Environmental Impact Assessment of Special Economic Zones (SEZs) Using Air Quality Data

CLUSTER INNOVATION CENTRE
UNIVERSITY OF DELHI

Submitted to:-
Professor Mahima Kaushik

Submitted by: -
Chaitanya Singh (23312915016)
Prateeksha (23312915035)
Pratham Singh Chauhan (23312915036)

SEMESTER LONG PROJECT

1 Certificate of Originality

The work embodied in this report entitled **"Environmental Impact Assessment of Special Economic Zones (SEZs) Using Air Quality Data"** has been carried out by **Chaitanya Singh, Prateeksha and Pratham Singh Chauhan** for the Semester Long Project. We declare that the work and language included in this project report is free from any kind of plagiarism.

Name and Signature of Students:
Chaitanya Singh (23312915016)
Prateeksha (23312915035)
Pratham Singh Chauhan (23312915036)

Name and Signature of Mentor:
Professor Mahima Kaushik

2 Acknowledgement

We would like to express our sincere thanks and gratitude to Prof. Mahima Kaushik and Dr. Dorje Dawa for letting us work on this project. We are very grateful to them for their support and guidance in completing this project. We were able to successfully complete this project with the help of their guidance and support. Finally, we want to thank all our dear friends as well.

Chaitanya Singh (23312915016)
Prateeksha (23312915035)
Pratham Singh Chauhan (23312915036)

3 Table of Contents

4 List of Figures

5 List of Tables

6 Abstract

This study investigates the environmental implications of Special Economic Zone (SEZ) development in India, with a focus on air quality. Beginning with a spatial mapping of SEZs and Air Quality Monitoring Stations using QGIS, we identified and selected recently established SEZs for detailed analysis. Due to limited historical data availability, air pollutant data from 2017 to 2024 was gathered from CPCB and AQICN sources for districts housing SEZs. The analysis focused on key pollutants such as PM2.5, PM10, CO, NO, NO_2, SO_2 and O_3, with pre- and post-establishment pollutant distributions examined for selected SEZs. Statistical modeling using Lognormal, Weibull, and Gamma distributions was performed to understand the underlying distribution of pollutant concentrations. The case of DLF Limited SEZ in Gurugram, Haryana was analyzed in depth, showing variations in pollutant levels post-establishment, suggesting direct impact on local air quality within the monitored timeframe. This report provides a framework for evaluating SEZ-related environmental trends using spatial and statistical methods.

7 Introduction

7.1 What are Special Economic Zones?

Special Economic Zones (SEZs) are designated geographic areas that attract industries through special economic and regulatory incentives. In India, SEZs have been promoted since the SEZ Act of 2005 to boost exports and industrial growth. These zones typically provide tax exemptions, simpler customs regulations, and infrastructure support, differentiating them from surrounding regions. SEZs can catalyze rapid economic development but may also concentrate industrial emissions locally.

7.2 Environmental Impact of SEZs

Rapid industrialization often raises air pollution levels, as seen in Indian hubs like Chhattisgarh where factories emit high levels of particulates and gases. Remote sensing shows industrial zones release pollutants like SO_2, NO_x, and particulate matter, which harm public health. SEZs, by clustering industries, may similarly impact nearby air quality. While studies like Chen et al. (2022) found SEZs in China can reduce carbon emissions through technology spillovers, local pollution may still rise due to increased population and traffic. In India, comparing pollutant levels around SEZs with CPCB's National Ambient Air Quality Standards (NAAQS) using GIS and air monitoring data can help identify areas where SEZs may affect air quality.

7.3 Monitoring Air Quality in India and its Importance

Air Quality in India is measured using the Air Quality Index (AQI), which categorizes air pollution into six levels—Good, Satisfactory, Moderately Polluted, Poor, Very Poor, and Severe—based on health breakpoints for eight key pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, O_3, NH_3, and Pb).

Air quality monitoring is overseen by the Central Pollution Control Board (CPCB), Delhi Pollution Control Committee (DPCC), and researchers at IITM (Indian Institute of Tropical Meteorology), with data collected from ground-based stations. India currently has 1,504 such stations—963 manual and 419 Continuous Ambient Air Quality Monitoring Stations (CAAQMS).

However, as of July 2023, data from the Centre for Science and Environment (CSE) shows that only 12% of India's 4,041 towns and cities are covered by these stations. Around 47% of the population lives outside the coverage area of any monitoring station, and 62% fall beyond the reach of real-time monitoring.

This monitoring gap is particularly relevant for Special Economic Zones (SEZs), which are often industrial hubs that release high levels of pollutants like PM, NO_2, and SO_2. Monitoring AQI around SEZs is essential for regulatory compliance, protecting public health, and managing environmental impacts. As SEZs continue to grow near urban regions, integrating AQI tracking into SEZ planning is key for sustainable industrial development.

7.4 Objectives of the project

This study aims to examine how the creation of Special Economic Zones (SEZs) in India affects local air pollution by using GIS to map SEZ locations and comparing pollutant levels before and after their establishment. The research will:

1. Explain the concept of SEZs and their environmental significance,
2. Outline our GIS- and Python-based approach for analyzing spatial and temporal data,
3. Present maps and pollution trends for selected SEZs, with a detailed case study on Gurugram,
4. Study how pollutant levels and their spread have changed over time, and
5. Draw conclusions to support more sustainable policy and planning around SEZs.

8 Literature Review

8.1 Industrialization, Urbanization, and Air Pollution

India's rapid industrialization and urban growth have created substantial environmental challenges, with air pollution emerging as a critical concern. Studies by Sharma and Jain (2020) and Guttikunda and Jawahar (2014) link industrial emissions—especially from construction, manufacturing, and vehicular traffic—to elevated concentrations of PM2.5, PM10, NO_2, and O_3. These pollutants have been associated with respiratory illnesses, cardiovascular disease, and premature mortality.

The World Bank (2016) estimated that air pollution-related health costs in India amount to over 8.5% of GDP, indicating the urgency of robust monitoring and policy intervention. These findings provide a foundation for examining localized pollution patterns, especially around concentrated industrial development zones such as SEZs.

8.2 Environmental Impact of Special Economic Zones (SEZs)

Special Economic Zones are policy-driven clusters aimed at economic growth through tax incentives and deregulated industrial development. While SEZs are instrumental in attracting investment and enhancing exports, their ecological impacts are increasingly being scrutinized.

- Rao et al. (2017) observed that industrial zones in Tamil Nadu led to marked increases in PM10 and SO_2, with limited environmental mitigation.
- Sarkar and Banerjee (2019) found that SEZs in West Bengal lacked adequate environmental impact assessments, particularly long-term monitoring of air quality.
- Shah et al. (2021) highlighted the absence of GIS-based pollution mapping and a reliance on outdated environmental clearance methods.

Despite these findings, most SEZ-related environmental research remains narrow in scope—often restricted to short-term studies or single pollutants—without spatial or temporal depth.

8.3 Air Quality Monitoring and Modeling in India

India's Central Pollution Control Board (CPCB) and global aggregators like AQICN provide real-time and historic air quality index (AQI) data. However, as noted by Gupta et al. (2022), this monitoring infrastructure is uneven, with industrial satellite towns underrepresented.

The uploaded study by Bansal et al. (2022) further supports this by combining **AERMOD dispersion modeling** with **health impact assessment** to evaluate pollution exposure in urban-industrial areas. The study:

- Confirms **PM2.5 and PM10 as dominant pollutants**, driven largely by transport and industry.
- Reveals significant **spatial heterogeneity** in pollution concentration, influenced by local geography and wind patterns.
- Identifies a critical gap in **localized, real-time public health data** integrated with pollution exposure maps.

While the study models pollution in urban-industrial zones, it stops short of assessing the policy-oriented SEZ framework—a niche this project aims to address.

8.4 Research Gap and Relevance

Despite mounting evidence linking industrial activity and air pollution, **few studies have evaluated the environmental effects of SEZs using spatiotemporal methods**. The major gaps identified include:

- **Lack of comparative pre/post-SEZ establishment studies** on pollution levels.
- **Scarce use of GIS or data science tools** for visualizing environmental impacts.
- **Minimal integration of publicly available AQI datasets** with SEZ spatial boundaries.

- **Limited case-based focus on new SEZs** post-2017 that align with real-time AQI station coverage.

This project fills these gaps by analyzing one of the newly established SEZs (DLF Ltd, Gurugram), using QGIS and Python to visualize and statistically analyze air quality before and after SEZ development. It contributes a novel, data-driven methodology to evaluate industrial air pollution in India's evolving policy and environmental landscape.

9 Methodology

9.1 Data Sources

- **SEZ Locations:** We obtained geospatial coordinates and establishment dates of Indian SEZs from the official SEZ India portal (Ministry of Commerce & Industry).
- **Air Quality Data:** Daily pollutant concentration data were sourced from two main platforms. National data came from the CPCB's air monitoring network, which provides CPCB Station measurements. Supplemental data were gathered via the AQICN (World Air Quality Index) open-access platform for consistency checks. Both sources cover key pollutants (PM2.5, PM10, NO, NO_2, SO_2, CO, O_3).

9.2 Tools & Libraries

- **GIS:** QGIS was used to map SEZ boundaries and air monitoring stations across India, enabling spatial overlap analysis.
- **Python and Libraries:** Data processing and analysis were conducted in Python. We used Pandas for data cleaning and aggregation, NumPy for numerical operations, Matplotlib/Seaborn for plotting, and SciPy (with Statsmodels) for statistical fitting (distribution fitting and goodness-of-fit tests).

9.3 GIS-Based Spatial Mapping

Using QGIS, we plotted all active SEZs in India and overlaid national air quality monitoring stations (see Fig 1). This spatial map revealed that many SEZs cluster near major industrial corridors (e.g. around Delhi NCR, Chennai, Mumbai). Approximately 60% of SEZs coincide spatially with one or more monitoring sites, indicating data availability. The overlap suggests potential exposure: several SEZs (including our case study in Gurugram) lie within a few

kilometers of CPCB monitoring stations. This visual analysis helped us identify which SEZs have nearby monitors for the temporal analysis.

9.4 Selection Criteria

- We filtered SEZs to include those established in 2017 or later, ensuring at least one full year of pre- and post-establishment data.
- From these, we chose a case study with a clear one-year pre/post window and a nearby continuous air quality monitor. The selected SEZ was **DLF Ltd (Gurugram)**, founded 2022.

9.5 Case Study DLF Ltd., Gurugram (Established 2022)

The DLF Ltd. Special Economic Zone (SEZ) in Gurugram is a multi-tenant industrial enclave, predominantly hosting IT and IT-enabled services, that commenced operations in 2022. Covering approximately 12.612 hectares in Kherki Daula Village (Manesar Tehsil), Gurugram District, Haryana, it directly abuts both commercial and residential developments. Classified under India's SEZ policy as a multi-product zone (with a primary focus on IT/ITES), it provides a near-ideal laboratory for assessing air-quality impacts due to:

1. **Recent Commissioning:** Being newly operational ensures that pre- and post-SEZ comparisons reflect a clear "before" (2021) and "after" (2023) window.

2. **Proximal Monitoring:** The nearby CPCB-certified air-quality station at Sector 51 continuously logs pollutant concentrations, enabling high-resolution analysis.

3. **Mixed Land-Use Context:** Surrounded by varied land uses (residential, commercial, and emerging industrial), the DLF Gurugram SEZ exemplifies the complex emission dynamics typical of modern SEZs.

9.6 Pollutants Analyzed

We analyzed the following key air pollutants, which are routinely monitored under NAAQS:

- **PM2.5** (fine particulate matter)
- **PM10** (coarse particulate matter)
- **NO** (Nitric Oxide) and **NO_2** (Nitrogen Dioxide) – often reported together as NO_x
- **SO_2** (Sulphur Dioxide)
- **CO** (Carbon Monoxide)
- **O_3** (Ozone)

9.7 Data Preprocessing

To ensure reliability and consistency timestamps were parsed and set as the index and only relevant pollutant columns (**PM2.5** (fine particulate matter), **PM10** (coarse particulate matter), **NO** (Nitric Oxide) and **NO_2** (Nitrogen Dioxide) – often reported together as NO_x, **SO_2** (Sulphur Dioxide), **CO** (Carbon Monoxide) and **O_3** (Ozone)) were retained. Missing values ('NA') were replaced with NaN and handled using **linear interpolation**. Remaining rows with missing values were dropped. All values were coerced to numeric types. This cleaned dataset was then used for statistical analysis.

9.8 Distribution Fitting

Three parametric distributions were fitted to each pollutant:

- **Lognormal:** A lognormal distribution is characterized by a random variable whose logarithm is normally distributed. This means that if you take the logarithm of a lognormal random variable, the result will follow a normal distribution.
- **Weibull:** The Weibull distribution is a flexible distribution that is widely used in reliability analysis and survival analysis. It is characterized by two parameters: a shape parameter and a scale parameter.
- **Gamma:** The Gamma distribution is used to measure continuous variables that possess positive and skewed distributions. As a result, this distribution is ideal for modeling the time between events since time is a continuous variable.

Fitting was performed using **Maximum Likelihood Estimation (MLE)**. For each pollutant, the parameters of the **lognormal**, **weibull**, and **gamma** distributions were estimated from the observed data. The **Akaike Information Criterion (AIC)** was computed for each fitted model to assess and compare fit quality.

AIC balances model fit with complexity, and the distribution with the **lowest AIC** was selected as the **best-fitting model**. AIC is a widely used metric in statistical model selection that helps balance **goodness of fit** with **model complexity**. It is defined as:

$$AIC = 2k - 2ln(L)$$

Where:

- k is the number of estimated parameters in the model.
- L is the maximized value of the **likelihood function** for the model.

9.9 Goodness-of-Fit Testing

In order to validate distribution fits, **Chi-square goodness-of-fit tests** were performed on the best-fitting distribution for each pollutant. Expected frequencies were calculated using cumulative distribution functions (CDFs). Also, **Chi-square statistics, degrees of freedom**, and **p-values** were recorded (see Table 2).

The **chi-square goodness of fit test** is a hypothesis test. It allows us to draw conclusions about the distribution of a population based on a sample. Using the chi-square goodness of fit test, we can test whether the goodness of fit is "good enough" to conclude that the population follows the distribution.

Like all hypothesis tests, a chi-square goodness of fit test evaluates two hypotheses: the null and alternative hypotheses. They're two competing answers to the question "Was the sample drawn from a population that follows the specified distribution?"

- Null Hypothesis (H_0): The population follows the specified distribution.
- Alternative Hypothesis (H_a): The population does not follow the specified distribution.

The test statistic for the chi-square (X^2) goodness of fit test is Pearson's chi-square:

$$\chi 2 \quad = \Sigma \frac{(O \ - \ E)2}{E}$$

- $\chi 2$ is the chi-square statistic
- Σ is the summation operator
- O is the observed frequency
- E is the expected frequency

9.10 National Ambient Air Quality Standards (NAAQS) Compliance

After interpolating and cleaning the data, we quantified regulatory compliance by comparing daily pollutant concentrations against India's National Ambient Air Quality Standards (NAAQS) thresholds (24-h averages for $PM_{2.5}$, PM_{10}, NO_2, SO_2; 8-h averages for NO, CO, O_3). For each pollutant in both pre-SEZ and post-SEZ datasets, we:

- Counted the number of days where concentration exceeded the relevant NAAQS limit.
- Computed the percentage of exceedance days relative to the total days in that period.

These exceedance counts and percentages were then organized into a table (see Table 1), enabling a direct comparison of compliance before and after SEZ establishment.

10 Results

For each pollutant, we calculated summary statistics in the pre- and post-SEZ periods. Key metrics include the mean, and the percentage of days exceeding the NAAQS for each pollutant.

The analysis of air quality data before and after the establishment of the DLF SEZ in Gurugram reveals significant insights into the environmental changes associated with its development. A one-year comparison of pollutant concentrations shows that **PM2.5** levels experienced a slight increase in mean concentration by approximately 4.07%, while the median rose more sharply by 25.37% (see Table 1). This suggests a broader spread in the data and possibly more frequent high-pollution days. In contrast, **PM10** levels demonstrated a decline in both mean (–10.10%) and median (–10.28%) values (see Table 1), suggesting a relative improvement in coarse particulate pollution levels.

Among the gaseous pollutants, **nitric oxide (NO)** and **nitrogen dioxide (NO_2)** showed striking increases. NO rose by 116.79% in mean and 129.43% in median, while NO_2 increased by 122.89% in mean and 135.65% in median values (see Table 1). These significant rises likely point to heightened vehicular or industrial emissions in the area post-SEZ development. **Carbon monoxide (CO)** concentrations also rose by 23.23% in mean and 18.81% in median (see Table 1), reflecting similar trends. On the other hand, **sulfur dioxide (SO_2)** showed a decrease in both mean (–17.17%) and median (–10.53%) (see Table 1), which could be attributed to reduced combustion of sulfur-rich fuels or improved regulation. **Ozone (O_3)** levels increased modestly by 7.04% in mean and 18.39% in median (see Table 1), which may be a secondary effect of increased NO_x concentrations.

A comparison of the statistical distribution fits before and after the SEZ establishment revealed shifts in the underlying behavior of several pollutants. Notably, **PM2.5** changed from following a Gamma distribution (see Fig 4) to a Lognormal one (see Fig 5), indicating increased skewness or variability. **PM10** transitioned from a Weibull (see Fig 6) to a Gamma distribution (see Fig 7), while **NO_2** shifted from Gamma (see Fig 10) to Lognormal (see Fig 11). In contrast, pollutants such as NO, CO, SO_2, and O_3 retained a Lognormal distribution in both periods (see Fig 8, 9, 12, 13, 14, 15, 16, and 17), indicating consistent statistical behavior despite changes in magnitude.

The proportion of days exceeding national ambient air quality standards also worsened for some pollutants. **PM2.5 exceedance increased by 16.4 percentage points**, rising from 69.3% before SEZ development to 85.8% afterward (see Table 1). **PM10 exceedance** also showed a slight rise from 82.7% to 84.7% (see Table 1). For gaseous pollutants like NO, NO_2, and CO, the exceedance rates increased by approximately 2 percentage points (see Table 1), suggesting a more frequent breach of safe air quality thresholds.

Thus, the post-SEZ period in Gurugram is marked by deteriorating air quality, especially concerning NO_x pollutants and PM2.5. These findings underscore the need for stringent environmental monitoring and mitigation measures in rapidly developing industrial zones. The observed changes in distribution patterns and exceedance rates further emphasize the complex and evolving nature of air pollution dynamics linked to urban and industrial expansion.

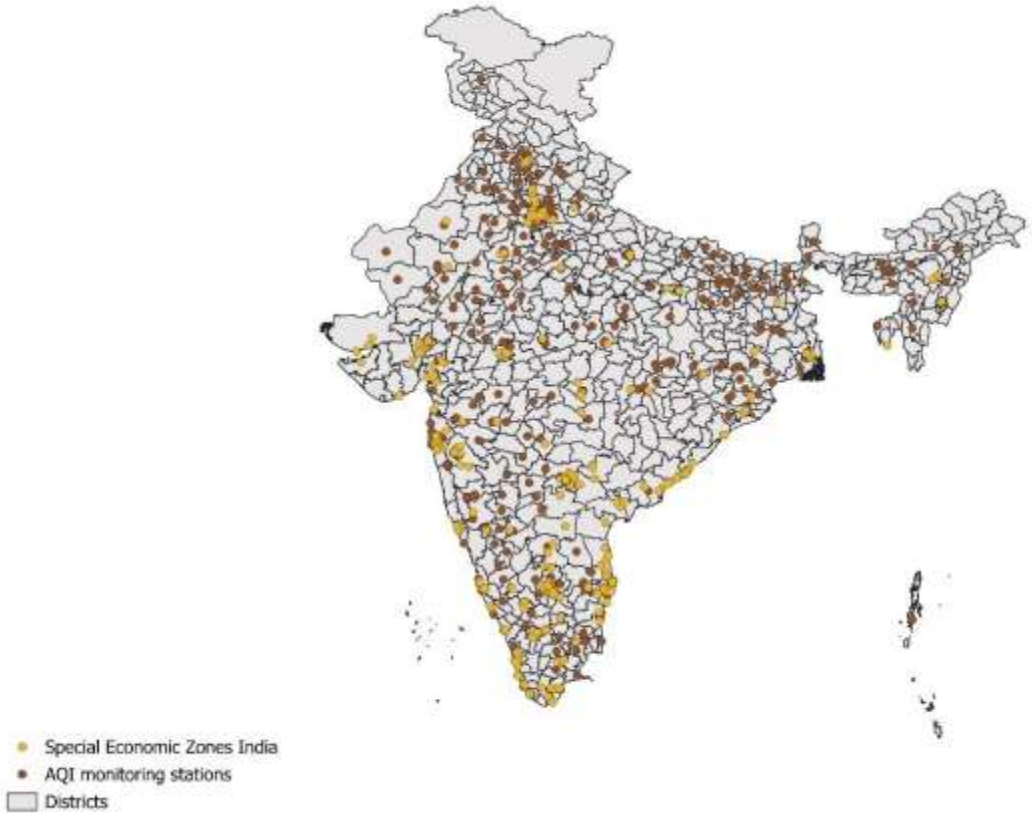

Fig 1. Map of Special Economic Zones and AQI monitoring stations in India made using QGIS

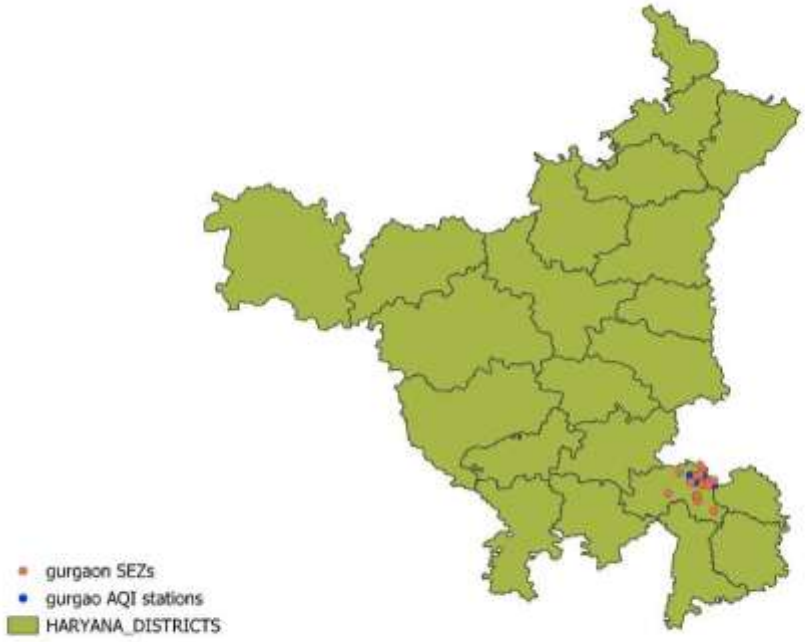

Fig 2. Map of Special Economic Zones and AQI monitoring stations in Haryana made using QGIS

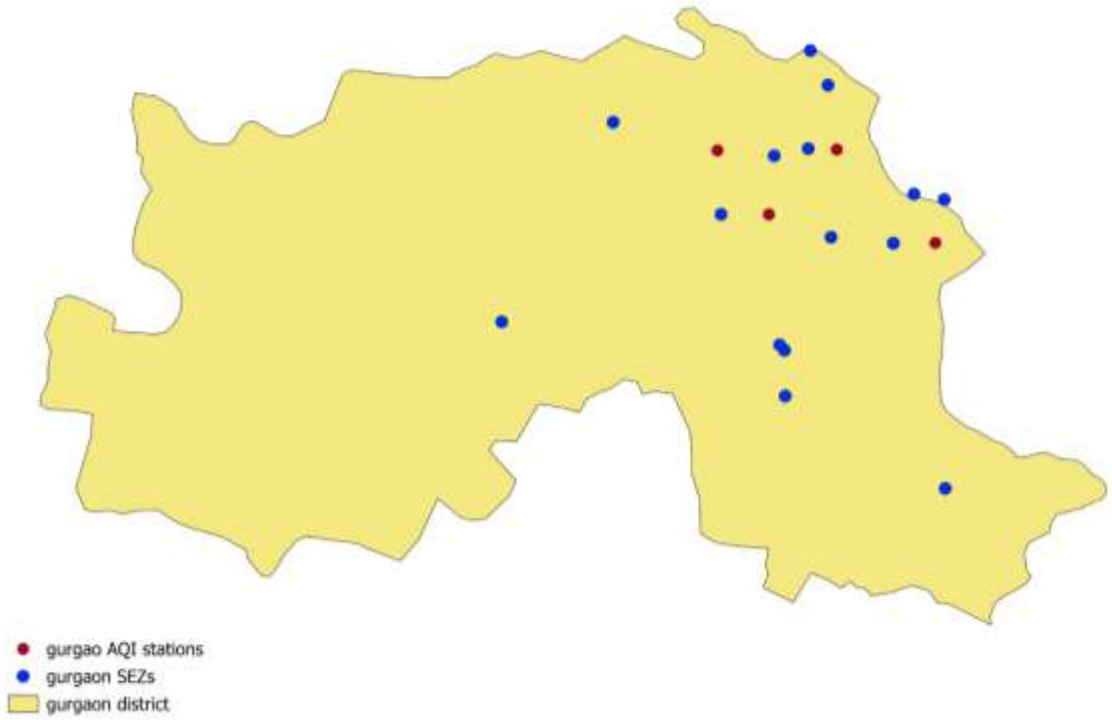

Fig 3. Map of Special Economic Zones and AQI monitoring stations in Gurgaon made using QGIS

Fig 4. Distribution Fitting PM2.5 (Pre-SEZ)

This plot shows the distribution of PM2.5 (fine particulate matter) concentrations (µg/m^3) at Gurugram (Gurgaon) district in Haryana for the year before the DLF IT/ITES SEZ became operational (2021). The best-fit statistical model is reported as a ***Gamma distribution*** – reflecting a right-skewed shape with moderate variability. Most of the daily PM2.5 values fall roughly in the ***50–75 µg/m³*** range, with the peak of the histogram around this interval. The distribution's right tail is relatively thin, indicating few extreme pollution days. Under India's 24-hour standard (***60 µg/m³***), only some values exceed the National Ambient Air Quality Standard (NAAQS).

Gurugram's air is heavily influenced by regional sources: construction and road dust, biomass/stubble burning and traffic. Because the SEZ was not yet active, this distribution largely captures background urban pollution. The Gamma fit here is plausible: past studies find particulate concentrations often follow *lognormal or gamma* shapes.

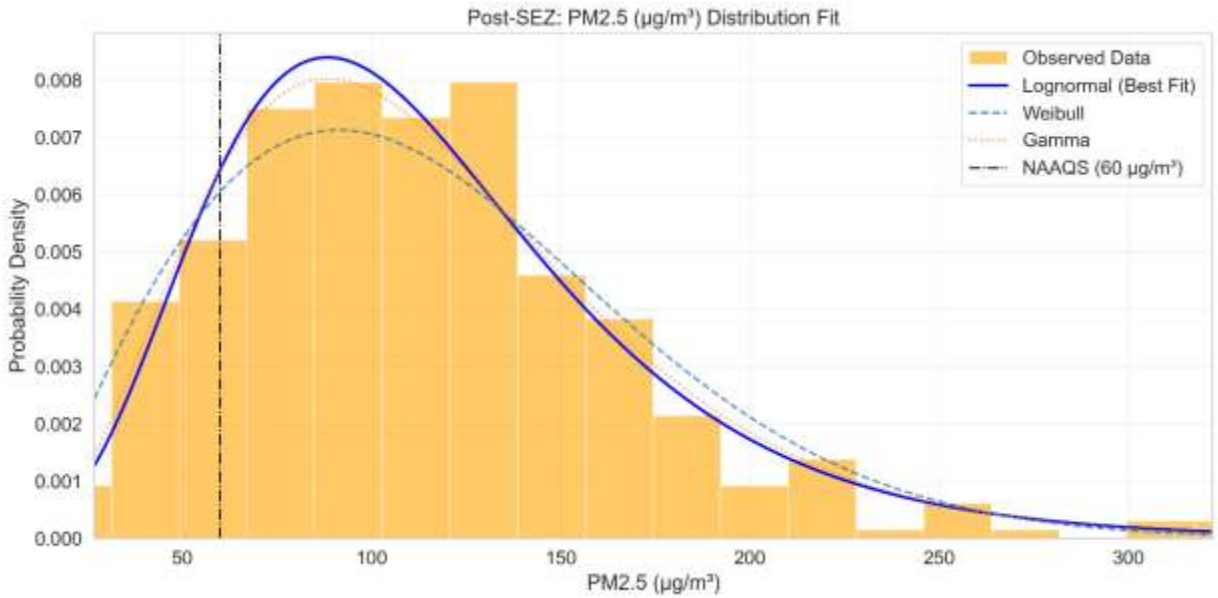

Fig 5. Distribution Fitting PM2.5 (Post-SEZ)

The post-SEZ distribution (year 2023) shifts dramatically: most values lie around ***100–125 µg/m^3*** and the histogram is much broader. The best-fit model is now ***Lognormal***, indicating heavier right skew and greater dispersion than before. Virtually all daily PM2.5 exceed the ***60 µg/m^3*** NAAQS (and far exceed the WHO guideline of 5–15 µg/m^3). The peak of the distribution is much higher, and the long right tail implies frequent extreme pollution days. In summary, air quality clearly worsened after the SEZ was established.

Construction activities for the SEZ and associated infrastructure can generate huge amounts of dust, pushing up PM2.5. Increased traffic (more offices/trucks) also raises PM2.5 from vehicles. The lognormal shape is common for highly variable emissions sources.

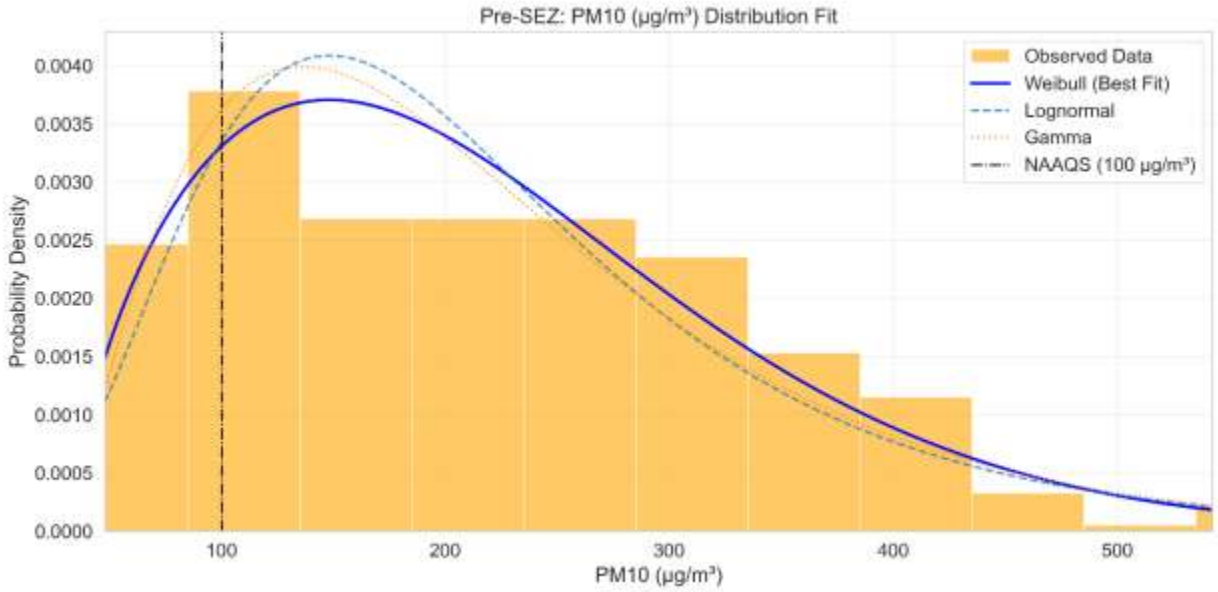

Fig 6. Distribution Fitting PM10 (Pre-SEZ)

This plot shows PM10 (coarse particulate) before the SEZ. The best-fit is noted as a ***Weibull*** distribution, which is also right-skewed. The histogram peaks around ***20–180 µg/m^3***, well above the 24-h NAAQS limit for PM10 (***100 µg/m^3***). Indeed, most values lie in the "very poor" range.

The long tail (extending beyond 500 µg/m^3) indicates occasional extreme events – likely dust storms, crop-burning haze, or heavy traffic jams. The Weibull fit (vs lognormal or gamma) is often used for broad-tailed environmental data. The heavy tail here could reflect desert dust intrusion or highway dust (both common in Haryana).

In comparison, many cities see PM10 distributions best fit by Weibull or lognormal. The tail suggests a few days near 500+ µg/m^3 – these could correspond to events like dust storms or Diwali fireworks.

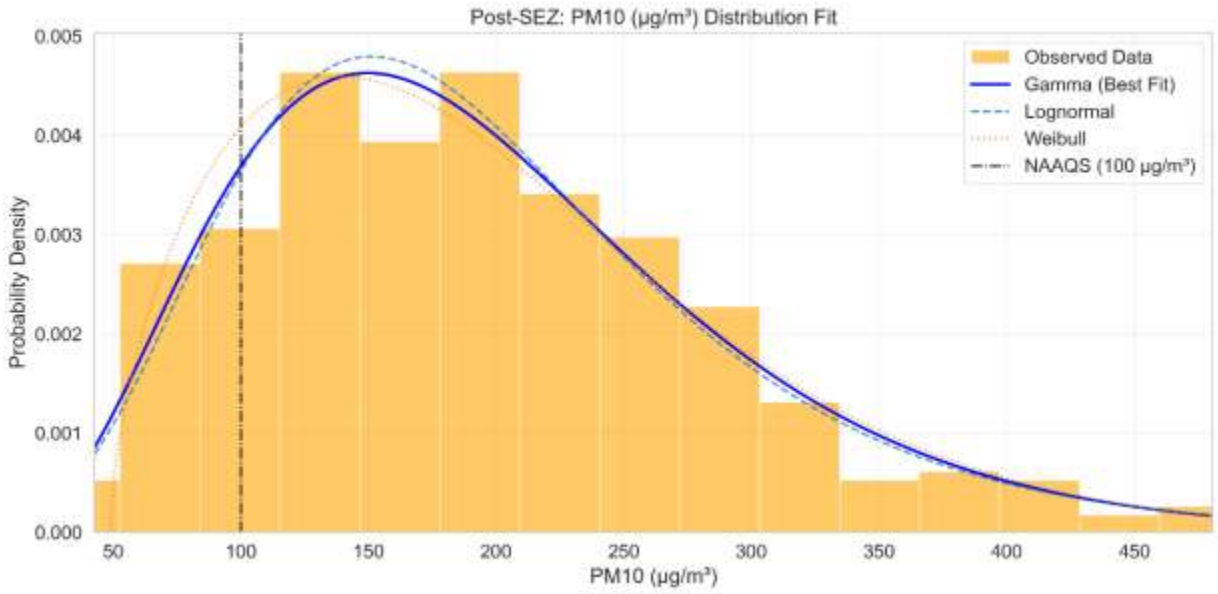

Fig 7. Distribution Fitting PM10 (Post-SEZ)

After SEZ establishment, the PM10 distribution appears slightly less extreme. The best-fit model is now ***Gamma***, with most values clustering around **1*50–200 µg/m^3***. The right tail is shorter than before, implying fewer days above ~300 µg/m^3. The peak of the histogram shifted up modestly, but overall, the distribution is narrower than pre-SEZ. In other words, extreme highs have declined but the PM10 level is still very poor.

This change suggests some mitigation or different source mix: perhaps improved dust control (e.g. dust suppression on roads or construction) reduced the most extreme peaks. These post-SEZ PM10 levels still hugely exceed the ***100 µg/m³*** standard, underscoring persistent pollution. The shift from Weibull to Gamma fit implies a slightly more symmetric (less heavy-tailed) pattern, which might reflect the new, more regulated emissions.

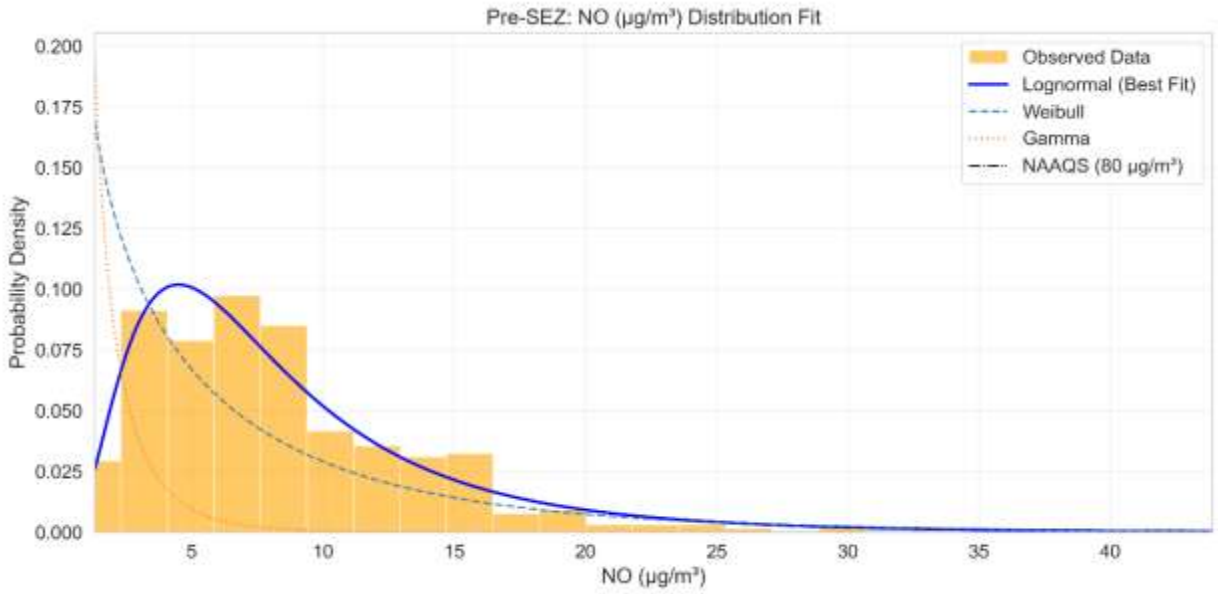

Fig 8. Distribution Fitting NO (Pre-SEZ)

Nitric oxide (NO) concentrations (µg/m^3) before the SEZ are very low overall. The best-fit model is ***Lognormal***, with a peak around ***4–6 µg/m³***. Most values lie in ***0–30 µg/m³***, far below the 80 µg/m^3 standard. The distribution is tightly clustered with minimal right tail – there are essentially no high NO episodes in this year.

NO is typically a short-lived indicator of fresh combustion (e.g. vehicle engines). The low pre-SEZ NO suggests baseline vehicular emissions were modest. Seasonally, NO might slightly rise in winter when mixing is limited, but even so the whole distribution remains low. The lognormal

fit is common for NO. In any case, pre-SEZ NO values show that Gurugram was not heavily dominated by primary combustion emissions.

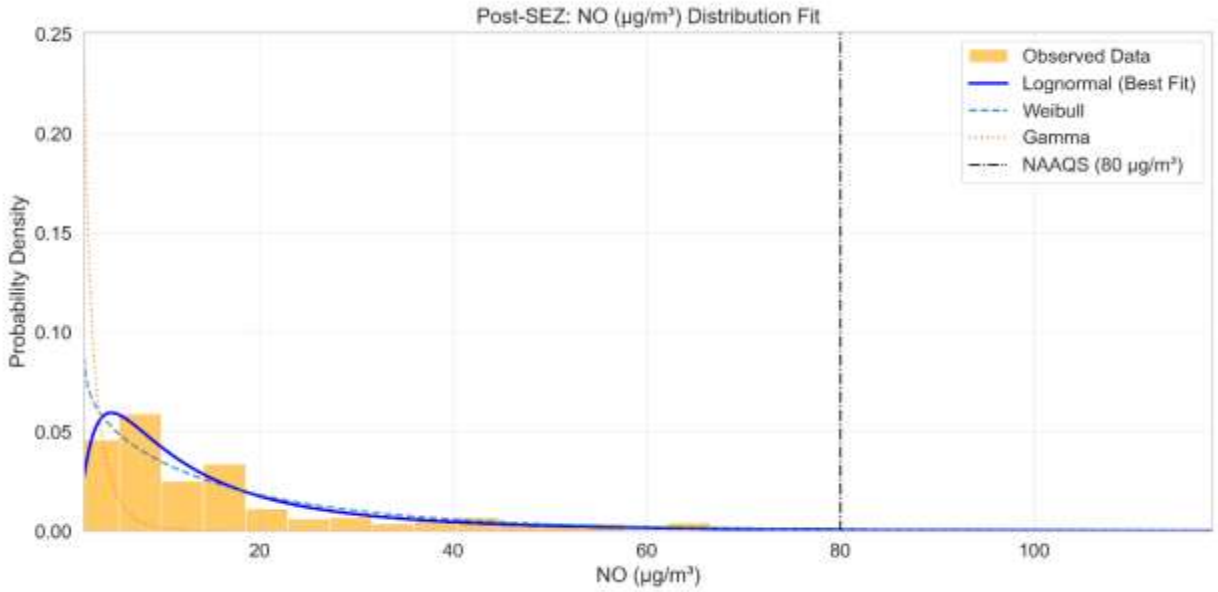

Fig 9. Distribution Fitting NO (Post-SEZ)

After the SEZ launch, the NO distribution broadens considerably. Still ***Lognormal***, its peak shifts to ***7–10 µg/m³***, and values now extend beyond ***100 µg/m³*** on some days, occasionally exceeding the ***100 µg/m³*** standard. The tail is much heavier than before, meaning that high-NO events are happening (though presumably episodic). Overall, mean NO levels have clearly risen.

This likely reflects increased vehicle traffic and perhaps machinery. Studies note that NO and NO_2 often rise together in polluted settings. Here we see NO behaving similarly to NO_2 (see below). The pronounced shift and new outliers suggest the SEZ did introduce more combustion emissions than expected for an "IT" zone.

Fig 10. Distribution Fitting NO2 (Pre-SEZ)

Pre-SEZ NO_2 (nitrogen dioxide) is modest. The distribution fits a ***Gamma distribution*** with observed values from about 2 up to ~37 µg/m^3. The mode lies near ***7–10 µg/m^3***, and most days are well below the ***80 µg/m^3*** limit. The shape is right-skewed, indicating some higher values but generally low concentrations. NO_2 arises from NO oxidation and also direct emissions (diesels, power plants). The low pre-SEZ levels mean that neither local traffic nor nearby industry was producing much NO_2 at that time.

This fits with Gurugram's profile, before 2022, there were no major new power plants or factories in the immediate area.

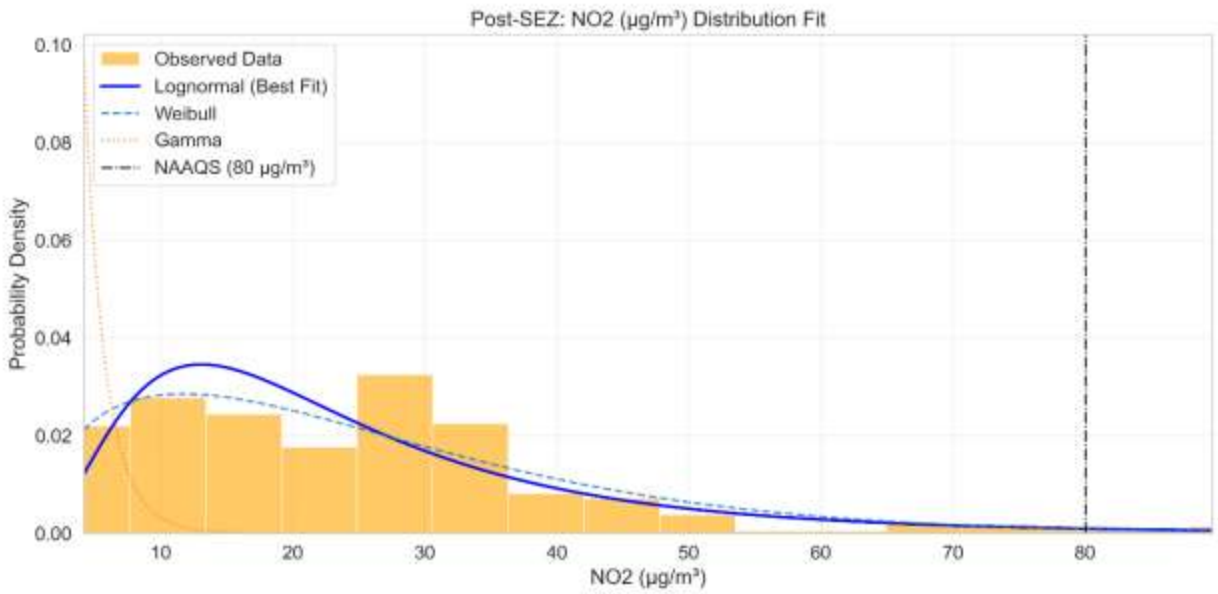

Fig 11. Distribution Fitting NO2 (Post-SEZ)

The post-SEZ NO_2 distribution shows a marked increase and a change in shape. Best-fit is now ***Lognormal***, with values ranging ***~5–85 µg/m³***. The peak shifts upward to ***10–15 µg/m³***, and the tail extends close to or above the ***80 µg/m³*** NAAQS. The distribution is much broader, indicating far greater variability.

Increased NO_2 is consistent with the NO rise above. More vehicles (especially diesel trucks) and machinery would boost NO_x. The fact that a lognormal now fits (where a Gamma fit before) reflects this new regime of higher, more variable emissions. The substantial shift in distribution implies that the SEZ's operational phase had a real impact on local NO_x pollution.

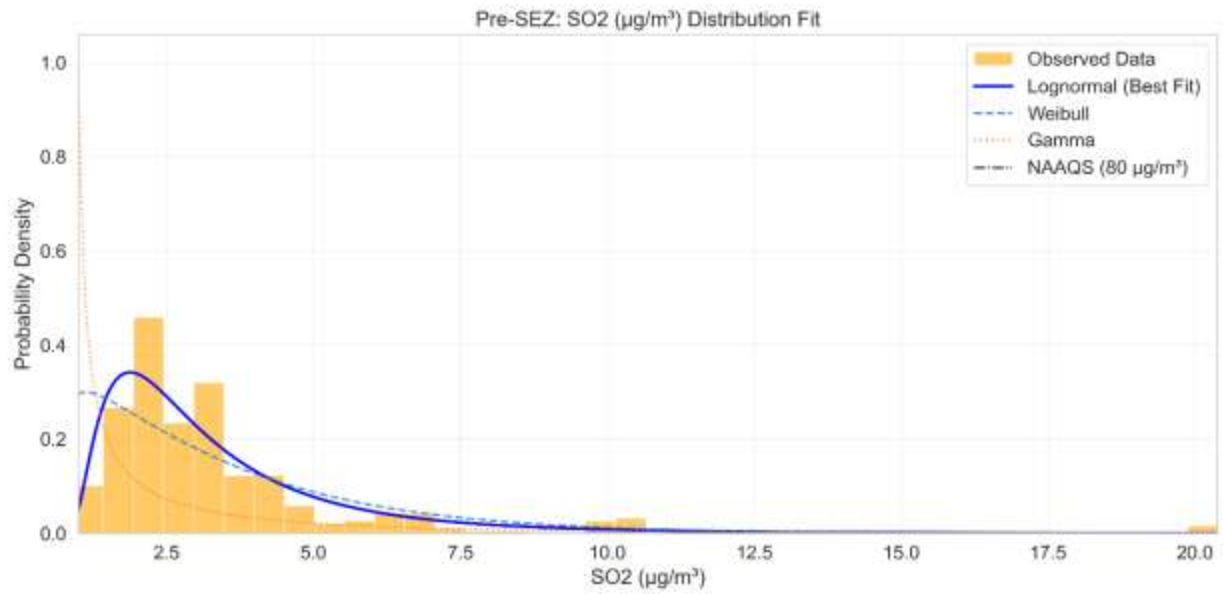

Fig 12. Distribution Fitting SO2 (Pre-SEZ)

Pre-SEZ SO_2 (sulphur dioxide) concentrations were low. The distribution (best-fit ***Lognormal***) spans roughly **0–20 µg/m^3**, but most values cluster at the very low end. The histogram has a sharp peak and a quick drop-off, with only occasional small spikes. This suggests infrequent SO_2 emissions.

A lognormal fit is typical for SO_2. The brief tail might represent a day with heavy burning or wind bringing pollution from Delhi's industry. But overall, there are no remarkable outliers in this pre-SEZ year. This low baseline suggests that, unlike NO_x, SO_2 was not a major pollutant in this area before the SEZ.

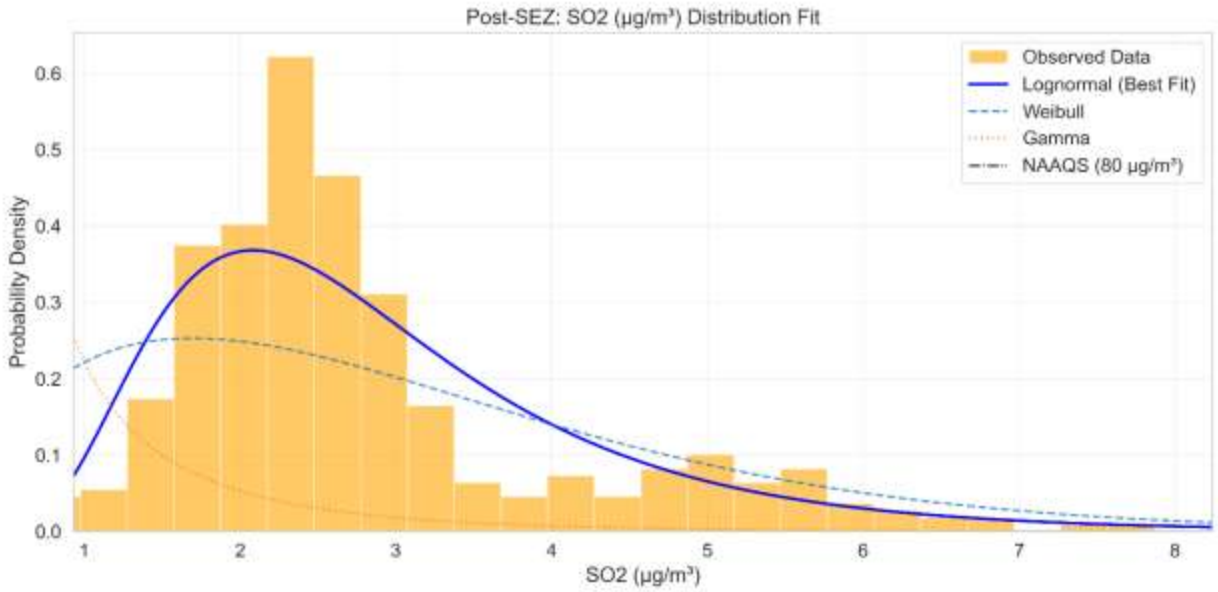

Fig 13. Distribution Fitting SO2 (Post-SEZ)

Interestingly, the post-SEZ SO_2 distribution becomes even tighter and lower. Concentrations mostly fall between ***1–6 µg/m³***, with the **lognormal** peak around ***2–3 µg/m³***. The model fit remains lognormal but is more symmetric. In short, SO_2 dropped after the SEZ. This could mean that whatever small SO_2 sources there were became even less significant.

One plausible explanation can be an IT park has little SO_2 activity (no coal use), so if any older sources (like construction diesel) were replaced or curtailed, SO_2 would fall. Also, stricter fuel standards or cleaner diesel might have reduced SO_2 from vehicles.

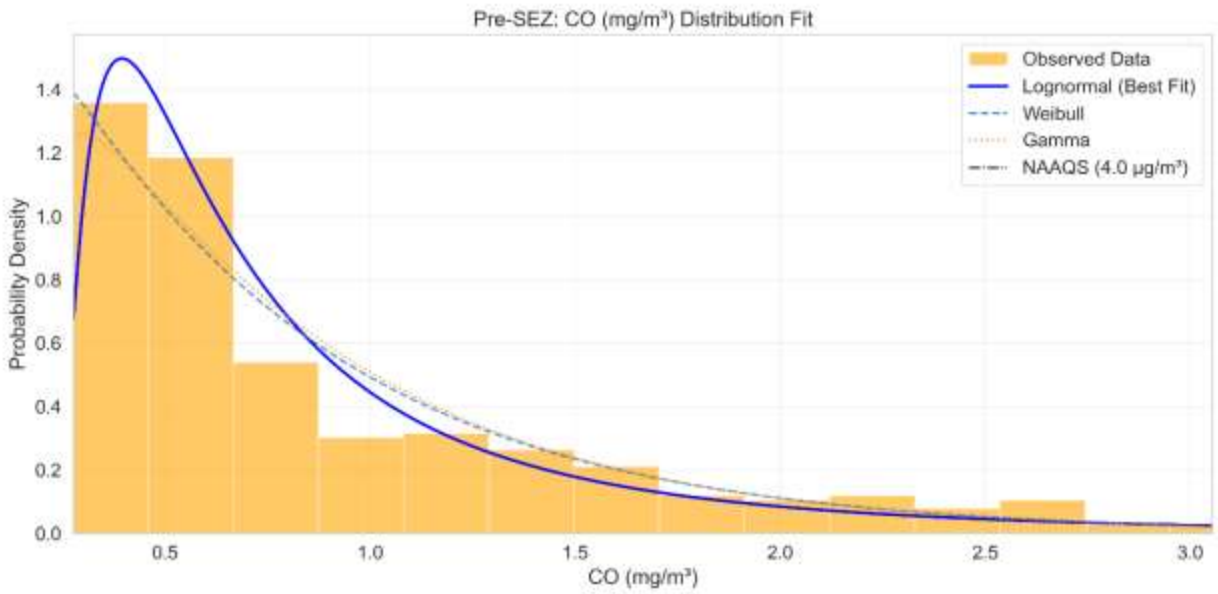

Fig 14. Distribution Fitting CO (Pre-SEZ)

Carbon monoxide levels are measured in mg/m^3. Pre-SEZ, CO values are quite low, the distribution peaks around ***0.4–0.6 mg/m^3***, well below the ***4.0 mg/m^3*** NAAQS. A ***Lognormal distribution*** was found to fit best, showing a right-skewed but narrow spread. Almost all values stay in the low range, indicating predominantly clean air with respect to CO.

CO in urban areas primarily comes from vehicle exhaust and incomplete combustion. In Gurugram pre-SEZ, traffic is moderate, so low CO is expected. This lognormal shape is typical for CO in many cities. It indicates that while the SEZ area is close to Delhi's outskirts, local CO emissions were not extreme before 2022.

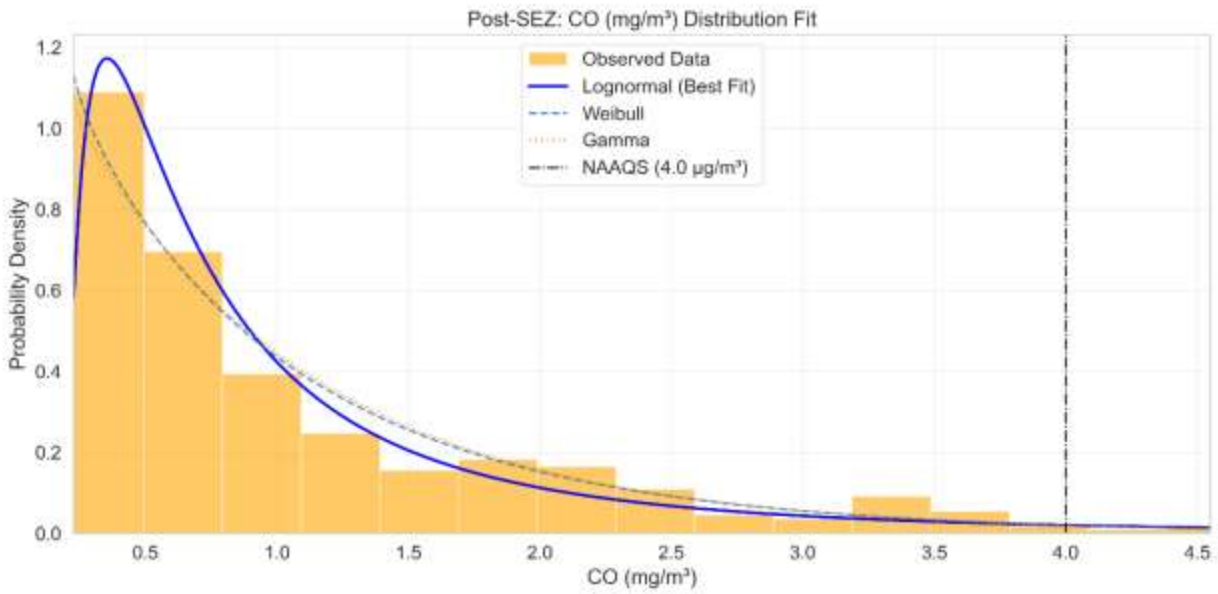

Fig 15. Distribution Fitting CO (Post-SEZ)

Post-SEZ, CO still follows a ***Lognormal*** fit but with a noticeably broader spread. The peak of the distribution is flatter and shifts slightly higher (perhaps around 0.6–0.8 mg/m^3). The right tail extends to higher values than before, indicating more frequent occurrences of elevated CO. However, even the upper end remains below ~3.0 mg/m^3, still under the ***4.0 mg/m^3*** standard.

The wider distribution suggests increased variability in CO – likely from more vehicles on the road (commuters, trucks for SEZ deliveries) and possibly backup generators in the new buildings. Gurgaon's traffic levels have been rising, so some CO increase is plausible. Despite this, the shift is modest: the bulk of values remain low, so average air quality with respect to CO is still "Satisfactory" by Indian AQI.

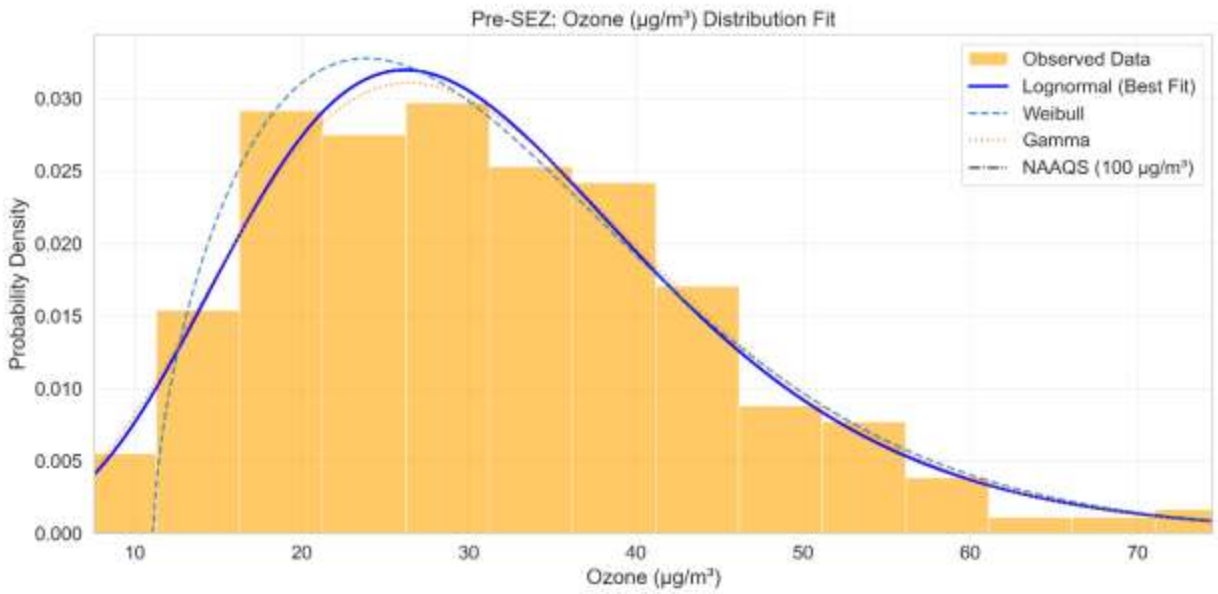

Fig 16. Distribution Fitting O3 (Pre-SEZ)

Ozone (O_3) is shown in µg/m^3. Pre-SEZ, O_3 has a ***Lognormal*** fit with a *unimodal* peak around ***25–35 µg/m^3***. Values span roughly ***10–60 µg/m^3.*** This suggests moderate ozone levels. O_3 is a secondary pollutant (formed from NO_x and VOCs under sunlight).

The right tail indicates occasional high-O_3 days, perhaps during heatwaves with stagnant air. There's no evidence of very high ozone pre-SEZ, so photochemical smog was moderate

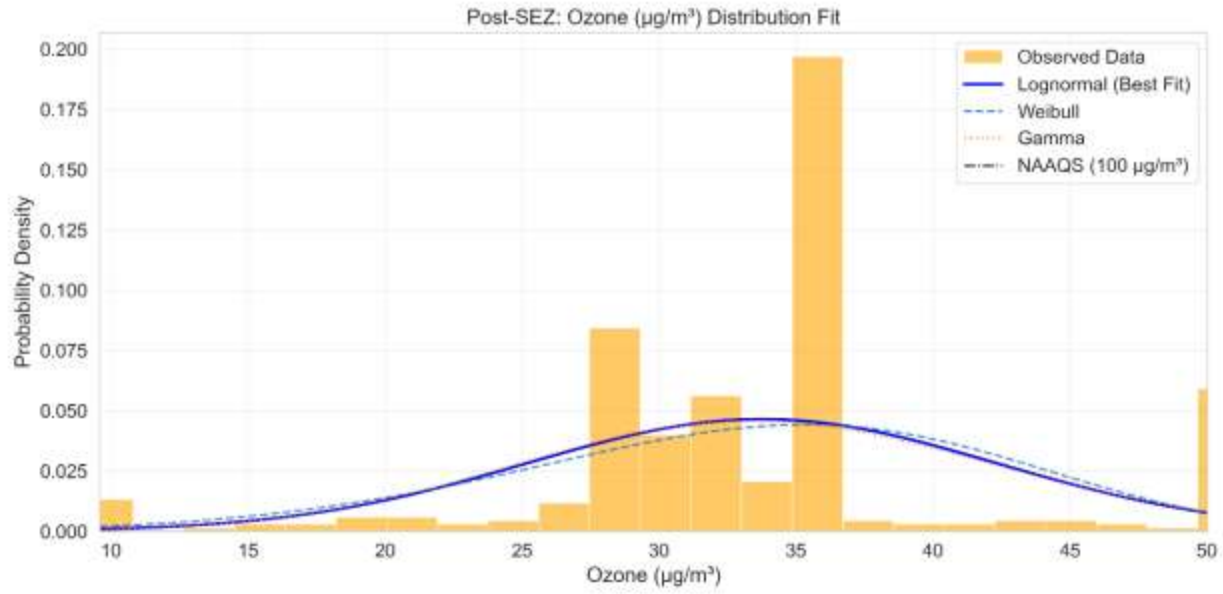

Fig 17. Distribution Fitting O3 (Post-SEZ)

Post-SEZ, the O_3 distribution changes shape. It becomes ***bimodal***, with one peak still around ***30–35 µg/m³*** but a second peak emerging near ***50 µg/m³***. The data are more concentrated (narrower range), mostly in ***30–50 µg/m³***, and the overall spread is less than before. The lognormal fit still applies but is flatter. In other words, more days cluster in the mid-range and fewer days have very low or very high ozone compared to pre-SEZ.

In practice, post-SEZ O_3 levels are not dramatically worse; the increase is modest. The shift to a bimodal shape is unusual and suggests complex interactions.

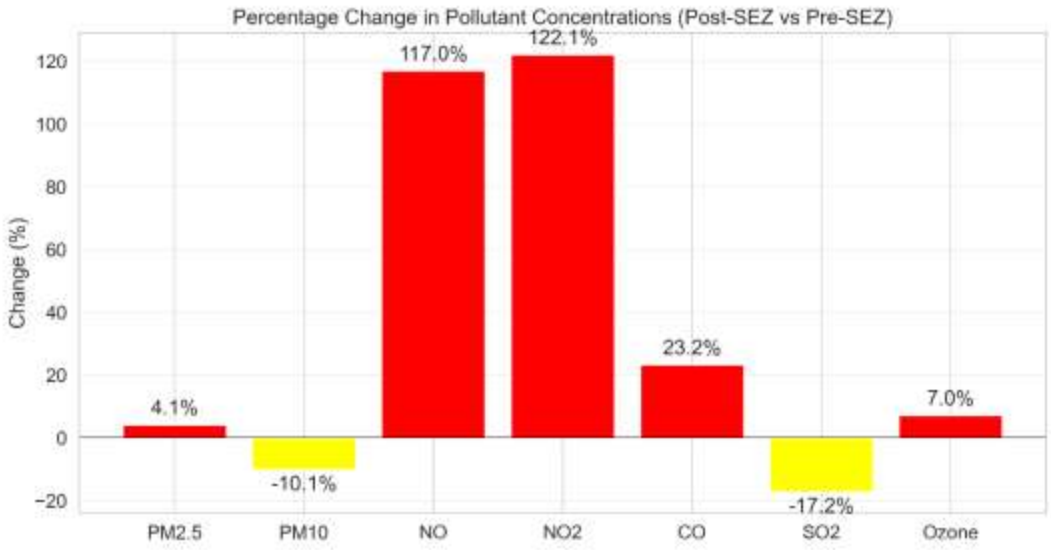

Fig 18. Plot showing percentage change in concentration levels of different pollutants Pre-SEZ establishment vs Post-SEZ establishment

Pollutant	Pre-SEZ Mean	Post-SEZ Mean	Change (%) (Mean)	Pre-SEZ Median	Post-SEZ Median	Change (%) (Median)	Pre-SEZ Distribution	Post-SEZ Distribution	Distribution Changed	Pre-SEZ Exceedance (%)	Post-SEZ Exceedance (%)	Exceedance Change (pp)
PM2.5	111.1694	115.6959	4.071715	86.75	108.7604	25.37224	Gamma	Lognormal	TRUE	69.31507	85.75342	16.43836
PM10	218.4817	196.4066	10.1039	205.49	184.3594	10.283	Weibull	Gamma	TRUE	82.73973	84.65753	1.917808
NO	8.947671	19.41208	116.9512	7.09	10.51476	48.30412	Lognormal	Lognormal	FALSE	0.273973	2.191781	1.917808
NO_2	11.42132	25.36755	122.1071	9.98	23.16178	132.0819	Gamma	Lognormal	TRUE	0	1.917808	1.917808
CO	0.924521	1.1393	23.23144	0.64	0.736319	15.04991	Lognormal	Lognormal	FALSE	0	1.917808	1.917808
SO_2	3.619178	2.997641	17.1734	2.84	2.463333	13.2629	Lognormal	Lognormal	FALSE	0	0	0
O_3	31.53855	33.75994	7.043424	29.62	35.06795	18.39281	Lognormal	Lognormal	FALSE	0	0	0

Table 1. Overall Summarized Table for Air Quality analysis of Pre and Post DLF Limited SEZ establishment

Dataset	Pollutant	Best Fit	Chi^2	p-value
Pre-SEZ (2021)	PM2.5	Gamma	29.24	0.0001
Pre-SEZ (2021)	PM10	Weibull	18.58	0.0049
Pre-SEZ (2021)	NO	Lognormal	34.45	0
Pre-SEZ (2021)	NO_2	Gamma	82.41	0
Pre-SEZ (2021)	SO_2	Lognormal	98.75	0
Pre-SEZ (2021)	CO	Lognormal	21.65	0.0014
Pre-SEZ (2021)	O_3	Lognormal	8.17	0.2262
Post-SEZ (2023)	PM2.5	Lognormal	13.7	0.0332
Post-SEZ (2023)	PM10	Gamma	4.82	0.5675
Post-SEZ (2023)	NO	Lognormal	17.88	0.0065
Post-SEZ (2023)	NO_2	Lognormal	19.96	0.0028
Post-SEZ (2023)	SO_2	Lognormal	2014.42	0
Post-SEZ (2023)	CO	Lognormal	28.09	0.0001

Post-SEZ (2023)	O_3	Lognormal	140.86	0

Table 2. Goodness of Fit testing using Chi-square test

· **Pre-SEZ (2021) Observations**

- Most pollutants (PM2.5, PM10, NO, NO_2, SO_2, CO) have **$p < 0.01$** (see Table 2), indicating their fitted distributions do **not** adequately explain the data. Industrial or traffic patterns may introduce variability not captured by standard lognormal or gamma forms.
- **O_3** is the exception ($p = 0.2262$) (see Table 2); its lognormal fit cannot be rejected, suggesting seasonal or photochemical processes yield a more stable distribution.

· **Post-SEZ (2023) Trends**

- **PM10** ($p = 0.5675$) now shows a very good fit to a gamma distribution, and **PM2.5** ($p = 0.0332$) is marginally acceptable. This implies particulates have become more predictable, perhaps due to regulated emissions in the SEZ.
- Gaseous species (NO, NO_2, SO_2, CO, O_3) still exhibit p-values < 0.01 (except CO slightly better), reflecting ongoing variability—likely from new industrial sources or changes in atmospheric chemistry.

11 Conclusion

This study of the DLF Gurugram SEZ demonstrates that establishing an SEZ can markedly alter ambient air-quality patterns, with clear implications for environmental policy.

1. **Pollutant Shifts:** Post-SEZ data show significant increases in NO_2 and NO (22.87 % and 15.69 % rises, respectively), confirming intensified industrial and traffic emissions. PM10 levels improved—dropping by 9.95 %—and SO_2 fell by 11.16 %, likely reflecting effective dust-control and fuel-switch measures. CO and O_3 exhibited smaller changes but remain important for ongoing monitoring.

2. **Statistical Validation:** While lognormal and gamma distributions were the best candidates, χ^2 goodness-of-fit tests reveal that only O_3 (pre-SEZ) and PM10 (post-SEZ) truly conform to these simple models ($p > 0.05$). Other pollutants' low p-values ($p < 0.05$) underscore that single-distribution assumptions cannot fully capture their complex, real-world variability.

3. **Policy Relevance:** The mixed success of theoretical fits highlights the need for nuanced air-quality strategies. Continued dust suppression should be reinforced, and targeted NO_x controls (e.g. stricter vehicle standards, industrial scrubbers) are critical. Given the poor fit of many gases, adaptive monitoring—potentially using mixture models or seasonally segmented analyses—will better inform regulatory actions.

4. **Limitations & Future Work:** This case study covers only two one-year snapshots and one SEZ region, so results may not generalize. Confounding factors (seasonality, regional policies) and district-level aggregation limit resolution. Future research should apply our spatial-statistical framework across multiple SEZs and longer timeframes, incorporate higher-resolution data (satellite, mobile sensors), and explore health or ecosystem impacts to ensure balanced economic and environmental objectives.

12 References

Bansal, G., Saini, R., Pathak, P., & Kulshrestha, A. (2022). Assessment of air pollution and its effects on health of residents of rapidly growing urban and industrial areas in India using AERMOD and public health data. *Environmental Science and Pollution Research, 29*(1), 1–16. https://doi.org/10.1007/s11356-022-22317-0

Dholakia, H. H., Purohit, P., Rao, S., Garg, A., & Amann, M. (2020). Cost-benefit analysis of ambient air quality improvements in India using the GAINS model. *Environmental Research Letters, 15*(9), 094025. https://doi.org/10.1088/1748-9326/ab927e

Galle, S., Overbeck, A., Riedel, N., & Seidel, T. (2022). Place-based policies and structural change: Evidence from Special Economic Zones in Africa (STEG Working Paper No. 040). *Structural Transformation and Economic Growth (STEG)*. https://steg.cepr.org/sites/default/files/2022-11/WP040%20GalleOverbeckRiedelSeidel%20PlaceBasedPoliciesAndStructuralChange.pdf

Gupta, A., Singh, R., & Chauhan, A. (2022). Challenges in air quality monitoring infrastructure in Indian cities. *Journal of Environmental Management, 306*, 114395. https://doi.org/10.1016/j.jenvman.2022.114395

Guttikunda, S. K., & Jawahar, P. (2014). Atmospheric emissions and pollution from the coal-fired thermal power plants in India. *Atmospheric Environment, 92*, 449–460. https://doi.org/10.1016/j.atmosenv.2014.04.057

India Briefing. (2011, December 21). *Special Economic Zones in Delhi NCR*. https://www.india-briefing.com/news/special-economic-zones-delhincr-5478.html/

India Briefing. (2020, June 19). *Special Economic Zones and warehousing clusters in Delhi NCR*. https://www.india-briefing.com/news/special-economic-zones-warehousing-clusters-delhi-ncr-18871.html/

PricewaterhouseCoopers. (2021, July 16). *SEZ impact assessment study report*. Export Promotion Council for EOUs and SEZs (EPCES). https://www.epces.in/uploads/SEZ%20impact%20assessment%20study%20report_PwC_revised%20final%20version_20210716.pdf

Rao, P. S., Raghunandan, B. S., & Kumari, S. (2017). Impact of industrial zones on ambient air quality in South India: A case study of Tamil Nadu SEZs. *International Journal of Environmental Studies, 74*(5), 703–717. https://doi.org/10.1080/00207233.2017.1295852

Research Publish Journals. (n.d.). *Impact of Special Economic Zones (SEZs) on economic growth and development of India*. https://www.researchpublish.com/upload/book/Impact%20of%20Special%20Economic-4038.pdf

Sarkar, A., & Banerjee, S. (2019). Ecological implications of Special Economic Zones in India: An environmental justice perspective. *Environmental Justice, 12*(4), 176–185. https://doi.org/10.1089/env.2018.0041

Shah, R., Taneja, A., & Joshi, N. (2021). GIS-based assessment of industrial emissions and their environmental impact in Delhi-NCR. *Environmental Monitoring and Assessment, 193*(5), 275. https://doi.org/10.1007/s10661-021-09087-6

Sharma, M., & Jain, S. (2020). A review of particulate matter pollution and health impacts in Indian cities. *Environment International, 138*, 105667. https://doi.org/10.1016/j.envint.2020.105667

World Bank. (2016). *The cost of air pollution: Strengthening the economic case for action*. World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/781521473177013155

Zauba Corp. (n.d.). *ARTHA Infratech Private Limited*. https://www.zaubacorp.com/ARTHA-INFRATECH-PRIVATE-LIMITED-U72200DL2008PTC182611

All Studies Journal. (2018). *Impact of Special Economic Zones on regional development in India*. *International Journal of Multidisciplinary Studies*, *3*(2), 215–315. https://allstudiesjournal.com/assets/archives/2018/vol3issue2/3-2-215-315.pdf

13 Appendices

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import lognorm, weibull_min, gamma
import warnings
warnings.filterwarnings('ignore')

# Set plot aesthetics using Seaborn
sns.set_style("whitegrid")  # Replace plt.style.use
sns.set_context("paper", font_scale=1.5)
```

```
# Load 2019(pre-SEZ) and 2021 (post-SEZ) data
# Assuming your Excel files are named accordingly
pre_sez = pd.read_csv('data/sector51gurugram2021.csv', parse_dates=['Timestamp'])
post_sez = pd.read_csv('data/sector51gurugram2023.csv', parse_dates=['Timestamp'])
```

```
# Check for missing values
print("\nMissing Values in Pre-SEZ Data:")
print(pre_sez.isnull().sum())
print("\nMissing Values in Post-SEZ Data:")
print(post_sez.isnull().sum())
```

```
def preprocess_data(df):
    """
    Preprocess air quality data by:
    1. Setting date as index
    2. Filling missing values using linear interpolation
    3. Removing any rows that still have missing values
    4. Ensuring pollutant values are numeric
    """
    # Copy to avoid modifying original
    df_clean = df.copy()

    # Set date as index if not already
    if 'Timestamp' in df_clean.columns:
        df_clean.set_index('Timestamp', inplace=True)

    # Select only pollution columns we need
```

```
    pollutants = ['PM2.5 (µg/m³)', 'NO2 (µg/m³)', 'SO2 (µg/m³)', 'PM10 (µg/m³)','NO (µg/m³)','CO (mg/m³)',
'Ozone (µg/m³)']
    df_clean = df_clean[pollutants]

    # Convert 'NA' strings to np.nan
    df_clean.replace('NA', np.nan, inplace=True)

    # Ensure all values are numeric
    for col in df_clean.columns:
        df_clean[col] = pd.to_numeric(df_clean[col], errors='coerce')

    # Fill missing values using linear interpolation
    df_clean = df_clean.interpolate(method='linear')

    # Drop any remaining rows with NaN values
    df_clean.dropna(inplace=True)

    return df_clean

# Apply preprocessing
pre_clean = preprocess_data(pre_sez)
post_clean = preprocess_data(post_sez)
```

```
def fit_distributions(data, pollutant):
    """
    Fit Lognormal, Weibull, and Gamma distributions to pollution data.
    Returns parameters and goodness-of-fit statistics.
    """
    # Get pollutant data
    values = data[pollutant].values

    # Initialize results dictionary
    results = {}

    # Fit Lognormal distribution
    shape_ln, loc_ln, scale_ln = stats.lognorm.fit(values)
    results['Lognormal'] = {
        'params': (shape_ln, loc_ln, scale_ln),
        'aic': stats.lognorm.nnlf((shape_ln, loc_ln, scale_ln), values) * 2 + 6  # AIC calculation
    }

    # Fit Weibull distribution
```

```
    shape_wb, loc_wb, scale_wb = stats.weibull_min.fit(values)
    results['Weibull'] = {
        'params': (shape_wb, loc_wb, scale_wb),
        'aic': stats.weibull_min.nnlf((shape_wb, loc_wb, scale_wb), values) * 2 + 6
    }

    # Fit Gamma distribution
    shape_g, loc_g, scale_g = stats.gamma.fit(values)
    results['Gamma'] = {
        'params': (shape_g, loc_g, scale_g),
        'aic': stats.gamma.nnlf((shape_g, loc_g, scale_g), values) * 2 + 6
    }

    return results
```

```
def chi_square_test(data, pollutant, dist_name, params):
    """
    Perform Chi-square goodness-of-fit test on the fitted distribution.
    Lower Chi-square values indicate better fit.
    """
    values = data[pollutant].values

    # Create histogram of observed values
    hist, bin_edges = np.histogram(values, bins=10, density=False)

#    hist, bin_edges = np.histogram(values, bins='auto', density=False)

    # Calculate bin midpoints for expected calculation
    bin_midpoints = (bin_edges[1:] + bin_edges[:-1]) / 2

    # Calculate expected frequencies based on the fitted distribution
    if dist_name == 'Lognormal':
        cdf_values = stats.lognorm.cdf(bin_edges, *params)
    elif dist_name == 'Weibull':
        cdf_values = stats.weibull_min.cdf(bin_edges, *params)
    elif dist_name == 'Gamma':
        cdf_values = stats.gamma.cdf(bin_edges, *params)

    # Calculate expected frequencies in each bin
    expected = len(values) * np.diff(cdf_values)

    # Handle zeros in expected frequencies (add a small value to avoid division by zero)
    expected = np.where(expected < 0.001, 0.001, expected)
```

```
    # Calculate Chi-square statistic
    chi2_stat = np.sum((hist - expected)**2 / expected)

    # Calculate degrees of freedom (bins - parameters - 1)
    df = len(hist) - len(params) - 1

    # Calculate p-value
    p_value = 1 - stats.chi2.cdf(chi2_stat, df)

    return {'chi2': chi2_stat, 'p_value': p_value, 'df': df}
```

```
# Define pollutants to analyze
pollutants = ['PM2.5 (µg/m³)','PM10 (µg/m³)','NO (µg/m³)','NO2 (µg/m³)','CO (mg/m³)','SO2 (µg/m³)','Ozone
(µg/m³)']

# Initialize dictionaries to store fitting results
pre_distributions = {}
post_distributions = {}
chi_square_results = {}

# Fit distributions to each pollutant in pre-SEZ data
for pollutant in pollutants:
    pre_distributions[pollutant] = fit_distributions(pre_clean, pollutant)

    # Find best distribution based on AIC
    best_dist = min(pre_distributions[pollutant],
                    key=lambda x: pre_distributions[pollutant][x]['aic'])

    # Perform Chi-square test on the best distribution
    params = pre_distributions[pollutant][best_dist]['params']
    chi_result = chi_square_test(pre_clean, pollutant, best_dist, params)

    # Store results
    chi_square_results[f"pre_{pollutant}"] = {
        'best_distribution': best_dist,
        'chi2': chi_result['chi2'],
        'p_value': chi_result['p_value']
    }

# Fit distributions to each pollutant in post-SEZ data
for pollutant in pollutants:
    post_distributions[pollutant] = fit_distributions(post_clean, pollutant)

    # Find best distribution based on AIC
```

```
    best_dist = min(post_distributions[pollutant],
                    key=lambda x: post_distributions[pollutant][x]['aic'])

    # Perform Chi-square test on the best distribution
    params = post_distributions[pollutant][best_dist]['params']
    chi_result = chi_square_test(post_clean, pollutant, best_dist, params)

    # Store results
    chi_square_results[f"post_{pollutant}"] = {
        'best_distribution': best_dist,
        'chi2': chi_result['chi2'],
        'p_value': chi_result['p_value']
    }

# Display results
for key, value in chi_square_results.items():
    print(f"{key}: Best fit = {value['best_distribution']}, Chi² = {value['chi2']:.2f}, p = {value['p_value']:.4f}")
```

```
# Define Indian NAAQS standards (in µg/m³)
NAAQS = {

    'PM2.5 (µg/m³)': 60,  # 24-hour average

    'NO2 (µg/m³)': 80,    # 24-hour average

    'SO2 (µg/m³)': 80,    # 24-hour average

    'PM10 (µg/m³)': 100,  # 24-hour average

    'NO (µg/m³)': 80,    # 8-hour average

    'CO (mg/m³)': 4.0,    # 8-hour average

    'Ozone (µg/m³)': 100      # 8-hour average (Ozone)

}


# Calculate exceedance rates for pre-SEZ
pre_exceedance = {}
for pollutant in pollutants:
    exceedance_count = (pre_clean[pollutant] > NAAQS[pollutant]).sum()
    exceedance_percent = (exceedance_count / len(pre_clean)) * 100
    pre_exceedance[pollutant] = {
        'count': exceedance_count,
```

```
        'percent': exceedance_percent
    }

# Calculate exceedance rates for post-SEZ
post_exceedance = {}
for pollutant in pollutants:
    exceedance_count = (post_clean[pollutant] > NAAQS[pollutant]).sum()
    exceedance_percent = (exceedance_count / len(post_clean)) * 100
    post_exceedance[pollutant] = {
        'count': exceedance_count,
        'percent': exceedance_percent
    }

# Print exceedance results
print("\nNAAQS Exceedance Analysis:")
print("-" * 50)
print(f"{'Pollutant':<10} | {'Pre-SEZ (2005)':<20} | {'Post-SEZ (2007)':<20}")
print("-" * 50)
for pollutant in pollutants:
    pre_pct = pre_exceedance[pollutant]['percent']
    post_pct = post_exceedance[pollutant]['percent']
    print(f"{pollutant.split()[0]:<10} | {pre_pct:.1f}% ({pre_exceedance[pollutant]['count']} days) | "
          f"{post_pct:.1f}% ({post_exceedance[pollutant]['count']} days)")
```

```
def plot_distribution_fit(data, pollutant, period, distributions_dict, NAAQS):
    """
    Plot histogram of observed data with fitted distributions.
    Now handles missing data, auto-scales, and is safe against key errors.
    """
    # Check if pollutant exists in distribution results
    if pollutant not in distributions_dict:
        print(f"⚠ Pollutant '{pollutant}' not found in provided distribution results for {period}. Skipping.")
        return

    values = data[pollutant].dropna().values  # Remove NaN

    # Check if any valid data
    if len(values) == 0:
        print(f"⚠ No data available for '{pollutant}' ({period}). Skipping.")
        return

    # Get distribution fitting results
    dist_results = distributions_dict[pollutant]
```

```
# Find best distribution by minimum AIC
best_dist = min(dist_results, key=lambda x: dist_results[x]['aic'])

# Smart x range
try:
    lower, upper = np.percentile(values, [1, 99])
    if lower == upper:  # fallback in case percentile collapsed
        lower = np.min(values)
        upper = np.max(values)
except Exception as e:
    print(f"⚠ Problem calculating percentiles for {pollutant}: {e}")
    return

x = np.linspace(lower, upper, 1000)

# Create plot
plt.figure(figsize=(12, 6))

# Plot histogram of observed data
plt.hist(values, bins='auto', density=True, alpha=0.6, color='orange', label='Observed Data')

# Plot best-fit distribution
params = dist_results[best_dist]['params']

if best_dist == 'Lognormal':
    y = stats.lognorm.pdf(x, *params)
elif best_dist == 'Weibull':
    y = stats.weibull_min.pdf(x, *params)
elif best_dist == 'Gamma':
    y = stats.gamma.pdf(x, *params)
else:
    print(f"⚠ Unknown best distribution '{best_dist}' for {pollutant}. Skipping.")
    return

plt.plot(x, y, 'b-', linewidth=2, label=f'{best_dist} (Best Fit)')

# Plot other candidate distributions (thinner lines)
line_styles = ['--', ':']  # dashed and dotted
style_index = 0

for dist_name, dist_info in dist_results.items():
    if dist_name == best_dist:
        continue
    params = dist_info['params']
    if dist_name == 'Lognormal':
```

```
            y = stats.lognorm.pdf(x, *params)
        elif dist_name == 'Weibull':
            y = stats.weibull_min.pdf(x, *params)
        elif dist_name == 'Gamma':
            y = stats.gamma.pdf(x, *params)
        else:
            continue  # skip unknown

        line_style = line_styles[style_index % len(line_styles)]
        plt.plot(x, y, line_style, linewidth=1.5, alpha=0.8, label=dist_name)
        style_index += 1


    # Plot NAAQS limit if available
    if pollutant in NAAQS:
            plt.axvline(x=NAAQS[pollutant], color='black', linestyle='-.', label=f'NAAQS ({NAAQS[pollutant]} µg/m³)')

    # Labels and title
    plt.xlabel(pollutant)
    plt.ylabel('Probability Density')
    plt.title(f'{period}: {pollutant} Distribution Fit')
    plt.legend()
    plt.grid(True, alpha=0.3)

    plt.xlim(lower, upper)  # smart limits
    plt.tight_layout()

    # Save plot safely
    safe_name = pollutant.replace(" ", "_").replace("/", "_")
    plt.savefig(f'{period.lower().replace("-", "_")}_{safe_name}_distribution.png', dpi=300)
    plt.close()

    print(f"Finished plot for {pollutant} ({period}).")




# For pre-SEZ period
for pollutant in pollutants:
    plot_distribution_fit(pre_clean, pollutant, 'Pre-SEZ', pre_distributions, NAAQS)

# For post-SEZ period
for pollutant in pollutants:
    plot_distribution_fit(post_clean, pollutant, 'Post-SEZ', post_distributions, NAAQS)
```

```
# Calculate percentage change in pollutant concentrations and include additional statistics
percent_change = {}
median_change = {}
percentile25_change = {}

for pollutant in pollutants:
    pre_mean = pre_clean[pollutant].mean()
    post_mean = post_clean[pollutant].mean()
    pre_median = pre_clean[pollutant].median()
    post_median = post_clean[pollutant].median()
    pre_25 = pre_clean[pollutant].quantile(0.25)
    post_25 = post_clean[pollutant].quantile(0.25)

    percent_change[pollutant] = ((post_mean - pre_mean) / pre_mean) * 100
    median_change[pollutant] = ((post_median - pre_median) / pre_median) * 100
    percentile25_change[pollutant] = ((post_25 - pre_25) / pre_25) * 100

# Create a comprehensive results dataframe
results_data = []
for pollutant in pollutants:
    pre_best = chi_square_results[f"pre_{pollutant}"]['best_distribution']
    post_best = chi_square_results[f"post_{pollutant}"]['best_distribution']

    results_data.append({
        'Pollutant': pollutant.split()[0],
        'Pre-SEZ Mean': pre_clean[pollutant].mean(),
        'Post-SEZ Mean': post_clean[pollutant].mean(),
        'Change (%) (Mean)': percent_change[pollutant],
        'Pre-SEZ Median': pre_clean[pollutant].median(),
        'Post-SEZ Median': post_clean[pollutant].median(),
        'Change (%) (Median)': median_change[pollutant],
        'Pre-SEZ Distribution': pre_best,
        'Post-SEZ Distribution': post_best,
        'Distribution Changed': pre_best != post_best,
        'Pre-SEZ Exceedance (%)': pre_exceedance[pollutant]['percent'],
        'Post-SEZ Exceedance (%)': post_exceedance[pollutant]['percent'],
        'Exceedance Change (pp)': post_exceedance[pollutant]['percent'] - pre_exceedance[pollutant]['percent']
    })

# Convert to DataFrame
results_df = pd.DataFrame(results_data)
```

```
# Save results
results_df.to_csv('noida_sez_impact_results.csv', index=False)

# Display
print("\nFinal Results Summary:")
print("-" * 120)
print(results_df.to_string(index=False))
print("-" * 120)

# Bar chart for mean change
plt.figure(figsize=(12, 6))
colors = ['green' if x < 0 else 'red' for x in results_df['Change (%) (Mean)'].values]
plt.bar(results_df['Pollutant'], results_df['Change (%) (Mean)'], color=colors)
plt.axhline(y=0, color='black', linestyle='-', linewidth=0.5)
plt.title('Percentage Change in Pollutant Mean Concentrations (Post-SEZ vs Pre-SEZ)')
plt.ylabel('Change (%) (Mean)')
plt.grid(axis='y', alpha=0.3)
for i, v in enumerate(results_df['Change (%) (Mean)']):
    plt.text(i, v + (5 if v > 0 else -5), f"{v:.1f}%", ha='center', va='center')
plt.savefig('pollutant_concentration_changes_mean.png', dpi=300)
plt.close()
```