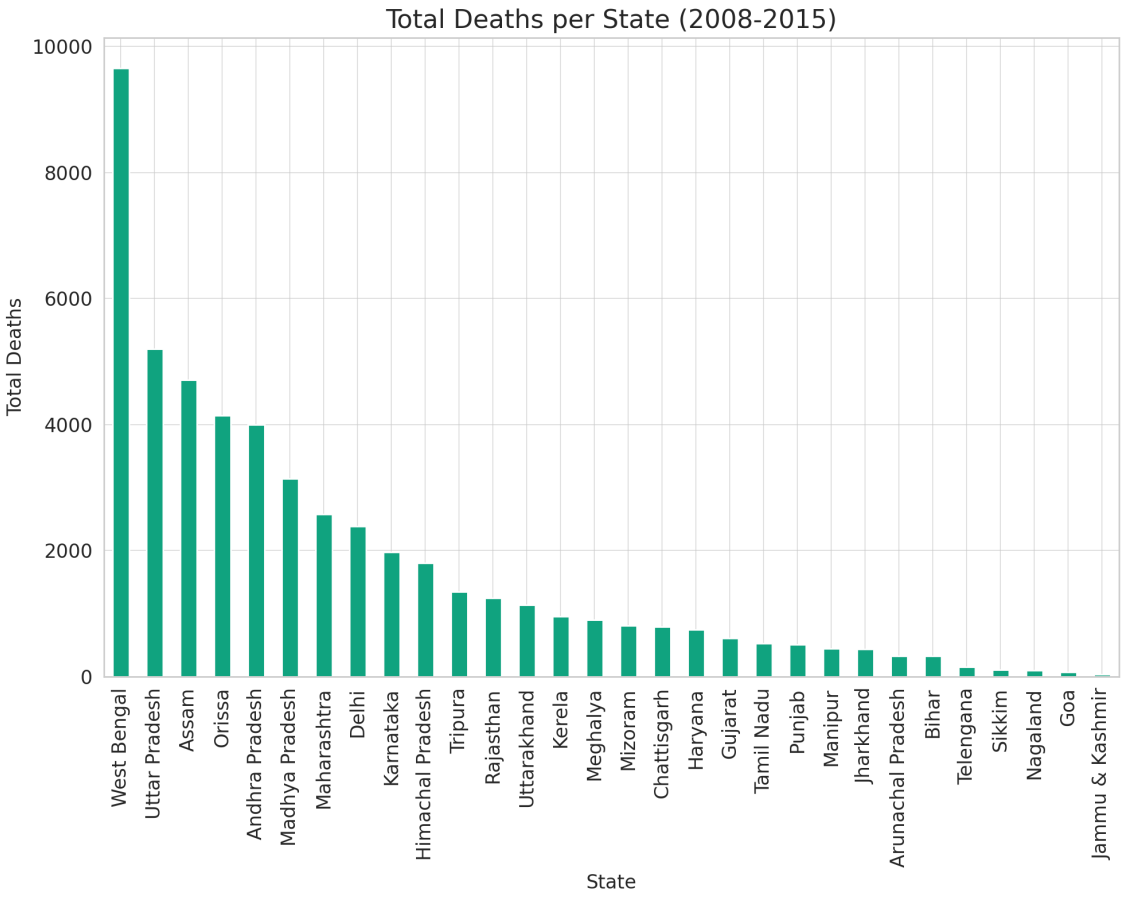


Figure 3.7: Total deaths due to diseases for each state in our dataset(2008-2015).

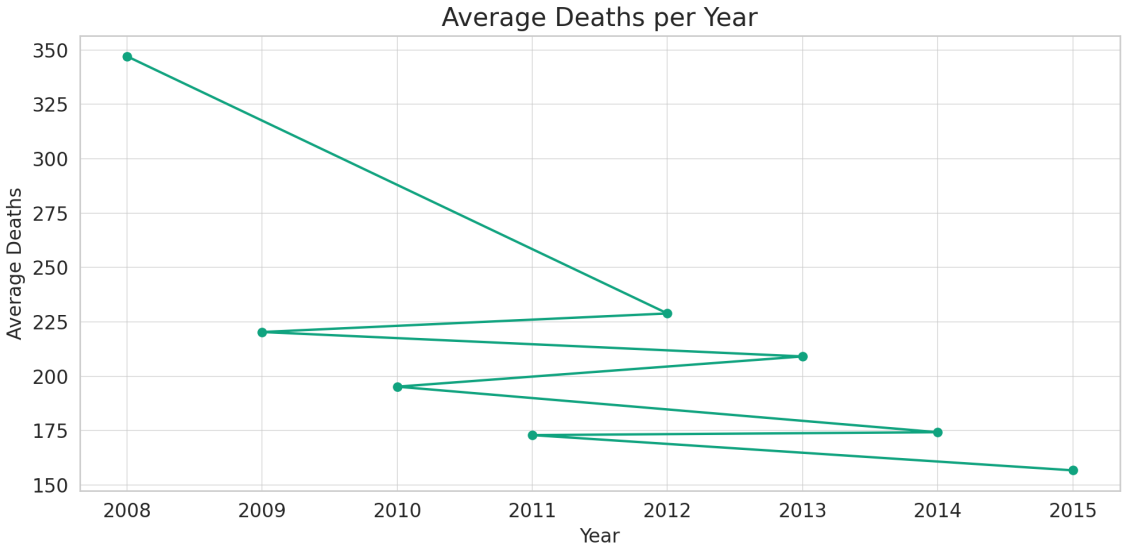


Figure 3.8: Average number of deaths due to diseases trend over the years.

Gaps and Recommendations for Future Research This study, while comprehensive in its approach to examining the relationships between environmental factors and socio-economic indicators, faces several inherent data limitations and methodological challenges. Utilizing yearly aggregated data across river pollution, nightlight intensity as a proxy for economic activity, and health outcomes may obscure finer temporal dynamics and seasonal variations, potentially masking immediate impacts. The accuracy of measurements within these datasets can vary, leading to potential inconsistencies, especially given the varied methodologies in pollution and health data collection. Specifically, employing nightlight intensity as an economic activity proxy assumes a uniform relationship between light emissions and economic output, a correlation that may not hold uniformly across diverse urban and rural contexts. Moreover, the broad categorization of health impacts, without distinguishing between diseases directly attributable to water pollution and other conditions, could dilute the clarity of pollution's health effects. The reliance on the GADM database for geographic information, while valuable, may not reflect the most current administrative or urban developments, affecting spatial analysis precision. Additionally, integrating these disparate data sources—ranging from various formats and geographic details to temporal alignments—necessitates extensive preprocessing to ensure data compatibility and reliability in regression models. Despite these challenges, the study endeavors to contribute valuable insights into the complex interplay of environmental degradation, economic conditions, and public health, highlighting areas for future research and methodological refinement.

4 Empirical Framework

This chapter elucidates the empirical framework underpinning the study of the multifaceted impacts of river pollution on public health outcomes and socio-economic indicators, specifically nightlight intensity. The chosen analytical approach is grounded in robust statistical methodologies designed to unravel the complex interplay between environmental degradation and its consequent effects on human health and economic activities. This chapter delineates the construction of the regression models, detailing the rationale behind the selection of dependent and independent variables, control variables, fixed effects, the implementation of log transformations, the incorporation of lags, and the choice of estimation techniques.

4.1 River pollution's effects on Health

The health implications of pollution in the Ganga River are profound and far-reaching, impacting not only the immediate environment but also the well-being and livelihood of the communities residing along its banks. In this analysis, we endeavor to quantify the extent to which the escalating levels of river pollution contribute to the overall number of deaths attributed to diseases. Our empirical strategy leverages pollution measurement data from the CPCB spanning the years 2012-2021, focusing on key indicators such as pH, temperature, nitrate levels, etc. Among these, BOD, Nitrate, and a derived variable labeled as not-drinkable have been selected to serve as the independent variables in our analysis. These indicators were chosen due to their direct implications on health, as evidenced by existing literature.

To gauge the impact of these pollution measures on health outcomes, we examine data on deaths due to diseases reported in the Statistical Book of India for the years 2008-2015. This dataset provides a consolidated account of deaths in each state on an annual basis, arising from a range of diseases such as Malaria, Acute Diarrheal Infections, among others. The rationale behind selecting death figures as our dependent variable stems from the critical role of health and family welfare in the socio-economic development of a country. Pollution, be it water or air, has undeniable health repercussions, which in turn have secondary effects on the socio-economic fabric of the population. Alternative health outcomes such as child mortality, hospitalizations, and the impact on sectors like fisheries, agriculture, and education were also considered but deemed beyond the scope of this specific analysis.

4.1.1 Regression Formula and Variables

The core of our analysis is predicated on a PanelOLS regression model that aims to elucidate the effects of river pollution on health outcomes. Specifically, we investigate how pollution indicators such as BOD, Nitrate levels, and a binary variable representing not-drinkable water status influence the number of deaths due to diseases. The generic form of our model is presented as follows:

$$\log(\text{Deaths due to diseases}_{it}) = \beta_0 + \beta_1 \text{Pollution Measure}_{it} + \gamma \mathbf{X}_{it} + \mu_i + \tau_t + \epsilon_{it} \quad (4.1)$$

where:

- $\log(\text{Deaths due to diseases}_{it})$ represents the natural logarithm of the number of deaths due to diseases in state i at time t, adjusted by adding 1 to account for zero death counts.
- $\text{Pollution Measure}_{it}$ denotes the primary independent variables of interest, including $\log(\text{BOD}_{it})$, $\log(\text{Nitrate}_{it})$, and $\text{not-drinkable}_{it}$, each lagged appropriately to capture delayed effects. The logarithmic transformation for continuous pollution measures helps linearize relationships and interpret coefficients as elasticities. $\text{Not-drinkable}_{it}$ is a binary indicator reflecting the not-drinkability status of water.
- $\mathbf{X}_{it}$ is a vector of control variables that might influence health outcomes, including demographic, socioeconomic, and environmental factors at district i and time t.
- μ_i represents district fixed effects, capturing unobserved heterogeneity across districts that could influence health outcomes.
- τ_t denotes year fixed effects, accounting for temporal trends and shocks that uniformly affect all districts.
- ϵ_{it} is the error term, capturing unexplained variation in the dependent variable.

Dependent Variable The dependent variable in our study is the logarithmic transformation of the number of deaths due to diseases ($\log(\text{Deaths due to diseases}_{it})$) in a given state i in year t. This transformation is applied to normalize the distribution of death counts, which typically exhibits right-skewness. Furthermore, adding 1 before taking the logarithm ensures that districts with zero reported deaths are included in the analysis without causing mathematical issues. The choice of this dependent variable allows us to assess the health impact of river pollution in terms of mortality due to diseases, providing a direct measure of public health outcomes.

Independent Variables Our primary independent variables are measures of river pollution: BOD, Nitrate levels, and a binary variable indicating not-drinkable water status ($\text{not-drinkable}_{it}$). These variables are chosen based on their established relevance to water quality and potential health implications.

- $\log(\text{BOD}_{it})$ and $\log(\text{Nitrate}_{it})$ capture the logarithmic values of BOD and Nitrate concentrations, respectively. The logarithmic transformation ensures a proportional relationship with the dependent variable, allowing us to interpret the coefficients as percentage changes in the death count for a percentage change in pollution levels.
- $\text{not-drinkable}_{it}$ is a binary variable that reflects whether the water quality in a district at a given time is below the safe drinking threshold as defined by relevant authorities. This variable captures the most severe instances of water pollution, where water is considered unsafe for human consumption.

Each of these pollution measures is lagged till 2 years to account for potential delayed health effects, acknowledging that the impact of pollution on health outcomes may not be immediate.

The decision to include up to two years of lagged pollution measures in our model is rooted in the desire to capture the delayed effects of pollution on health outcomes. Health impacts resulting from pollution exposure, such as respiratory or waterborne diseases, may not manifest immediately. The two-year lag structure allows us to observe potential long-term health outcomes following exposure, enhancing our model's sensitivity to capturing these delayed effects. This lag structure was validated through preliminary analyses that indicated significant relationships between pollution levels in previous years and current health outcomes, reinforcing the importance of considering these temporal dynamics in our study.

Control Variable In our model, temperature serves as the key control variable, represented as $\log(\text{Temperature}_{it})$. The inclusion of temperature as a control variable is pivotal for several reasons:

- **Direct health impacts:** Temperature has a well-documented effect on health outcomes, influencing the spread of temperature-sensitive diseases, heat-related illnesses, and general mortality rates. By controlling for temperature, we aim to isolate the specific impact of water pollution on health outcomes from the broader environmental influences of climatic conditions.
- **Water quality interaction:** Temperature also affects the chemical and biological processes in water bodies, potentially altering the toxicity of pollutants. For instance, higher temperatures can increase the rate of BOD and affect the solubility of gases like oxygen, thereby influencing the aquatic ecosystem's health. Controlling for temperature allows us to account for these interactions and better understand the direct impact of pollution levels on human health.
- **Seasonal variation:** Including temperature as a control variable helps to account for seasonal variations in health outcomes that might coincide with changes in pollution levels. This is crucial for accurately assessing the relationship between pollution and health, independent of seasonal effects that could confound the analysis.

By controlling for temperature, we enhance the precision of our model and ensure that the estimated effects of pollution measures on health outcomes are not biased by the confounding effects of temperature variability. This careful consideration strengthens the validity of our findings, highlighting the specific contribution of water pollution to the observed health outcomes in the districts along the Ganga River.

By carefully selecting and justifying our dependent, independent, and control variables, we aim to construct a robust empirical framework capable of shedding light on the complex dynamics between river pollution and health outcomes.

State Fixed Effects The incorporation of state fixed effects, denoted by μ_i in our model, plays a crucial role in enhancing the robustness of our analysis. These fixed effects are instrumental in controlling for unobservable, time-invariant characteristics specific to each state that might influence the health outcomes of interest. This includes factors such as long-term economic policies, healthcare infrastructure, and cultural practices that could affect mortality rates but are not directly captured by other variables in the model. By including state fixed effects, we effectively control for these constant differences across states, allowing us to isolate the impact of variations in pollution levels and other covariates on health outcomes within states over time. This methodological approach ensures that our estimated coefficients for pollution measures are not confounded by omitted state-level characteristics, providing a more accurate and credible estimate of the effects of river pollution on health.

In our PanelOLS regression model, we adopt a sophisticated approach by employing district-level pollution measures (Biochemical Oxygen Demand, Nitrate levels, and not-drinkable water status) while examining their impact on state-level health outcomes, specifically the number of deaths due to diseases. This methodological choice is driven by the objective to capture the local environmental conditions as accurately as possible, acknowledging that pollution can exhibit significant spatial variability even within a single state.

The challenge of aligning district-level environmental data with state-level health outcomes is addressed by maintaining the granularity of the pollution data while using the state-level aggregated deaths. This setup increases the number of observations and enriches the analysis by leveraging the variability in pollution levels across districts within the same state. However, it necessitates the careful consideration of how to interpret the effects of local pollution on broader health metrics.

To reconcile the use of district-level pollution data in predicting state-level health outcomes, we employ entity fixed effects at the state level. This approach allows us to control for unobserved, time-invariant heterogeneity across states that could influence health outcomes, thereby isolating the effect of local environmental conditions on state-wide mortality rates due to diseases.

Year Fixed Effects Similarly, year fixed effects, represented by τ_t in our model, are essential for accounting for temporal trends and nationwide shocks that might affect all districts simultaneously. These fixed effects capture year-specific influences such

as national health campaigns, climatic anomalies, or economic recessions that could impact health outcomes across the board. By controlling for these universal time effects, we can more accurately assess the temporal dynamics of pollution's impact on health, independent of these overarching temporal factors. The inclusion of year fixed effects ensures that the variations in health outcomes we observe are not simply reflections of broader temporal trends or events but are indeed associated with changes in pollution levels. This addition significantly enhances the temporal validity of our analysis, allowing us to draw more precise conclusions about the impact of environmental degradation on public health over the study period.

Heteroskedasticity and Standard Error Clustering: To mitigate concerns of heteroskedasticity and autocorrelation within state over time, we employ standard error clustering at the state level. This approach acknowledges the potential for uneven variance across state and the correlation of errors within the same state, enhancing the robustness of our estimations.

4.1.2 Fixed effects model with log-log transformations

In our study, we employ a fixed effects model with log-log transformations to meticulously analyze the effects of river pollution on health, specifically focusing on the complex relationship between district-level pollution measures and state-level health outcomes. This approach is integral to our empirical framework, enabling us to capture the nuanced impacts of environmental degradation on public health while accounting for unobserved heterogeneity and ensuring the robustness of our findings. Below, we detail the rationale and advantages of this methodological choice in the context of our study.

Fixed effects model Fixed effects models are paramount in studies like ours, where the goal is to isolate the effect of variables that vary over time within entities (in our case, states), from those that are constant within entities but vary across them. By incorporating state-level entity fixed effects (μ_i) and year fixed effects (τ_t) into our model, we control for time-invariant characteristics of each state that could influence health outcomes, such as geographical factors, healthcare infrastructure, and long-term policy impacts. Similarly, year fixed effects account for temporal trends and external shocks affecting all states uniformly, such as nationwide health initiatives or epidemics. This methodology is particularly beneficial in our context for several reasons:

- Local Environmental Variability: It captures the local environmental conditions affecting public health by using district-level pollution data, providing a more granular and accurate depiction of pollution exposure.
- State-level Health Outcomes: By examining state-level health outcomes, we acknowledge the broader impact of local pollution on statewide public health metrics, facilitating a comprehensive understanding of environmental health dynamics.
- Unobserved Heterogeneity: Fixed effects mitigate the potential bias arising from

omitted variables that are constant over time within states but vary across them, ensuring a more accurate estimation of the pollution-health relationship.

Log-Log transformations The adoption of log-log transformations in our model serves multiple purposes:

- Linearization: It linearizes the exponential relationships between pollution levels and health outcomes, enabling a simpler and more interpretable linear regression analysis.
- Elasticity Estimation: Coefficients in a log-log model can be directly interpreted as elasticities, offering insights into the percentage change in health outcomes resulting from a one percent change in pollution levels.
- Handling Zeroes: By applying the transformation $\log(x+1)$ to both the dependent variable (deaths due to diseases) and independent variables (BOD, Nitrate), we accommodate observations with zero deaths or pollution measures, ensuring their inclusion in the analysis without skewing the results.

4.2 River pollution's effects on Economic Activity through Nightlight Intensity

The interplay between environmental degradation and economic activity forms the crux of our investigation, wherein we endeavor to quantify the extent to which pollution in the Ganga River impacts the economic vibrancy of adjacent communities. Our analytical lens focuses on nightlight intensity per capita as a proxy for economic activity, a novel approach that leverages satellite imagery to infer economic development levels. This section outlines the rationale behind our variable selections and methodological choices, mirroring the structure of our analysis on health outcomes but with a distinct emphasis on economic implications.

4.2.1 Regression Formula and Variables

Adopting the PanelOLS regression model, our analysis seeks to explore the relationship between river pollution measures and nightlight intensity per capita. The adapted model for this analysis is represented as follows:

$$\log(\text{Nightlight per Capita}_{it}) = \beta_0 + \beta_1 \text{Pollution Measure}_{it} + \gamma \mathbf{X}_{it} + \delta_i + \lambda_t + \epsilon_{it} \quad (4.2)$$

where:

- $\log(\text{Nightlight per Capita}_{it})$ signifies the natural logarithm of the nightlight intensity per capita in district i at time t, calculated as the mean nightlight intensity divided by the population (from the 2011 census) to adjust for population density.

- Pollution Measure$_{it}$ our primary variables of interest, log(BOD$_{it}$), log(Nitrate$_{it}$), and not_drinkable$_{it}$, reflecting the pollution levels within each district, lagged to capture delayed effects.
- $\mathbf{X}_{it}$ encompasses control variables, with temperature being a pivotal factor for its direct and indirect effects on economic activities.
- δ_i and λ_t denote district fixed effects and year fixed effects, respectively, controlling for unobserved heterogeneity across districts and temporal trends or shocks that might uniformly affect the districts.
- ϵ_{it} is the error term.

Dependent Variable The decision to employ nightlight intensity per capita as the dependent variable stems from its effectiveness in capturing economic activity, especially in regions where conventional economic data may be sparse or unreliable. By dividing the mean nightlight intensity by the population figures from the 2011 census, we aim to control for population size, thus providing a more accurate reflection of economic density rather than absolute levels of light emission. While more current population data would have been ideal, the 2011 census figures offer a viable approximation for our analysis.

Independent Variables Our primary independent variables remain consistent with the health outcomes analysis, focusing on BOD, Nitrate, and not-drinkable to capture the multifaceted nature of water pollution. These variables are crucial for assessing the environmental conditions that could influence economic activities in the districts along the Ganga River.

For the economic activity analysis, the inclusion of three years of lagged pollution measures acknowledges the potential delayed impact of environmental conditions on economic performance. Economic responses to pollution, including shifts in agricultural productivity, tourism, and health-related work absences, might unfold over several years. The choice of a three-year lag structure is supported by empirical evidence suggesting that the full economic implications of pollution exposure may not be immediate but accumulate over time. This approach ensures that our analysis captures not just the short-term but also the more prolonged economic impacts of pollution, providing a comprehensive view of its effects on nightlight intensity as a proxy for economic activity.

Control Variable Temperature, as a control variable, plays a significant role in this analysis too. It not only impacts economic activities directly, such as agricultural productivity and energy consumption but also influences the ecological balance of the river, thereby affecting the economic reliance on riverine resources.

District Fixed Effects Employing district-level fixed effects (δ_i) allows us to control for intrinsic characteristics of each district that might influence economic activities, such as geographical advantages, infrastructure, and local governance quality. This adjustment

ensures that the observed effects of pollution on economic activity are not confounded by these unobserved district-specific factors.

Year Fixed Effects Year fixed effects (λ_t) capture the nationwide or global economic trends, policy changes, or environmental regulations affecting all districts similarly over time. By including these fixed effects, our model can isolate the impact of pollution on economic activities from broader temporal influences.

Heteroskedasticity and Standard Error Clustering: To mitigate concerns of heteroskedasticity and autocorrelation within districts over time, we employ standard error clustering at the district level. This approach acknowledges the potential for uneven variance across districts and the correlation of errors within the same district, enhancing the robustness of our estimations.

Endogenous Variables: While our model focuses on the direct impacts of pollution on nightlight intensity, we remain cognizant of potential endogeneity issues, such as reverse causality or omitted variable bias. To address these concerns, our selection of lagged pollution measures aims to mitigate temporal endogeneity, ensuring that current economic activity is not influencing past pollution levels.

Our analysis acknowledges the potential for endogeneity, particularly the concern that health outcomes might simultaneously influence pollution levels, creating a reverse causality scenario. Additionally, omitted variable bias, where unobserved factors simultaneously affect pollution levels and health outcomes, could distort our model's estimates. To address these concerns, we utilized lagged pollution measures, presuming that current health outcomes are unlikely to influence past pollution levels, thereby mitigating temporal endogeneity. This strategy, while not eliminating all potential endogeneity, reduces the likelihood of reverse causality affecting our findings. Future work could explore instrumental variable (IV) approaches to more rigorously address endogeneity issues.

4.2.2 Fixed effects model with log-log transformations

The utilization of a fixed effects model with log-log transformations is particularly apt for dissecting the nuanced impacts of river pollution on economic activity, as measured by nightlight intensity per capita. This methodological choice facilitates an analysis that is sensitive to the proportionate changes in economic activity in response to variations in pollution levels, allowing for a refined understanding of the environmental-economic nexus.

Rationale for PanelOLS Our choice of PanelOLS, reflective of the approach taken in the health outcomes analysis, is predicated on its appropriateness for panel data, encompassing observations across districts over time. PanelOLS, with its capacity to incorporate fixed effects, aligns perfectly with our dataset's structure and the nuanced nature of our research question. It ensures that our findings are robust, controlling for

both spatial and temporal heterogeneity, and provides a clear, interpretable framework to elucidate the relationships under investigation.

Limitations In analyzing the relationship between river pollution and economic activity, potential biases such as the misrepresentation of economic activity through nightlight intensity, and limitations related to the use of outdated population data, must be acknowledged. Nightlight intensity, while a novel proxy for economic activity, may not capture the full spectrum of economic changes, especially in areas with low electrification or where economic activities occur during daylight hours. The reliance on 2011 census data for population estimates further limits the temporal accuracy of our per capita calculations. Despite these challenges, our methodological choices, including the consideration of temperature as a control variable and the use of district-level fixed effects, are designed to enhance the robustness of our findings within these constraints.

Endogeneity poses a significant challenge in analyzing the impact of pollution on economic activity, with potential reverse causality where economic conditions could influence environmental policies and pollution levels. Omitted variable bias is another concern, as unmeasured factors might simultaneously drive both economic development and pollution. Our approach, incorporating lagged pollution measures, is designed to reduce the risk of reverse causality, assuming that past pollution levels are exogenous to current economic conditions. While this strategy helps mitigate some endogeneity concerns, it does not eliminate them entirely. Instrumental variable (IV) techniques or difference-in-differences (DID) approaches could provide alternative methods to more robustly address endogeneity in future analyses.

5 Results and Discussions

This chapter elucidates the impacts of river pollution on health and socio-economic outcomes for populations residing near the river Ganga. Building upon the empirical framework established in the preceding chapter—which detailed the selection of regression variables and the methodology for discerning the direct impacts between dependent and independent variables—this chapter presents the findings from our PanelOLS analyses.

The discourse is bifurcated into two primary sections. Section 5.1 delves into the health ramifications of various pollution measures, specifically focusing on mortality rates attributable to diseases. Conversely, Section 5.2 contemplates the repercussions of these pollution measures—namely Nitrate, BOD, and the designation of water as not-drinkable—on nightlight per capita, serving as a proxy for economic activity.

In exploring the complex interplay between river pollution and its effects on communities within the Ganga basin, this study is guided by the following hypotheses, each designed to shed light on different dimensions of pollution's impact:

- **Pollution vs Health:** It is hypothesized that an increase in river pollution is positively correlated with an increase in deaths due to diseases, suggesting a direct adverse impact on public health.

- **Pollution vs Economic Activity (Negative Correlation):** This hypothesis posits that an increase in river pollution negatively affects economic activity, potentially due to deteriorating environmental conditions making the area less appealing for habitation and business operations.

- **Pollution vs Economic Activity (Positive Correlation):** Conversely, we also explore the possibility that increased river pollution could inadvertently correlate with heightened economic activity. This scenario could emerge if regions with lax environmental controls attract more industrial activities due to lower compliance costs.

The inclusion of both negative and positive correlations in the context of pollution's impact on economic activity acknowledges the multifaceted nature of this relationship, reflecting the varying dynamics across different sectors and regions. The forthcoming sections detail the outcomes of our analyses concerning these hypotheses, offering insights into the broader implications of river pollution on the Ganga basin's socio-economic fabric and public health.

5.1 Regression results River pollution vs Deaths due to diseases

This analysis delves into the stark realities of river pollution's impact on human health, as manifested through increased mortality rates from diseases. By examining the effects of water quality indicators such as BOD, Nitrate levels, and the derived variable not-drinkable, we endeavor to uncover the profound implications of environmental degradation on public health. This section exclusively addresses the correlation between pollution measures and death rates, elucidating the environmental determinants of health outcomes.

In ensuring the integrity of this analysis, both state and year fixed effects are meticulously incorporated, serving to mitigate the influence of unobserved heterogeneity and capture temporal variations. This methodological rigor, complemented by robust standard errors clustered by state, underpins the reliability of the findings, safeguarding against potential data distortions.

5.1.1 Biological Oxygen Demand and Mortality Rates

Table 5.1: Impact of BOD on Death Rates

	Dependent Variable: Log Deaths			
	(1)	(2)	(3)	(4)
Log BOD $_{t-1}$	0.0876	0.1009		
	(0.0618)	(0.0712)		
Log Temperature $_{t-1}$		0.0052		
		(0.3079)		
Log BOD $_{t-2}$			0.0770*	0.1057*
			(0.0437)	(0.0606)
Log Temperature $_{t-2}$				-0.0425
				(0.2360)
State Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	740	696	735	691
R-squared	0.0148	0.0169	0.0116	0.0188

Robust standard errors are clustered at the state level and are in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The first set of regressions explores the relationship between BOD and death rates due to diseases. Table 5.1 presents these findings, revealing significant lagged effects of BOD levels on mortality rates. This regression table is set up to analyze the impact of BOD and temperature on mortality rates, with the dependent variable being the logarithm of deaths. Each column represents a different model variation, with coefficients for the lagged variables of maximum BOD and temperature. Here's how to interpret the key elements of this table:

- Log BOD $t-1$: In models (1) and (2), a one-unit increase in the log of maximum BOD from the previous year ($t-1$) is associated with a 0.0876 and 0.1009 increase in the log of death rates, respectively. The standard errors, given in parentheses (0.0618 and 0.0712), indicate the precision of these estimates. The absence of a statistical significance indicator (e.g., *, **, ***) suggests that these increases are not statistically significant at the conventional levels, meaning we are less confident that the true effect is not zero.

- Log BOD $t-2$: In models (3) and (4), a one-unit increase in the log of maximum BOD from two years prior ($t-2$) is associated with a 0.0770 and 0.1057 increase in the log of death rates, with * indicating statistical significance at the 10% level. This suggests there's a slightly higher confidence in these effects.

- Log Temperature: The coefficients for temperature are mixed, with some positive and some negative, indicating variable effects on death rates. The large standard errors (e.g., 0.3079 for $t-1$ temperature in model 2) suggest considerable uncertainty around these estimates.

The analysis attempts to determine how previous years' water quality, as measured by BOD, impacts mortality rates, with adjustments for temperature and controlled for state-specific and time-specific factors. While we observe a trend where increases in BOD are associated with increases in death rates (especially for BOD levels two years prior), the evidence is not strong enough to conclusively say these increases are significant across all models, partly due to the high variability indicated by the standard errors. Additionally, the low R-squared values suggest that many other factors not captured by these models likely play a significant role in determining mortality rates. Essentially, while there might be a relationship between higher BOD levels and increased deaths, these models alone do not provide a definitive or strong explanation of mortality rates due to diseases.

BOD is a critical indicator of organic pollution in water bodies, directly correlating with human health impacts. Research, including findings by Wen et al., 2017[50] and observations from the Ganga basin, anticipates a significant rise in populations exposed to hazardous BOD levels by 2050. This pollution is not only a threat to aquatic ecosystems but also to human health, evidenced by the linkage between water quality and diseases such as gastroenteritis[69] and conditions induced by cyanobacterial toxins[70]. Further compounding these concerns, Li Lin et al.[71] highlight the severe global health burden posed by waterborne diseases, with millions annually succumbing to diarrheal illnesses linked to poor water quality. Elevated BOD levels are associated with the discharge of untreated urban sewage containing pathogens that cause diseases like diarrhoea, which is a leading cause of illness and death globally[50]. The intensity of industrial organic water pollution has been found to be positively correlated with infant mortality and child mortality in less developed regions, highlighting the serious consequences of water pollution on human health and disease heterogeneity[71].

The relationship between high BOD levels and adverse health effects underscores the urgent need for comprehensive water management and treatment solutions. The United Nations[72] outlines how elevated BOD levels can signify fecal contamination and organic carbon increases from various sources, complicating water use and necessitating costly treatments. Moreover, the production of safe drinking water becomes more challenging with high organic carbon concentrations, especially when chlorination leads to toxic by-products like trihalomethanes.

Table 5.2: Impact of Nitrate on Death Rates

	Dependent Variable: Log Deaths			
	(1)	(2)	(3)	(4)
Log Nitrate $_{t-1}$	0.1345	0.1770		
	(0.1306)	(0.1565)		
Log Temperature $_{t-1}$		0.0742		
		(0.3169)		
Log Nitrate $_{t-2}$			0.1493	0.1955
			(0.1159)	(0.1387)
Log Temperature $_{t-2}$				0.0788
				(0.2294)
State Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	585	550	581	546
R-squared	0.0215	0.0356	0.0277	0.0454

Robust standard errors are clustered at the state level and are in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5.1.2 Nitrate Levels and Mortality Rates

The section on the impact of Nitrate levels on mortality rates delves into how variations in Nitrate pollution affect death rates due to diseases. Here's how to interpret the key findings and statistical data from the provided table 5.2.

- Lag $t-1$ Nitrate Levels: The regression models (1) and (2) assess the impact of Nitrate levels from the previous year on mortality rates. The coefficients (0.1345 and 0.1770) suggest that a one-unit increase in the log of Nitrate levels from the previous year is associated with an increase in the log of death rates by these amounts, respectively. However, the relatively large standard errors (0.1306 and 0.1565) indicate a degree of uncertainty around these estimates, making it challenging to draw firm conclusions about the significance of these effects.

- Lag $t-2$ Nitrate Levels: Models (3) and (4) show a significant positive association between Nitrate levels from two years prior and mortality rates, with coefficients of 0.1493 and 0.1955, respectively. The smaller standard errors (0.1159 and 0.1387) compared to the $t-1$, especially for model (4) with a p-value of 0.0423, suggest

a more statistically reliable relationship at this lag, indicating that increases in Nitrate pollution can have a delayed but significant effect on increasing death rates.

The analysis investigates how Nitrate pollution in water affects death rates over time, revealing that higher Nitrate levels, particularly two years prior, are significantly associated with increased mortality rates due to diseases. The immediate past year's Nitrate levels also show a potential for increasing death rates, but with less statistical certainty. The variable effects of temperature and the significant positive association found at the $t-2$ lag underscore the delayed health impacts of Nitrate pollution, pointing to the need for long-term environmental and health policy strategies that address and mitigate Nitrate pollution sources. The presence of State and Year Fixed Effects strengthens the analysis by controlling for potential confounders, but the relatively low R-squared values indicate that additional factors not captured by these models are also influencing mortality rates.

The levels of nitrate in drinking water have been linked to various health concerns, including colorectal cancer, thyroid disease, and neural tube birth defects. Studies have shown that elevated nitrate levels in water can lead to adverse health outcomes such as methemoglobinemia, especially in infants, and potential associations with other health effects like nausea, headaches, and abdominal cramps[73][74].

The immediate effects of nitrate pollution in rivers on human health include the risk of methemoglobinemia, especially in infants[75], which can lead to serious health issues like skin discoloration and, if left untreated, potentially fatal outcomes[74]. Regarding the lagged effects of nitrate pollution in rivers on human health, it's important to note that there are delays in the movement of nitrates into groundwater, impacting the effectiveness of interventions aimed at mitigating nitrate pollution[76]. These lag times can affect the success of initiatives to improve water quality and may lead to challenges in achieving policy objectives within specific timeframes. Understanding and addressing these lagged effects are crucial for developing strategies to protect water resources and public health from the long-term consequences of nitrate pollution.

5.1.3 Not-drinkable and Mortality Rates

To interpret the table regarding the impact of water being classified as not-drinkable on death rates, it's important to consider several key components of the regression output, including coefficients, standard errors, fixed effects, and R-squared values. Here's a

detailed breakdown of the results:

Table 5.3: Impact of Not-drinkable water on Death Rates

	Dependent Variable: Log Deaths			
	(1)	(2)	(3)	(4)
Not-drinkable $_{t-1}$	0.0757	0.1301		
	(0.0546)	(0.1629)		
Temperature $_{t-1}$		0.0046		
		(0.0159)		
Not-drinkable $_{t-2}$			-0.0792	-0.0949
			(0.0720)	(0.1251)
Temperature $_{t-2}$				-0.0016
				(0.0143)
State Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	758	700	753	695
R-squared	0.0005	0.0037	0.0006	0.0011

Robust standard errors are clustered at the state level and are in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

- Lag One (t-1) Impact: The coefficient for 'not-drinkable' at lag one (t-1) suggests that an increase from 0 to 1 in the not-drinkable variable is associated with a 0.0757 to 0.1301 increase in the log of death rates across models (1) and (2), respectively. The standard errors (0.0546 for model 1 and 0.1629 for model 2) reflect the precision of these estimates. Despite the positive coefficients, which imply that access to not-drinkable water could increase mortality rates, the results are not statistically significant (p-value > 0.05), indicating uncertainty about the strength of this relationship.

- Lag Two (t-2) Impact: For the lag two (t-2) models, the coefficients are negative (-0.0792 and -0.0949), suggesting a counter-intuitive relationship where an increase in the not-drinkable variable could be associated with a decrease in the log of death rates. However, like the lag one results, these coefficients are not statistically significant, and the standard errors (0.0720 and 0.1251) highlight the variability in these estimates.

The analysis attempted to explore how access to water classified as 'not-drinkable' affects mortality rates. Although the models suggest a potential increase in death rates associated with not-drinkable water at one year lag, and a decrease at two years lag, none of these findings are statistically significant, underscoring considerable uncertainty in these relationships. The minimal impact of temperature and the very low R-squared values further highlight that factors beyond the scope of this analysis likely play a significant role in determining mortality rates. Overall, while the direction of the coefficients

might suggest some relationship between water quality and health outcomes, the lack of statistical significance and the low explanatory power of the models call for cautious interpretation and underline the complexity of environmental health impacts.

Dwivedi et al., 2018[77] showed that in Ganga, the level of pesticides has gone way down in the last decade after government interventions, however, the levels of organochlorines a,d ingorganic pollutants is still many times the admissable amounts. The organochlorines are majorly from textile and agriculture, which is in plethora along the Ganga basin. They also check the effects on human health consuming water which surpasses permissable levels, and find high carciogenic risk from metals and reisdue of DDT and HCHs. Survey from The Energy and Resources Institute(TERI)[78] says that 80% of Varanasi respondents feel that pollution of Ganga has an impact on their health. Moreover, 96% agree that the water is not consumable without treatment, similar to our not-drinkable variable. Also, during the summer and monsoon, hospital wards teem with children who need treatment for waterborne diseases due to swimming in the polluted river[79].

In synthesizing the results from the above tables , it is evident that river pollution, through varied indicators such as BOD, Nitrate levels, and the presence of not-drinkable water, directly impacts mortality rates due to diseases. These insights not only underscore the pressing environmental challenges faced but also highlight the imperative for concerted efforts in pollution control and water quality management. The significant associations observed across different pollutants and lags underscore the complex interplay between environmental factors and health outcomes, reinforcing the need for a holistic approach to environmental health.

5.2 Regression Results: River Pollution vs. Nightlight Per Capita

This section measures the influence of river pollution on nightlight intensity per capita, encapsulating the economic activities within the vicinity of the Ganges River. By leveraging pollution metrics such as BOD, Nitrate levels, and the prevalence of not-drinkable water, we aim to uncover the interplay between environmental degradation and economic vibrancy. The analysis exclusively focuses on the correlation between pollution measures and nightlight per capita (nightlight intensity/total population of the district), offering a window into how environmental factors might echo through economic manifestations.

In conducting this analysis, district and year fixed effects were methodically included to neutralize unobserved heterogeneity and temporal shifts, respectively. This methodological choice aims to refine the accuracy of the pollution impact estimation, addressing both spatial and chronological variances. Robust standard errors, clustered by district, strengthens the statistical inference process, ensuring the resilience of the findings against potential data perturbations.

Table 5.4: BOD impact on Log Nightlight per Capita

	(1)	(2)	(3)	(4)
	Dependent Variable: Log Nightlight per Capita			
Log BOD $_{t-1}$	-0.0302*			-0.0150
	(0.0162)			(0.0178)
Log Temperature $_{t-1}$	0.2080**			0.1265
	(0.1233)			(0.1055)
Log BOD $_{t-2}$		-0.0267*		0.0063
		(0.0154)		(0.0169)
Log Temperature $_{t-2}$		0.2488**		0.1972**
		(0.1261)		(0.0974)
Log BOD $_{t-3}$			-0.0500**	-0.0309
			(0.0259)	(0.0290)
Log Temperature $_{t-3}$			0.3888***	0.3682***
			(0.1235)	(0.1128)
District Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	705	703	701	654
R^2	0.0084	0.0132	0.0397	0.0461

Robust standard errors are clustered at district level and are in parentheses.

Significance levels: * $p < 0.1$, ** $p < 0.05$, ***$p < 0.01$

5.2.1 BOD and Nightlight per Capita

Analyzing the relationship between BOD and nightlight intensity per capita sheds light on how environmental pollution influences economic activity, as measured by nightlight intensity—a proxy often used in economic studies to estimate economic growth and activity levels. The regression results, as outlined in Table 5.4, provide a quantitative assessment of this relationship, emphasizing the effect of pollution levels at different lags on economic vibrancy. Here's a detailed interpretation of the key findings and how to understand them:

- Log BOD (t-1): The coefficient for BOD levels from the previous year (t-1) shows a negative impact on nightlight per capita, with a one-unit increase in the log of BOD leading to a -0.0302 decrease in the log of nightlight per capita in model (1), and a -0.0150 decrease in model (4). The presence of a * next to the coefficient in model (1) indicates statistical significance at the 10% level, suggesting a modest confidence in the negative impact of pollution on economic activities.
- Log Temperature: The coefficients for temperature indicate its positive effect on nightlight per capita in the same and subsequent years, with significant increases observed (0.2080** in model (1) and 0.2488** in model (2) for t-1 and t-2, respectively). The statistical significance (** for $p<0.05$, *** for $p<0.01$) of these effects underscores the robustness of temperature's impact on nightlight per capita.
- Lagged Effects: The analysis also reveals significant lagged effects for BOD, particularly for the t-2 and t-3 lags, highlighting the delayed impact of pollution on economic activity. This is especially evident with a -0.0500** decrease in nightlight per capita for a one-unit increase in BOD at lag three (t-3), suggesting that the adverse effects of pollution on economic activity can persist over time.

This analysis illustrates the intricate ways in which pollution, specifically BOD levels, can affect economic activity within a region. A key takeaway is the negative correlation between BOD levels and nightlight per capita, indicating that higher pollution levels are likely to dampen economic vibrancy. This is evidenced by the decrease in nightlight intensity with increases in BOD, particularly in the previous years. The significant lagged effects highlight that the impact of pollution is not immediate but unfolds over time, affecting economic activity with a delay. Additionally, the analysis accounts for temperature variations, which show a positive relationship with nightlight intensity, further enriching our understanding of the environmental determinants of economic outcomes.

The inclusion of state and year fixed effects ensures that the analysis controls for unobserved factors that could bias the results, enhancing the credibility of the findings. However, the relatively low $R\hat{2}$ values across models suggest that while BOD and temperature are important, they are not the sole determinants of economic activity, as captured by nightlight. This points to a complex interplay of factors influencing economic conditions, underscoring the need for a holistic approach to understanding and addressing the economic implications of environmental pollution.

This study underscores the complex relationship between environmental pollution, as evidenced by BOD levels, and economic activity, proxied through nightlight intensity. Consistent with existing literature, our findings suggest a nuanced interaction between economic development and environmental quality.

Joshi's work[80] through GMM estimations on BOD highlights a pivotal concern: economic development can exacerbate water pollution, reversing potential gains in environmental quality. This assertion is particularly relevant in contexts where industrial expansion contributes significantly to BOD levels, challenging the stabilization of aquatic ecosystems. Similarly, Chapagain et al., 2022[81] provide an insightful analysis within the Balinese economy, identifying economic sectors as primary contributors to BOD through both direct and indirect emissions.

Further, the notion that economic prosperity might improve water quality—reflected through increased dissolved oxygen levels in studies such as Pandit et al., 2016[82] introduces a compelling argument for the potential of economic interventions in mitigating pollution. This perspective is echoed in discussions around the Environmental Kuznets Curve for BOD, suggesting an initial increase in pollution with rising income levels, followed by a decline as economic conditions improve.

However, the World Bank[9] presents a stark reminder of the economic costs associated with high BOD levels, illustrating significant GDP losses in both moderately and heavily polluted areas. This correlation underscores the broader economic implications of environmental degradation, reinforcing the importance of integrating pollution control measures within economic planning and development strategies.

Our analysis aligns with these studies, offering further evidence of the adverse impacts of BOD on economic vibrancy. By exploring the lagged effects of pollution, we contribute to a deeper understanding of its prolonged economic repercussions. Acknowledging the complexity of these dynamics is essential for devising effective pollution control policies and fostering sustainable economic growth.

5.2.2 Nitrate and Nightlight per Capita

The analysis focusing on the impact of Nitrate levels on nightlight intensity, as shown in Table 5.5, reveals insightful patterns about how environmental pollutants influence economic conditions, as inferred from nightlight data. Similar to the analysis involving BOD, this examination of Nitrate's influence introduces a nuanced view of pollution's economic implications.

- Log Nitrate (t-1): The positive coefficients for Nitrate levels from the previous year (t-1) in models (1) and (4) indicate a potential increase in nightlight intensity per capita with higher Nitrate pollution. Specifically, a one-unit increase in the log of Nitrate at lag one is associated with increases of 0.0515 and 0.0994 in the log of nightlight per capita. These figures, though positive, are not marked with significance indicators, suggesting variability in their impact across different contexts.

Table 5.5: Nitrate impact on Log Nightlight per Capita

	(1)	(2)	(3)	(4)
	Dependent Variable: Log Nightlight per Capita			
Log Nitrate $_{t-1}$	0.0515			0.0994
	(0.0725)			(0.0776)
Log Temperature $_{t-1}$	0.4287**			-0.0917
	(0.1698)			(0.1296)
Log Nitrate $_{t-2}$		0.1095		0.0211
		(0.0703)		(0.0698)
Log Temperature $_{t-2}$		0.2927		0.2606
		(0.1781)		(0.1617)
Log Nitrate $_{t-3}$			-0.0784	0.0438
			(0.0913)	(0.0822)
Log Temperature $_{t-3}$			0.4654***	0.6007***
			(0.1425)	(0.1611)
District Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	547	550	551	388
R^2	0.0283	0.0187	0.0447	0.1105

Robust standard errors are clustered at district level and are in parentheses.

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

- Log Nitrate (t-2) and (t-3): At lag two (t-2) and three (t-3), the coefficients again show variable impacts of Nitrate on nightlight intensity, with some positive and one negative value (-0.0784) for t-3 in model (3)). The presence of both positive and negative effects across lags suggests the complex relationship between Nitrate pollution and economic activity.
- Log Temperature: The coefficients for temperature across different lags consistently show a positive relationship with nightlight intensity, with significant increases in models (1), (2), and especially marked significance in models (3) and (4) for (t-3) temperature, indicating the robust impact of temperature on economic conditions as measured by nightlight data.

The investigation into Nitrate pollution's effect on nightlight intensity per capita offers intriguing insights into how environmental quality might interact with economic vibrancy. Unlike BOD, which showed a generally negative impact on nightlight intensity, Nitrate levels demonstrate a more varied influence, with both positive and potentially negative associations depending on the time lag considered. This suggests that the economic repercussions of Nitrate pollution are not straightforward and may reflect the dual role of Nitrate as both a pollutant and a byproduct of agricultural and industrial activities, which could, in turn, drive economic activity in certain areas.

Significantly, the consistent positive effect of temperature across models underscores the multifaceted nature of environmental influences on economic indicators like nightlight intensity. The mixed results for Nitrate, coupled with the clear impact of temperature, highlight the complexity of drawing direct correlations between specific pollutants and economic outcomes. While increased Nitrate levels at certain lags hint at a possible boost to economic activity, likely through agricultural runoff enhancing local economic conditions, the overall picture remains intricate. These findings emphasize the need for a nuanced understanding of pollution's economic impacts, taking into account the variability and context-specific nature of these relationships.

The study by Pandit et al., 2016[82] aligns with our findings, suggesting that income levels can influence nitrate pollution through changes in agricultural practices. As income rises, the efficiency in nitrogen application improves, initially increasing but eventually reducing nitrate pollution, reflecting the Environmental Kuznets Curve hypothesis. This dynamic indicates that economic development might lead to better environmental management over time.

Mathewson et al., 2020[83] highlight the economic and health costs of nitrate pollution in Wisconsin, emphasizing the significance of addressing nitrate contamination to mitigate its adverse outcomes. This perspective underlines the economic implications of environmental policies targeting water quality.

Balazs's research, 2011[84] points out the socio-economic disparities in nitrate exposure, suggesting that economic conditions influence pollution levels and their health impacts. This complements our findings, suggesting that economic activity can both contribute to and be affected by nitrate levels.

Table 5.6: Not-drinkable Impact on Log Nightlight per Capita

	(1)	(2)	(3)	(4)
	Dependent Variable: Log Nightlight per Capita			
Not-drinkable $_{t-1}$	0.2262**			0.2237**
	(0.1169)			(0.0964)
Temperature $_{t-1}$	0.0058			0.0026
	(0.0036)			(0.0032)
Not-drinkable $_{t-2}$		0.0862		0.0443
		(0.0980)		(0.0749)
Temperature $_{t-2}$		0.0078**		0.0052
		(0.0041)		(0.0035)
Not-drinkable $_{t-3}$			-0.0684	-0.0825
			(0.1380)	(0.1360)
Temperature $_{t-3}$			0.0145***	0.0125***
			(0.0042)	(0.0035)
District Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	709	707	705	658
R^2	0.0174	0.0115	0.0349	0.0513

Robust standard errors are clustered at district level and are in parentheses.

Significance levels: * $p < 0.1$, ** $p < 0.05$, ***$p < 0.01$

Contrastingly, Rao's work, 2017[85] presents an alternative view where higher economic activity is associated with increased nitrate concentrations. Our study's initial positive correlation between nitrate levels and nightlight intensity supports this, although the long-term effects suggest a more complex relationship.

These literature insights corroborate our study's findings, illustrating the intricate links between economic activity, agricultural practices, and nitrate pollution. The variability in nitrate's impact on economic conditions underscores the importance of considering both immediate and extended effects in environmental and economic planning.

5.2.3 Not-drinkable and Nightlight per Capita

The analysis of the relationship between the availability of not-drinkable water (classified as not-drinkable) and nightlight intensity per capita unveils critical insights into how water quality impacts economic activity within a region. Table 5.6 systematically assesses this relationship, revealing significant findings that underscore the broader implications of environmental health on economic vibrancy.

- Not-drinkable (t-1) and (t-2): The analysis indicates a statistically significant positive relationship between the presence of not-drinkable water in the immediate and previous year (lags one and two) and nightlight intensity per capita. Specifically, a

one-unit increase in the 'Not-drinkable' variable at these lags is associated with increases in nightlight intensity, marked by coefficients of 0.2262 and 0.2237 (models (1) and (4) for (t-1) and a milder positive effect for (t-2) . These outcomes suggest that in the short term, regions with water quality issues may paradoxically see an uptick in economic activity, as measured by nightlight intensity. This could be due to increased industrial or remediation activities that inadvertently contribute to nightlight observations.

- Not-drinkable (t-3): Conversely, at lag three (t-3), the coefficients turn negative (-0.0684 and -0.0825), indicating a potential long-term dampening effect of not-drinkable water on economic activity. Although not statistically significant, this shift hints at the eventual economic downturn that may follow the initial uptick, as persistent water quality issues begin to outweigh any short-term economic boosts.

- Temperature: Across all lags, temperature shows a consistently positive effect on nightlight intensity, with significant associations observed for the (t-3) lag. This underscores the broader role of environmental conditions in shaping economic outcomes, beyond the direct impacts of water quality.

This analysis illuminates the intricate relationship between the quality of water and economic dynamics within a region. Initially, the presence of not-drinkable water correlates with an increase in nightlight intensity, potentially reflecting a surge in activities linked to industrial growth or efforts to combat water pollution. However, this short-term economic indicator does not capture the full story. As time progresses, the negative impact at the three-year lag suggests that the enduring presence of water quality issues may begin to erode economic resilience, highlighting a delayed adverse effect on economic desirability and activity.

The contradiction of short-term gains against long-term losses in economic activity underlines the complex consequences of environmental degradation. While immediate responses to pollution might temporarily elevate economic indicators like nightlight intensity, sustained environmental neglect could ultimately stifle economic potential. The role of ambient temperature in this dynamic further accentuates the multifaceted nature of environmental influences on economic health.

We have seen that Ridzuan[59] shows that increasing water pollution widens the gap in income inequality in India. But he does not find any evidence for environmental Kuznets curve hypothesis for water pollution in India, which is visible in our results when the relationship between not-drinkable changes from positive to negative in the third year.

A study from southern Spain assesses the socio-economic impacts of water quality changes in the Guadiana Estuary[86], focusing on tourism related to bathing. It employs an integrated simulation model linking ecological and socio-economic components through the Blue Flag Award, which is dependent on fecal bacterial thresholds. The Economic Base Model quantifies the impact of water quality on employment and resident

population due to changes in coastal water quality. A Cost-Benefit Analysis, incorporating a monetary valuation of water quality changes on human welfare through the Contingent Valuation Method, provides scenario evaluations. The study highlights the importance of water quality for tourism, suggesting that improving it can have significant positive socio-economic effects.

A similar survey based study was conducted for the river Yamuna, which is the biggest tributary of Ganga in 2020[87]. They explore local ideas from river communities about the livelihood and health impacts river Yamuna's pollution has on heir lives. Most respondents were aware of the negative impacts of poor water quality but did not want to attribute it to the river Yamuna, showcasing the religious role rivers play in India. The same faith mindset can be applied for river Ganga in large parts as it is also very much plays a religious part in the states.

We see the impacts of tourism on the pollution levels in Ganga[[88], [89]]. But we don't see any work done to find the impacts of water pollution on the tourism industry, which is a major economic driver in some districts from which the Ganga river passes like Haridwar, Rishikesh, Varanasi etc.

This chapter has detailed the outcomes of Panel OLS regressions analyzing the effects of river pollution measures, sourced from the CPCB, on health outcomes (specifically, deaths due to diseases) and economic activity (proxied through nightlight intensity per capita). It delves into the magnitude and direction of each pollution variable's impact on these dependent variables, including how these effects may persist or change over time. Our models are fortified with rigorous statistical techniques, including fixed effects to control for unobserved heterogeneity across districts and years, clustering standard errors to address potential autocorrelation, and incorporating control variables to refine our estimates. We also see the broader impacts of our results and see where they lie in the existing literature.

In the forthcoming chapter, we will conclude our findings.

6 Conclusion

This research is focused on studying the health and socio-economic impacts of river pollution, specifically in the Ganga river in India. Ganga is the 5th largest polluted river and the basin is densely populated. Our goal is to find the repercussions of increasing pollution in the river to the communities living nearby. To that end, we utilize water pollution data released by CPCB every year for water pollution measures. To study the health impacts, we used the official number of deaths and cases due to diseases in Indian states. And to study the socio-economic impacts, we utilize the nightlight data as a proxy for economic activity in the region. We run 2-way fixed effects PanelOLS regressions between selected pollutants and an overall indicator of pollution on the number of deaths and on the nightlight per capita of a district for 2 years lag and 3 years lag respectively. These linear Panel regressions after controlling for temperature and fixed effects give the following results:

- We see that the relation between water pollution and deaths is almost always positive in our findings supporting our hypothesis. The effects for specific pollutants is small but statistically significant in some lags (BOD after 2 years). We observe from the literature that water pollution is indeed directly linked with human health, accounting for 1.4 million deaths each year globally. Specifically BOD is a measure of organic pollutants in the water discharged mainly from untreated urban sewage, animal farming, and industrial pollution. Organic pollutants are a cause of diarrhoea and other gastrointestinal diseases which are a leading cause of deaths, especially in children. We also see in the literature that the effects are disproportional biased towards infant mortality and child mortality in less developed regions.

- The second level impact of river water pollution is between the pollutants and the economic activity proxied through nightlight per capita. We find most stark results in BOD where in each year till 3 years of lag, we see a negative relationship between the dependant and the independent variables with statistically significant results. This is in favor of our original hypothesis. This is most evident at lag 3 with a 0.050 decrease in nightlight per capita for a one-unit increase in BOD. These results signify the negative relationship is not only present but also persistent over time.

- The relation between Nitrate and not-drinkable water on nightlight is more nuanced than BOD. For the first two years, the results support our alternate hypothesis of a positive correlation. In both these cases, we find that the results follow an

Environmental Kuznets Curve kind of relationship, which states that the environmental pollution and economic growth are initially positively correlated, however, when the income crosses a threshold, the economic growth allows for environmental remediation. We see a similar relationship in our opposite study, where a rise in pollution is linked with a rise in economic activity, but when the pollution crosses a certain threshold, it causes a deterioration in economic activity.

In conclusion, this thesis has systematically explored the multifaceted impacts of river pollution on health and socio-economic conditions along the Ganga river. Our findings affirm the detrimental effects of pollution on community health, evidenced by a significant correlation between water pollutants and mortality rates. Additionally, the economic analysis, through nightlight intensity, unveils a nuanced relationship between pollution and economic activity, with BOD consistently showing a negative impact. This comprehensive study not only corroborates existing literature but also offers new insights into the dynamics of pollution and its broader implications, underscoring the urgent need for integrated environmental and health policy interventions.

Bibliography

[1] Environmental Protection Agency, "Drinking water," 2019.

[2] Wikipedia contributors, "Central pollution control board — wikipedia, the free encyclopedia," 2024. [Online; accessed 18-February-2024].

[3] S. Asher, T. Lunt, R. Matsuura, and P. Novosad, "Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in india using the shrug open data platform," *The World Bank Economic Review*, vol. 35, no. 4, 2021.

[4] J. V. Henderson, A. Storeygard, and D. N. Weil, "A Bright Idea for Mesuring Economic Growth," *American Economic Review*, 2011.

[5] V. Schippers and W. Botzen, "Uncovering the veil of night light changes in times of catastrophe," *Natural Hazards and Earth System Sciences*, vol. 23, pp. 179–204, Jan. 2023.

[6] S. Asher, T. Lunt, R. Matsuura, and P. Novosad, "Development research at high geographic resolution: An analysis of night-lights, firms, and poverty in india using the shrug open data platform," *World Bank Economic Review*, vol. 35, no. 4, pp. 845–871, 2021.

[7] Ministry of Statistics and Programme Implementation, "Statistical year book india 2015." Ministry of Statistics and Programme Implementation.

[8] Central Pollution Control Board, "Integrated waste management," 5 2011. Accessed: 2023-03-26.

[9] J. Russ, E. Zaveri, S. Desbureaux, R. Damania, and A.-S. Rodella, "The impact of water quality on gdp growth: Evidence from around the world," *Water Security*, vol. 17, p. 100130, 2022.

[10] M. T. H. van Vliet, J. Thorslund, M. Strokal, N. Hofstra, M. Flörke, H. Ehalt Macedo, A. Nkwasa, T. Tang, S. S. Kaushal, R. Kumar, A. van Griensven, L. Bouwman, and L. M. Mosley, "Global river water quality under climate change and hydroclimatic extremes," *Nature Reviews Earth & Environment*, vol. 4, no. 10, pp. 687–702, 2023.

[11] World Bank, "Global water security and sanitation partnership (gwsp) annual report 2023," 2023.

[12] U.S. Environmental Protection Agency, "Nutrient pollution ¿ the effects: Economy," 2014.

[13] M. Kummu, H. de Moel, P. J. Ward, and O. Varis, "How close do we live to water? a global analysis of population distance to freshwater bodies," *PLoS One*, vol. 6, no. 6, p. e20578, 2011.

[14] A. K. Biswas, *History of hydrology*. North-Holland and American Elsevier, 1970.

[15] M. Achtnicht, M. Borell, K. Gantert, M. Kappler, B. Müller, B. Boockmann, G. Klee, R. Krumm, K. Neugebauer, G. Hunya, H. Vidovic, and R. Römisch, "Socio-economic assessment of the danube region: State of the region, challenges and strategy development. final report part i," zew gutachten/forschungsberichte, Mannheim u.a., 2014.

[16] R. Damania, S. Desbureaux, A.-S. Rodella, J. Russ, and E. Zaveri, *Quality Unknown: The Invisible Water Crisis*. Washington, DC: World Bank, 2019. License: CC BY 3.0 IGO.

[17] S. K. Sahoo and S. Goswami, "Theoretical framework for assessing the economic and environmental impact of water pollution: A detailed study on sustainable development of india," pp. 23–34, 01 2024.

[18] The Water Project, "The water crisis: The importance of clean water to health." `https://thewaterproject.org/why-water/health`, 2021. Retrieved 13 February 2024.

[19] United Nations Children's Fund (UNICEF) and World Health Organization (WHO), "Progress on household drinking water, sanitation and hygiene 2000–2022: special focus on gender," tech. rep., United Nations Children's Fund (UNICEF) and World Health Organization (WHO), New York, 2023.

[20] "These are the challenges facing india's most sacred river," 2019.

[21] S. Ridzuan, "Inequality and water pollution in india," *Water Policy*, vol. 23, 06 2021.

[22] United Nations, "Climate change: Water." `https://www.un.org/en/climatechange/science/climate-issues/water`. Accessed: 13 February 2024.

[23] E. R. Jones, M. F. P. Bierkens, P. J. T. M. van Puijenbroek, L. R. P. H. van Beek, N. Wanders, E. H. Sutanudjaja, and M. T. H. van Vliet, "Sub-saharan africa will increasingly become the dominant hotspot of surface water pollution," *Nature Water*, vol. 1, no. 7, pp. 602–613, 2023.

[24] L. Liao, M. Du, and Z. Chen, "Environmental pollution and socioeconomic health inequality: Evidence from china," *Sustainable Cities and Society*, vol. 95, p. 104579, 2023.

[25] B. Adelodun, F. O. Ajibade, J. O. Ighalo, G. Odey, R. G. Ibrahim, K. Y. Kareem, H. O. Bakare, A. O. Tiamiyu, T. F. Ajibade, T. S. Abdulkadir, K. A. Adeniran, and K. S. Choi, "Assessment of socioeconomic inequality based on virus-contaminated water usage in developing countries: A review," *Environmental Research*, vol. 192, p. 110309, Jan 2021.

[26] National Institute for Transforming India (NITI) Aayog, "Composite water management index," 2018. Accessed: 2023-03-26.

[27] A. Tyagi and G. Hutton, "Economic impacts of inadequate sanitation in india," tech. rep., Water and Sanitation Program (WSP), World Bank, New Delhi, India, 01 2011.

[28] P. Hirani and V. Dimble, "Water pollution in india: Data is the first step towards a solution," 10 2019. Programme director, Water-to-Cloud, Tata Centre for Development at UChicago; Assistant director, research and strategy, Tata Centre for Development at UChicago.

[29] R. Damania, S. Desbureaux, A.-S. Rodella, J. Russ, and E. Zaveri, *Quality unknown: the invisible water crisis*. World Bank Publications, 2019.

[30] T. Schellenberg, V. Subramanian, G. Ganeshan, D. Tompkins, and R. Pradeep, "Wastewater discharge standards in the evolving context of urban sustainability–the case of india," *Frontiers in Environmental Science*, vol. 8, 2020.

[31] J. C. Morris, I. Georgiou, E. Guenther, and S. Caucci, "Barriers in implementation of wastewater reuse: Identifying the way forward in closing the loop," *Circular Economy and Sustainability*, vol. 1, no. 1, pp. 413–433, 2021.

[32] Wikipedia contributors, "Ganges — wikipedia, the free encyclopedia," 2024. [Online; accessed 18-February-2024].

[33] National Mission for Clean Ganga, "Ganga basin." [Online; accessed 18-February-2024].

[34] H. C. Bonsor, A. M. MacDonald, K. M. Ahmed, W. G. Burgess, M. Basharat, R. C. Calow, A. Dixit, S. S. D. Foster, K. Gopal, D. J. Lapworth, M. Moench, A. Mukherjee, M. S. Rao, M. Shamsudduha, L. Smith, R. G. Taylor, J. Tucker, F. van Steenbergen, S. K. Yadav, and A. Zahid, "Hydrogeological typologies of the indo-gangetic basin alluvial aquifer, south asia," *Hydrogeology Journal*, vol. 25, no. 5, pp. 1377–1406, 2017.

[35] D. S. Bhargava, "Nature and the ganga," *Environmental Conservation*, vol. 14, no. 4, p. 307–318, 1987.

[36] K. Khairnar, "Ganges: special at its origin," *Journal of Biological Research-Thessaloniki*, vol. 23, p. 16, 2016.

[37] C. Kumar, A. Ghosh, M. Debnath, P. Bhadury, *et al.*, "Seasonal dynamicity of environmental variables and water quality index in the lower stretch of the river ganga," *Environmental Research Communications*, vol. 3, no. 7, p. 075008, 2021.

[38] Editors of Encyclopaedia Britannica, "Gomati river," 10 2023.

[39] B. Das, A. Ray, T. Nirupada Chanu, R. Baitha, and S. BAYEN, "Fisheries of the river ganga: Present vs past," *Science and Culture*, vol. 88, 10 2022.

[40] "Ganges river," 2023.

[41] United Nations Environment Programme, "Restoring india's holiest river," 2021.

[42] National Green Tribunal, U.P, Lucknow, "Updated report of oversight committee in compliance of order of hon'ble national green tribunal passed in o.a 200 of 2014 in re: M.c. mehta versus union of india & ors," 2020.

[43] Wikipedia contributors, "Pollution of the ganges." `https://en.wikipedia.org/wiki/Pollution_of_the_Ganges`, 2023.

[44] A. Mathur, "Namami gange scheme-a success or mere propaganda?," *GLS Law Journal*, vol. 2, no. 2, pp. 54–64, 2020.

[45] A. Balkrishna, S. K. Singh, R. Pathak, and V. Arya, "Namami gange: An opinion based framework and possible resolution," *Authorea Preprints*, 2022.

[46] A. Kumar, A. Ajay, B. Dasgupta, P. Bhadury, and P. Sanyal, "Deciphering the nitrate sources and processes in the ganga river using dual isotopes of nitrate and bayesian mixing model," *Environmental Research*, vol. 216, no. Pt 4, p. 114744, 2023. Epub 2022 Nov 9.

[47] M. H. Ward, R. R. Jones, J. D. Brender, T. M. de Kok, P. J. Weyer, B. T. Nolan, C. M. Villanueva, and S. G. van Breda, "Drinking water nitrate and human health: An updated review," *International Journal of Environmental Research and Public Health*, vol. 15, no. 7, p. 1557, 2018.

[48] C. N. Sawyer, P. L. McCarty, and G. F. Parkin, *Chemistry for Environmental Engineering and Science*. New York: McGraw-Hill, 5 ed., 2003.

[49] V. O, G. B, Z. M, D. C, B. F, A. A, and P. A no. KJ-NA-29451-EN-N (online),KJ-NA-29451-EN-C (print),KJ-NA-29451-EN-E (ePub), 2018.

[50] Y. Wen, G. Schoups, and N. van de Giesen, "Organic pollution of rivers: Combined threats of urbanization, livestock farming and global climate change," *Scientific Reports*, vol. 7, no. 1, p. 43289, 2017.

[51] O. Malve, S. Tattari, J. Riihimäki, E. Jaakkola, A. Voß, R. Williams, and I. Bärlund, "Estimation of diffuse pollution loads in europe for continental scale modelling of loads and in-stream river water quality," *Hydrol. Process.*, vol. 26, pp. 2385–2394, 2012.

[52] United States Environmental Protection Agency, "Water: Monitoring and Assessment: 5.2 Dissolved Oxygen and Biological Oxygen Demand." `https://www.epa.gov/`, 2012.

[53] U.S. Department of the Interior, U.S. Geological Survey, "Water Quality Information by Topic: Biological Oxygen Demand (BOD) and Water." `https://www.usgs.gov/`, Accessed 12 Mar. 2020.

[54] National Federation of Group Water Schemes, "Guide to the Parameters in the European Communities. What's in your water?." `https://www.nfgws.ie/`, 2007. S. I. No. 278 of 2007.

[55] Central Pollution Control Board (CPCB), "Guidelines for water quality monitoring." `https://cpcb.nic.in/wqm/Guidelines_Water_Quality_Monitoring_2017.pdf`, 2017.

[56] Central Pollution Control Board, "Bulletin vol-i, july 2016." Ministry of Environment and Forests, Govt. of India, Parivesh Bhawan, Delhi, 2016.

[57] V. Dutta, D. Dubey, and S. Kumar, "Cleaning the river ganga: Impact of lockdown on water quality and future implications on river rejuvenation strategies," *The Science of the Total Environment*, vol. 743, p. 140756, 2020.

[58] M. Kumar, "The ganga pollution is the big problem,"

[59] Wikipedia contributors, "Water pollution in india," 2023.

[60] J. V. Henderson, A. Storeygard, and D. N. Weil, "Measuring economic growth from outer space," *The American Economic Review*, vol. 102, no. 2, pp. 994–1028, 2012.

[61] U. Dorji, C. Siripanpornchana, N. Surasvadi, A. Plangprasopchok, and S. Thajchayapong, "Exploring night light as proxy for poverty and income inequality approximation in thailand," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pp. 1082–1087, 2019.

[62] C. Elvidge, K. Baugh, S. Anderson, P. Sutton, and G. Tilottama, "The night light development index (nldi): A spatially explicit measure of human development from satellite data," *Social Geography*, vol. 7, 07 2012.

[63] A. Mveyange, "Night lights and regional income inequality in africa," Tech. Rep. 085, Helsinki, Finland, September.

[64] A. Singhal, S. Sahu, S. Chattopadhyay, A. Mukherjee, and S. N. Bhanja, "Using night time lights to find regional inequality in india and its relationship with economic development," *PLoS One*, vol. 15, no. 11, p. e0241907, 2020.

[65] Global Administrative Areas, "GADM database of Global Administrative Areas, version 2.0." `http://www.gadm.org`, 2012.

[66] Central Pollution Control Board (CPCB), "Functions." `https://cpcb.nic.in/functions/`.

[67] National Mission for Clean Ganga, "Details of districts on river ganga and tributaries (riverwise)," 2017.

[68] C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh, "VIIRS night-time lights," *International Journal of Remote Sensing*, vol. 38, pp. 5860–5879, 2017.

[69] L. Volterra, M. Boualam, A. Menesguen, J. Duguet, J. Duchemin, and X. Bonnefoy, "Eutrophication and health," *World Health Organization & European Commission. Luxembourg.[Online]. Available in http://www. ypeka. gr/LinkClick. aspx*, 2002.

[70] W. Scott, R. Van Steenderen, D. Welch, W. R. Commission, *et al.*, "Health aspects of eutrophication," *Available from the National Technical Information Service, Springfield VA*, vol. 22161, 1985.

[71] L. Lin, H. Yang, and X. Xu, "Effects of water pollution on human health and disease heterogeneity: A review," *Frontiers in Environmental Science*, vol. 10, 2022.

[72] United Nations Department of Economic and Social Affairs, "Biochemical oxygen demand - methodology sheets." `https://www.un.org/esa/sustdev/natlinfo/indicators/methodology_sheets/freshwater/biochemical_oxygen_demand.pdf`.

[73] M. H. Ward, R. R. Jones, J. D. Brender, T. M. de Kok, P. J. Weyer, B. T. Nolan, C. M. Villanueva, and S. G. van Breda, "Drinking water nitrate and human health: An updated review," *International journal of environmental research and public health*, vol. 15, no. 7, p. 1557, 2018.

[74] L. Grout, T. Chambers, S. Hales, M. Prickett, M. G. Baker, and N. Wilson, "The potential human health hazard of nitrates in drinking water: a media discourse analysis in a high-income country," *Environmental health : a global access science source*, vol. 22, no. 1, p. 9, 2023.

[75] S. Fossen Johnson, "Methemoglobinemia: Infants at risk," *Current problems in pediatric and adolescent health care*, vol. 49, no. 3, pp. 57–67, 2019.

[76] Minnesota Department of Health, "Nitrate in well water." `https://www.health.state.mn.us/communities/environment/water/wells/waterquality/nitrate.html`, 2023.

[77] S. Dwivedi, S. Mishra, and R. D. Tripathi, "Ganga water pollution: A potential health threat to inhabitants of ganga basin," *Environment International*, vol. 117, pp. 327–338, 2018.

[78] The Energy and Resources Institute (TERI), "Title of the environmental survey." `https://www.teriin.org/environmentalsurvey/`, 2015.

[79] BBC News, "Title of the article." `https://www.bbc.com/news/magazine-28112403`, 2014. Accessed: 2023-03-26.

[80] P. Joshi and K. A. Beck, "Biological oxygen demand and economic growth: An empirical investigation," *Water Economics and Policy*, vol. 01, no. 02, p. 1550001, 2015.

[81] S. K. Chapagain, G. Mohan, A. B. Rimba, C. Payus, I. M. Sudarma, and K. Fukushi, "Analyzing the relationship between water pollution and economic activity for a more effective pollution control policy in bali province, indonesia," *Sustainable Environment Research*, vol. 32, no. 1, p. 5, 2022.

[82] M. Pandit and K. P. Paudel, "Water pollution and income relationships: A seemingly unrelated partially linear analysis," *Water Resources Research*, vol. 52, no. 10, pp. 7668–7689, 2016.

[83] P. D. Mathewson, S. Evans, T. Byrnes, A. Joos, and O. V. Naidenko, "Health and economic impact of nitrate pollution in drinking water: a wisconsin case study," *Environmental Monitoring and Assessment*, vol. 192, no. 11, p. 724, 2020.

[84] C. Balazs, R. Morello-Frosch, A. Hubbard, and I. Ray, "Social disparities in nitrate-contaminated drinking water in california's san joaquin valley," *Environmental Health Perspectives*, vol. 119, no. 9, pp. 1272–1278, 2011.

[85] E. Prakasa Rao, K. Puttanna, K. Sooryanarayana, A. Biswas, and J. Arunkumar, "21 - assessment of nitrate threat to water quality in india," in *The Indian Nitrogen Assessment* (Y. P. Abrol, T. K. Adhya, V. P. Aneja, N. Raghuram, H. Pathak, U. Kulshrestha, C. Sharma, and B. Singh, eds.), pp. 323–333, Elsevier, 2017.

[86] M. H. E. Guimarães, A. Mascarenhas, C. Sousa, T. Boski, and T. P. Dentinho, "The impact of water quality changes on the socio-economic system of the guadiana estuary: an assessment of management options," *Ecology and Society*, vol. 17, no. 3, 2012.

[87] N. Maurya, P. Hirani, H. Sharma, A. Balhara, P. Pandey, S. Mundlay, A. Kumar, R. Chauhan, and S. Guha, "River yamuna: Deteriorating water quality its socio-economic impact. voices from the ground," 08 2020.

[88] S. Sharma and M. Agrawal, "The river ganga and its pollution- tourists' perception visiting ghats of varanasi," *Turizam*, vol. 25, pp. 55–71, 07 2021.

[89] R. Madan, S. Chaudhry, M. Sharma, and S. Madan, "Determination of water quality index of indraprastha estate region and the vicinity area in delhi, india," *International Journal for Environmental Rehabilitation and Conservation*, vol. IX, no. 1, pp. 204–216, 2018. Open Access Publication. This work is licensed under Attribution-Non Commercial 4.0 International (https://creativecommons.org/licenses/by/4.0/).

7 Appendix

The Appendix chapter serves as a comprehensive repository for additional analyses that complement the main findings of the study. This includes a series of subset regressions that explore the nuanced relationships between pollution and nightlight intensity across various categorizations of districts along the Ganges river. These subsets are defined based on geographical positioning, industrial activity levels, and specific regional characteristics, enabling a detailed examination of the socio-economic impacts of river pollution under different conditions.

7.1 Framework of Subset Regressions

This section outlines the framework for the subset regressions included in the Appendix, providing insights into the differential impacts of pollution on economic activity as captured by nightlight intensity.

7.1.1 Districts with More Industries than Average

Objective: Investigate how industrial activity levels influence the relationship between pollution and economic activity as indicated by nightlight intensity.

Data and Prepossessing For this analysis, we leveraged the 2023 dataset detailing the number of industries per district in India, obtained from The National Data and Analytics Platform (NDAP). NDAP's comprehensive dataset provides a valuable resource for understanding the distribution of manufacturing, services, and trading sectors across India's diverse administrative landscape.

Methodology The data was grouped by state and district and then merged with the combined river pollution data and nightlight data. To focus on the districts with substantial industrial activity, districts with a total number of industries above the mean were identified and flagged (more_than_mean).

The number of districts categorized as having more than the mean industrial activity totaled 236. This categorization allowed for a targeted regression analysis to examine the effects of industrial activity levels on the economic vitality of a region, as evidenced by nightlight intensity.

To ensure the robustness of the analysis, both log-transformed and lagged variables were created for the key metrics, including the BOD, Nitrate and temperature, alongside the dependent variable, the annual mean nightlight intensity.

Regressions This PanelOLS regression framework allowed for controlling both entity and time-fixed effects, providing a nuanced understanding of the impacts of industrial activities on nocturnal light emissions. The inclusion of lagged variables aimed to capture the delayed effects of pollution and temperature on economic activities. The regression formula employed is as follows for various lag periods and a combination of pollution measures:

General Formula:

$$\begin{aligned} \log(\text{VIIRS Annual Mean}) \sim\ & \log(\text{Pollution Measure}_{\text{lag 1}}) + \\ & \log(\text{Temperature}_{\text{lag 1}}) + \\ & \log(\text{Pollution Measure}_{\text{lag 2}}) + \\ & \log(\text{Temperature}_{\text{lag 2}}) + \\ & \log(\text{Pollution Measure}_{\text{lag 3}}) + \\ & \log(\text{Temperature}_{\text{lag 3}}) + \\ & \text{EntityEffects} + \text{TimeEffects} \end{aligned} \tag{7.1}$$

Results The regression results are summarized in the table below:

Table 7.1: Regression Results for Districts with More Than Mean Population

Variable	**Coefficient**	**Std. Err.**	**T-stat**	**P-value**
	Lag 1 Variables:			
log(BOD Max lag 1)	-0.0347	0.0227	-1.524	0.128
log(Temperature Max lag 1)	0.0099	0.1235	0.080	0.936
	Lag 2 Variables:			
log(BOD Max lag 2)	-0.0290	0.0209	-1.391	0.165
log(Temperature Max lag 2)	-0.0028	0.1227	-0.023	0.982
	Lag 3 Variables:			
log(BOD Max lag 3)	-0.0266	0.0200	-1.326	0.186
log(Temperature Max lag 3)	0.2204	0.1165	1.892	0.060
F-test for Poolability: 58.064; P-value: <0.001				
Included effects: Entity, Time				

Key findings The regression analyses across different subsets of districts reveal the nuanced impact of pollution on nightlight intensity, serving as a proxy for economic activity. The inclusion of temperature as a control variable across all regressions ensures a more accurate understanding of pollution's effects, accounting for its influence on aquatic life's metabolic activities and biochemical processes in the Ganges River.

The negative coefficient for the BOD lagged by two years suggests that increased levels of organic pollution negatively impact economic activity in these industrial districts. This could imply that the detrimental environmental effects of industrial pollution, such as water contamination, may outweigh the immediate economic benefits derived from

industrial activities. The lagged impact indicates that the adverse effects of pollution on economic activity may not be immediate but become more pronounced over time, possibly due to the cumulative nature of environmental degradation and its broader impacts on public health, agricultural productivity, and overall quality of life.

On the other hand, the positive coefficients associated with temperature in the regressions hint at the complex role environmental factors play in economic activities. Higher temperatures could be indicative of climatic conditions favorable to certain industries or may reflect broader, long-term climatic trends that impact economic productivity, such as longer growing seasons in agricultural districts or increased energy consumption for cooling purposes.

However, the overall low R-squared values and the mixed significance levels across the regressions suggest that while there is some relationship between river pollution, temperature, and economic activity, it is not strong and is influenced by a myriad of other unaccounted factors. The results highlight the complexity of isolating the impact of river pollution on economic activity, especially in industrial districts where economic outputs are influenced by various factors, including but not limited to environmental conditions.

In conclusion, the subset analysis provides valuable insights but also underscores the challenges in directly correlating river pollution with economic activity in industrially dense districts. It suggests a need for more comprehensive models that include a wider range of variables to better understand these dynamics. For policymakers and economic planners, these findings emphasize the importance of adopting sustainable industrial practices that mitigate environmental impact to ensure long-term economic resilience and growth.

7.1.2 Districts with More than Mean Population

Objective The aim of this subsection is to explore the relationship between river pollution and nightlight intensity, serving as a proxy for economic activity, in districts with populations exceeding the mean population, based on the 2011 census data.

Data and Prepossessing Utilizing the 2011 census data as a benchmark—given the absence of more recent comprehensive population data—we categorized districts based on their population size relative to the mean. This analysis subset consisted of 345 districts, representing areas with relatively higher population densities. Such categorization allows us to examine whether and how the interplay between pollution and economic activity differs in more densely populated regions.

Methodology The analysis followed a systematic approach, starting with the conversion of key pollution measures and nightlight intensity data into numeric formats and applying logarithmic transformations to ensure normality and reduce skewness. Subsequently, we generated lagged variables for pollution measures to capture potential delayed effects on economic activity. This preparation enabled the employment of Pan-

elOLS regression models to quantify the impacts, accounting for both entity and time-fixed effects to isolate the effects of interest from other confounding factors.

Regressions The regression analysis was structured around the hypothesis that the relationship between river pollution and economic activity, as proxied by nightlight intensity, might be more pronounced or distinct in more populous districts. By incorporating lags of up to three years for key pollution metrics, alongside temperature as a control variable, the models aimed to discern both immediate and delayed effects, as well as to account for environmental factors potentially influencing this relationship.

$$\begin{aligned}\log(\text{VIIRS Annual Mean}) = \beta_0 &+ \beta_1 \log(\text{BOD Max}_{\text{lag}_1}) + \beta_2 \log(\text{Temperature Max}_{\text{lag}_1}) + \\ &\beta_3 \log(\text{BOD Max}_{\text{lag}_2}) + \beta_4 \log(\text{Temperature Max}_{\text{lag}_2}) + \\ &\beta_5 \log(\text{BOD Max}_{\text{lag}_3}) + \beta_6 \log(\text{Temperature Max}_{\text{lag}_3}) + \\ &\text{EntityEffects} + \text{TimeEffects} + \varepsilon\end{aligned} \tag{7.2}$$

Results The regression outputs indicate that the relationship between river pollution and nightlight intensity in districts with higher-than-average populations is nuanced. The variations in R-squared values and the significance of coefficients across different lag periods suggest that while pollution does impact economic activity, the strength and direction of this impact are influenced by multiple factors, including population density, the temporal lag of pollution effects, and ambient temperature conditions.

Table 7.2: Regression Results for Districts with More Than Mean Population

Variable	**Coefficient**	**Std. Err.**	**T-stat**	**P-value**
	Lag 1 Variables:			
log(BOD Max lag 1)	-0.0347	0.0227	-1.524	0.128
log(Temperature Max lag 1)	0.0099	0.1235	0.080	0.936
	Lag 2 Variables:			
log(BOD Max lag 2)	-0.0290	0.0209	-1.391	0.165
log(Temperature Max lag 2)	-0.0028	0.1227	-0.023	0.982
	Lag 3 Variables:			
log(BOD Max lag 3)	-0.0266	0.0200	-1.326	0.186
log(Temperature Max lag 3)	0.2204	0.1165	1.892	0.060
F-test for Poolability: 58.064; P-value: <0.001				
Included effects: Entity, Time				

Key findings The regression analyses conducted on districts with populations exceeding the 2011 mean reveal nuanced insights into the relationship between river pollution and economic activity, as proxied by nightlight intensity. A key finding is the subtle

yet discernible impact of pollution levels, specifically the BOD, on nightlight intensity, suggesting that higher pollution levels might be associated with lower economic activity in more densely populated districts. This relationship is nuanced, as indicated by the varying significance and coefficients of the lagged pollution and temperature variables across different regression models.

Particularly, the negative coefficients associated with BOD in certain lags point to a detrimental effect of pollution on economic vitality. However, the presence of positive coefficients for temperature in some models highlights the complex interplay between environmental factors and economic outcomes. These mixed results suggest that while pollution may have an adverse effect on economic activity, factors like temperature and possibly other unobserved variables also play significant roles in shaping economic conditions in these regions.

Moreover, the low R-squared values across the models indicate that while there is a relationship between pollution, temperature, and nightlight intensity, it is not overwhelmingly strong, suggesting the influence of other factors not captured in the model. This complexity underscores the challenges of isolating the effects of pollution on economic activity and highlights the need for a comprehensive approach that considers a broader range of environmental, social, and economic variables to fully understand the dynamics at play.

Overall, the findings suggest that in districts with higher populations, pollution does have an observable impact on economic activity, but this relationship is modulated by other factors, including environmental conditions such as temperature. This emphasizes the importance of adopting multifaceted strategies that address pollution while considering the broader socio-economic and environmental context to foster sustainable economic growth.

7.1.3 Upstream vs. Downstream States (Uttarakhand vs. West Bengal)

Objective: The analysis aims to understand the differential impact of river pollution on nightlight per capita—a proxy for economic activity—between the upstream state of Uttarakhand and the downstream state of West Bengal.

Data and Preliminary Steps: An examination of the dataset reveals stark differences in pollution levels, with mean fecal coliform levels in West Bengal being significantly higher than those in Uttarakhand. This sets the stage for comparing the two states in terms of how pollution levels relate to nightlight per capita, serving as a proxy for economic activity.

Methodology: The methodology involves running a series of PanelOLS regressions for each state, using pollution measures (such as BOD and fecal coliform) lagged by one period, with temperature as a control variable. The regression accounts for both entity effects (differences between districts within the state) and time effects (year-to-year changes).

Regression Uttarakhand Analysis:

Uttarakhand, being at the upstream, shows varied results across the pollution parameters:

BOD: The results indicate a weak and non-significant association with nightlight per capita. Fecal Coliform: The positive coefficient was small and statistically significant, suggesting some economic activities may thrive despite higher pollution levels. Nitrate: No significant relationship with nightlight per capita was observed, implying the complexity of pollution impact. not-drinkable: A non-significant relationship with nightlight per capita suggests economic activity may not be immediately affected by water potability. West Bengal Analysis:

In contrast, West Bengal, located downstream, exhibits different dynamics:

BOD: The negative but non-significant coefficient suggests a minimal impact on economic activity as indicated by nightlight per capita. Fecal Coliform: Contrary to Uttarakhand, a statistically significant positive relationship was found, which could indicate industries or facilities contributing to nightlight despite high pollution levels. Nitrate: Similarly, no significant impact on economic activity was identified. not-drinkable: There is no significant relationship with economic activity, which might indicate an economy that has adapted to high levels of pollution or other unaccounted factors at play.

Key findings Comparing the two states, it appears that river pollution's effect on economic activity is complex and nuanced. In Uttarakhand, the relationship between pollution and nightlight per capita is not strongly defined. In contrast, West Bengal shows some significant positive relationships, which may be attributed to economic resilience or adaptation to higher pollution levels due to its downstream position.

7.1.4 Districts with More Literacy Rate vs Less Literacy Rate

Data and Preliminary Steps The data from the SHRUG platform was utilized, where the literacy rate pc11_pca_p_lit for each district was examined. The literacy rate variable represents the population count of literate individuals in a district. Summary statistics of this variable were observed as follows:

- Average literate population across the districts was approximately 1.19 million,, with a standard deviation of about 1.07 million.
- The districts with literacy numbers at the lower end had a population of around 4,436 literate individuals.
- The median literacy population was approximately 957,346.

District Categorization Districts were dichotomized into 'above_mean_literacy' and 'below_mean_literacy' based on whether their literacy rate was above or below the mean literacy rate of the sample.

- The initial dataset contained 640 districts, out of which 263 were categorized as 'above_mean_literacy' and 377 as 'below_mean_literacy' after merging with the pollution and nightlight data.

Key Observations The categorization based on literacy revealed that a majority of the districts fell into the 'above_mean_literacy' category, suggesting that these districts have a higher number of individuals who can read and write compared to the 'below_mean_literacy' districts.

Implications for Analysis Understanding the literacy distribution along the Ganges is crucial as it may influence the interpretation of nightlight data. Literacy rates can impact economic development, as a more literate population may have better access to information, higher employment rates, and potentially more resilience to environmental issues such as river pollution. Moreover, literate districts might be more proactive in addressing pollution due to greater awareness and organizational capacities.

In conclusion, this literacy-based division provides a starting point for further analysis on the socio-economic dynamics affecting the Ganges River regions. It underscores the need for multifaceted development approaches that consider education as a critical component in environmental and economic planning.

7.2 Common Limitations and Considerations:

Across all subset analyses, certain limitations are acknowledged, reflecting the inherent challenges of econometric studies. A recurring consideration is the reliance on nightlight intensity as a singular proxy for economic activity, which might not fully capture the breadth of economic dynamics, particularly in non-luminous sectors or daytime economies. Additionally, the employment of lagged variables as proxies for the temporal impact of pollution may not fully capture immediate effects or the diverse temporal responses across different economic activities. While the study meticulously controls for observable confounders, the possibility of omitted variable bias remains, necessitating caution in interpreting the results as causal relationships. We have also used old data for population and literacy studies from 2011 which might not capture the results today due to absence of new census information.These limitations underscore the importance of adopting a cautious stance when extrapolating the findings and highlight the necessity for comprehensive modeling that incorporates a broader spectrum of socio-economic and environmental factors to unravel the complex tapestry of interactions between river pollution and economic outcomes.

7.3 Technical Considerations

For each subset regression: - The selection criteria for districts within each subset are explicitly defined to ensure clarity and replicability. - Control variables, including temperature, are consistently included across all models to account for potential confounders.

- The specifications of the regression models, including the treatment of lagged variables and interaction terms, are detailed to facilitate interpretation of the results.

The Appendix chapter, through its focused subset regressions, enriches the study's findings by unveiling the differential impacts of river pollution on economic activities across diverse regions along the Ganges River. By dissecting these relationships within specific contexts—geographical positioning, industrialization levels, and regional distinctions—the study provides a multifaceted understanding of the socio-economic consequences of environmental degradation, offering valuable insights for targeted policy interventions and future research directions.