**Analyzing and Predicting Air Quality Trends: A Case Study on Seasonal and Policy Impacts**

Parker Pratt

Department of Mathematics, Oregon Tech

DSAI 443: Deep Learning

Prof. Joseph Reid

March 19, 2025

**Introduction & Background**

Air pollution is a growing global concern with significant public health, environmental, and economic implications. Poor air quality can lead to respiratory issues, cardiovascular diseases, and increased mortality rates, presenting major challenges for public health systems and policy makers (Air Pollution, 2024). Business analytics plays a critical role in this area by providing insights into pollution trends, identifying anomalies, and predicting future pollution levels. Through effective data analysis, businesses and governments can design targeted interventions, improve regulatory frameworks, and enhance public health outcomes.

This study uses air quality data from the World Air Quality Historical Database (World Air Quality), focusing on two major cities: Delhi, India (2457 rows) and London, England (4080 rows). The key pollutants analyzed include:

* **PM2.5** (ppb) – Particulate matter less than 2.5 microns in diameter
* **PM10** (ppb) – Particulate matter less than 10 microns in diameter
* **O3** (ppb) – Ozone levels
* **NO2** (ppb) – Nitrogen dioxide levels
* **SO2** (ppb) – Sulfur dioxide levels
* **CO** (ppm) – Carbon monoxide levels

The dataset includes daily air quality readings in either parts per billion (ppb) or parts per million (ppm). The data spans multiple years and each point also has the date that it was recorded. Raw, the data points are strings, but I later convert them to floats so that they are continuous, and missing values can be interpolated. Below are the distributional summaries for each city:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Delhi | COUNT | N/A % | MIN | MEAN | MEDIAN | MODE | MAX | STD | SKEW |
| PM25 | 2436 | 0.85 | 21 | 183.56 | 165 | 157 | 677 | 91.06 | 1.14 |
| PM10 | 2428 | 1.20 | 7 | 194.78 | 163 | 76 | 999 | 128.84 | 1.39 |
| O3 | 2414 | 1.75 | 1 | 25.51 | 23 | 14 | 93 | 17.15 | 0.90 |
| NO2 | 2427 | 1.22 | 1 | 17.55 | 16 | 11 | 109 | 8.84 | 1.47 |
| SO2 | 2425 | 1.30 | 1 | 7.47 | 6 | 5 | 36 | 4.33 | 1.40 |
| CO | 2421 | 1.47 | 1 | 10.22 | 9 | 8 | 112 | 5.97 | 4.13 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| London | COUNT | N/A % | MIN | MEAN | MEDIAN | MODE | MAX | STD | SKEW |
| PM25 | 4070 | 0.25 | 13 | 57.32 | 54 | 52 | 329 | 21.92 | 1.71 |
| PM10 | 4069 | 0.27 | 4 | 23.74 | 22 | 19 | 89 | 9.85 | 1.33 |
| O3 | 4033 | 1.15 | 1 | 24.95 | 25 | 27 | 91 | 9.47 | 0.77 |
| NO2 | 4036 | 1.08 | 3 | 31.20 | 29 | 20 | 92 | 15.01 | 0.53 |
| SO2 | 3074 | 24.66 | 1 | 3.47 | 3 | 1 | 27 | 2.47 | 2.07 |
| CO | 3783 | 7.28 | 1 | 4.36 | 4 | 3 | 19 | 2.71 | 1.02 |

**Related Works**

Several studies have explored the analysis and prediction of air quality trends using statistical and machine learning methods. The growing availability of historical air quality data has enabled researchers to identify patterns, forecast pollution levels, and assess the effectiveness of environmental policies.

Liu et al. (2023) provided a comprehensive review of business analytics techniques applied to environmental data, highlighting the importance of combining descriptive, predictive, and prescriptive approaches. Their work underscores the role of machine learning models, including time-series forecasting methods like Long Short-Term Memory (LSTM) networks, in improving the accuracy of air quality predictions.

A study conducted in Visakhapatnam, India, applied machine learning techniques to forecast the Air Quality Index (AQI), focusing on 12 contaminants and 10 meteorological parameters over a five-year period. The research highlighted the efficacy of machine learning models in predicting air quality and emphasized the importance of incorporating comprehensive environmental and meteorological data (Patra et al., 2023).

Another study introduced a time-based-spatial (TBS) forecasting framework that integrates spatial and temporal information using machine learning techniques. This approach demonstrated improved accuracy in forecasting the AQI in urban environments, underscoring the potential of advanced modeling techniques in air quality prediction (Liu et al., 2024).

Research analyzing the seasonal distribution patterns of six key pollutants found that temperature, humidity, air pressure, and atmospheric conditions significantly influence pollutant concentrations. This underscores the necessity of incorporating seasonal dynamics into air quality forecasting models to enhance their accuracy and reliability (Chen et al., 2024) .

A study evaluated the impacts of winter heating and clean heating policies on air quality in China using an observation-based causal inference approach. The findings demonstrated that clean heating policies led to a reduction in PM2.5 concentrations, highlighting the effectiveness of targeted interventions in improving air quality during specific seasons (Wang et al., 2023) .

A comprehensive analysis compared various statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. The study provided recommendations for assessing the impacts of anthropogenic emission changes on air quality, emphasizing the importance of selecting appropriate statistical approaches for accurate trend analysis (Smith et al., 2022) .

Another study compared statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentrations. The research highlighted the strengths and limitations of different modeling approaches in capturing spatial and temporal variations in pollutant levels, providing insights into effective methods for air quality assessment (Williams et al., 2019) .

**Objectives**

My analytical objectives within the dataset are to examine trends in air quality in different places around the world and the trends of different particles in the air. I will also use the information I have to predict how the air quality might change in the future. In addition, I will look at how seasons, policies, regulations, and other outside factors might affect air quality.

Some questions that my analytical objectives will answer are:

* How has air quality changed over time?
* How have policies and regulations affected air quality over time if at all?
* Do seasons have a noticeable effect on air quality?
* How do I predict air quality will change in the future?
* What is the reason for air quality being different around the world?

My datasets can answer these questions through various analysis techniques. Some of these that I will employ are basic statistics, anomaly detection, and long short-term memory models. Basic statistics will help with having a base-level understanding. This can help me answer the first question above and some of the second question. Anomaly detection will give me additional information related to the second question and will answer the third question. LSTM’s can answer the fourth question through machine learning and the fifth question can be answered through a combination of all of these and some outside resources. By answering these questions, I would have developed a very deep understanding of the data and the results that I will have.

These questions can be very useful for business analytics in multiple ways. By understanding how air quality changes by season, businesses in areas such as construction can plan projects during periods of improved air quality for the safety of their workers. Being able to predict how air quality will be in the future can also help with this and help policymakers investigate the causes of extreme events. Determining if certain policies and regulations are effective in improving air quality is also important. This can give companies an idea of things they should change for the safety of people, and get ahead of possible policies and regulations that might be put in place in the future.

**Analytical Plan**

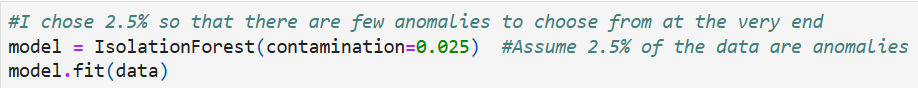
This analytical plan outlines the methods and techniques used to analyze and predict air quality trends based on historical data from the World Air Quality Historical Database. The analysis consists of three main components: descriptive analysis, predictive analysis, and prescriptive analysis as they are described by Liu, et al, in 2023. Descriptive analysis will be performed with t-tests and anomaly detection methods. This will be used to examine historical trends and variations in air quality. Predictive analysis aims to forecast future air quality levels using machine learning models such as the LSTM that I employed. Prescriptive analysis includes using t-tests to evaluate the impact of policies on air quality and recommend improvements. The analysis focuses on six key pollutants (PM2.5, PM10, O3, NO2, SO2, CO) across two major cities (Delhi, and London), with particular attention to seasonal variations and policy changes.

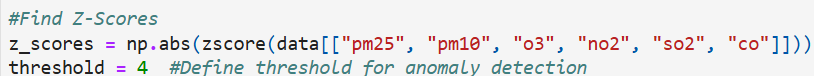
To prepare the data for analysis, several cleaning and transformation steps were necessary.

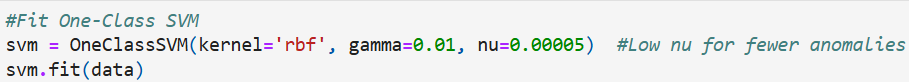
A screenshot of a computer program

AI-generated content may be incorrect.

Missing values were imputed using the median of the column if it was in the first row and the interpolate method for time series data for all the points after that. The ‘date’ field was made into the index of the data frame, and the month was used to classify the season (spring, summer, fall, winter). Pollutant levels were converted to floats and scaled to ensure consistency and improve model performance. These preprocessing steps were essential to ensure that the data was properly formatted for statistical and machine learning techniques.

A combination of statistical and machine learning methods was employed to conduct the analysis. Descriptive analysis involved calculating summary statistics, including mean, median, standard deviation, and skewness, to understand the underlying patterns in the data. Anomaly detection was also performed using three complementary methods: Isolation Forest, Z-Score, and One-Class SVM.





Isolation Forest was effective at detecting outliers by using random partitioning, Z-Score identified values that significantly deviated from the mean, and One-Class SVM captured complex patterns in normal pollution levels. Combining these methods improved the robustness of the anomaly detection process, and points flagged by all three methods were considered highly significant anomalies.

Predictive analysis was conducted using a Long Short-Term Memory (LSTM) model, which is well-suited for time-series forecasting.

A screenshot of a computer code

AI-generated content may be incorrect.

The data was split into a training set (80%) and a testing set (20%) to evaluate the model’s effectiveness.

A computer screen shot of a program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

The LSTM model was tuned to optimize accuracy and minimize overfitting, and its performance was evaluated using Mean Squared Error (MSE).

To assess the effectiveness of policy interventions through prescriptive analysis, t-tests were conducted on air quality data before and after the implementation of London's Ultra-Low Emission Zone (ULEZ) in 2019.

The analysis was conducted using Python and several key libraries. Data manipulation and preprocessing were performed using pandas and NumPy, while machine learning models were implemented using scikit-learn and PyTorch/Keras. Data visualization was handled with matplotlib, and the analysis workflow was documented in Jupyter Notebooks. These tools provided a flexible and powerful environment for conducting both statistical and machine learning-based analysis.

Interpretation guidelines were established to ensure consistent and meaningful analysis. For anomaly detection, points identified by all three methods were considered highly significant and investigated further to identify environmental or policy-related causes. Predictive analysis was evaluated based on the model’s ability to track overall trends and minimize forecasting error. Statistical significance was assessed using p-values, with values below 0.05 indicating a meaningful difference between pre- and post-policy data.

The selected methods were chosen based on their suitability for the type of data and the analytical objectives. Isolation Forest, Z-Score, and One-Class SVM were selected for anomaly detection due to their ability to capture different types of outliers and deviations. LSTM was chosen for predictive analysis because of its ability to model complex time-series data with long-term dependencies. And t-tests were used for statistical inference to measure the impact of policy changes.

**Results**

First, I am going to perform anomaly detection on the Delhi, India dataset. From this, we will be able to understand the reasoning for unusually bad air quality and when this usually occurs.

I will start with descriptive analysis and will be using the Isolation Forest method to find anomalies by season. This is plotted as PM2.5 x PM10 because these are the particles most associated with the AQI (Air Quality Index) and anomalies will be outlined in black.

**A diagram of a number of dots

AI-generated content may be incorrect.**

As you can see, PM2.5 is heavily influenced by season. Spring and summer usually have much lower values than in fall and winter. PM10 doesn’t seem to be as dependent on the season but fall and winter still have slightly larger values on average.

After doing some research into the reason behind this, I found that “This can be attributed to multiple factors including temperature inversion in which a layer of cool air forms near the ground, trapping pollutants and preventing them from dispersing. Also, low wind speeds during the winters reduce the dispersion of pollutants” (Khan, 2024). I also found that during late October and November, farmers in Punjab, Haryana, and Uttar Pradesh often burn crop residues after the harvest. This practice releases large amounts of smoke and particulate matter into the atmosphere, which can drift to Delhi and significantly degrade air quality (Ellis-Petersen, 2024).

Anomaly detection analyses identify extreme pollution events that could be linked to environmental changes or data anomalies. Each method provides unique insights:

* Isolation Forest effectively detected outliers by using random partitioning.
* Z-Score highlighted values significantly different from the mean.
* One-Class SVM captured complex boundaries of normal pollution patterns.

Combining these methods allows for robust anomaly detection in air quality monitoring.

Because of this, I will now employ One-Class SVM and Z-Score anomaly detection methods in addition to the Isolation Forest method to narrow down the number of anomalies even more so that specific reasons can be found. The figure below shows the results of the three methods with red points being the anomalies and the blue points being considered “normal”.

**A diagram of a number of red dots

AI-generated content may be incorrect.**

Merging these methods together, we can color the points blue if they are detected by none of the methods as anomalies, yellow for one method labeling it as an anomaly, orange for two, and red for three.

**A diagram of a number of colored dots

AI-generated content may be incorrect.**

There are four red points in the above plot. Because they were labeled as anomalies by all three methods, they must really be something special. We’ll output the rows associated with these points so we can properly see what might’ve occurred.

**A table with numbers and a number of objects

AI-generated content may be incorrect.**

Looking at these four rows we can see that most of them are in either fall or winter, which makes sense considering the first plot that we made that compared the points by season. The first row is most likely due to the extremely high levels of nitrogen dioxide, the second is probably a combination of PM10 and ozone, and the third and fourth are a combination of PM2.5 and PM10. As found earlier in the article by Ellis-Petersen, the first, third, and fourth are most likely extremely bad cases of stubble burning mixed with weak wind. But what is the case for the point in May of last year? Well according to an article on ET Online, the anomaly in May was partly due to a severe dust storm that passed through in addition to the effects that that had on reducing the year-round sources of emissions from things such as vehicles, industrial activities, and construction dust.

In April of 2019, London, England implemented a plan to reduce air pollution through vehicle emissions. This “Ultra-Low Emission Zone” was part of a goal to reduce nitrogen dioxide levels in the air so that it would be safer to breathe. (City of London, 2019).

To answer the questions about air quality over time and whether policies like these are effective, I will be using an LSTM (long short-term memory) model. After performing a train-test split where 80% of the data is used to train and the last 20% is used to test effectiveness, I am given the plots below as the results. This shows blue lines as the real data that the model was trained on, green lines as the real data that the model was tested on, red dotted lines as the predicted results from the model, and a vertical orange dotted line as the year that the policy went into effect so that we can see if there was any change from it.

**A screenshot of a graph

AI-generated content may be incorrect.**

**A screenshot of a graph

AI-generated content may be incorrect.**

First off, we can see that although the model doesn’t quite reach the highest heights or the lowest lows, it is still very effective at following the trend for all the air particles.

Regarding whether the policy was effective, the naked eye can see that indeed, NO2 levels decreased after the policy went into effect. The other particles aren’t as obvious, however. To solve this, we will perform a t-test on the data where we split from before the policy and after the policy for each of the particles to determine if there was a significant change or not. Here are the results from the test:

A white paper with numbers and letters

AI-generated content may be incorrect.

Because of the amount of data present, the test is especially sensitive; even to what seems like very small changes. So even though it looked like there wasn’t much change for most of the particles, the test found that all of them except ozone was significantly different after the policy. In fact, the nitrogen dioxide levels were so significantly different that the computer couldn’t compute a number that small and output a straight 0.0.

**Discussion**

A few strengths of this project are the robustness of the anomaly detection model, and the accuracy of the LSTM model. By combining Isolation Forest, One-Class SVM, and Z-Score, a very robust model that finds anomalies in many ways is created that can find anomalies with significant and important insights into air quality. Even with there being room for improvement for the LSTM model, being able to create something that can capture the gist for each particle is extraordinary considering the somewhat small amount of data used to train it. T-tests were also extremely effective in classifying whether policies had a significant impact on air quality or not, even for slight differences that aren’t noticeable to the naked eye.

The weaknesses of this project are that many places that I looked at analyzing and even the two that I did analyze have a decent number of missing values. This can cause inaccuracies in analyses and predictions and can make some places impossible to interpret. Another weakness was that the LSTM model was not great at predicting extreme peaks or dips even if it did follow the general pattern.

Over the course of working on this project, I was given a few recommendations. Some of these recommendations I included, and some I didn’t. One of these recommendations that I included was that I should investigate the differences in air quality season-by-season and find policies or regulations so that I can determine if they are effective at improving air quality or not. I really liked this recommendation, and it ended up becoming one of the most important pieces of my project. A piece of feedback that I didn’t get to include was that I should look at how it differs between urban and rural areas. This is a great idea and could be very beneficial in finding more reasons behind how air quality is in certain areas, but due to lack of data in non-urban areas, I was unfortunately unable to implement this idea.

Through this project, I gained deeper insights into the challenges of working with environmental data and time-series forecasting. I learned how to apply machine learning models to real-world problems and the importance of combining multiple methods for robust anomaly detection. The project also reinforced the value of statistical validation in measuring policy effectiveness. This project has taught me a lot about air quality and the reasons why some places have cleaner air than others. Overall, I have learned how powerful data can be—and all the different ways that the same data can be used to learn different things.

**Recommendations**

An idea for future work in this area includes incorporating weather and traffic data to better understand the data and improve predictive models. Another recommendation is to extend analysis to additional cities that have other policies and regulations to determine if there is a specific type that is most effective in improving air quality.

**Conclusion**

This study demonstrated the value of business analytics in monitoring and improving air quality. Descriptive analysis provided a clear understanding of pollution trends, and that air quality seems to change with season; but going back as far as I can with this data, air quality has not changed much overall, unless certain policies are put into effect. Predictive analysis helped forecast future pollution levels and led me to predict that air quality will most likely stay in the same cycle unless policies and regulations such as London’s “ULEZ” are put in place. Prescriptive analysis measured the effectiveness of policy changes and determined that London’s “ULEZ” was effective in improving air quality and that other places in the world might benefit from something like this being implemented. The results also indicate that more than weather and population are important in finding the differences in air quality around the world; and that the reason for some of these differences is because of practices or policies in that area like India’s stubble burning and London’s “ULEZ”. Future work in this area could involve expanding the analysis to other cities and incorporating additional environmental factors like traffic and weather data to improve predictive accuracy.

**References**

*Air Pollution and Your Health. (2024, August 6). National Institute of Environmental Health Sciences; National Institute of Environmental Health Sciences.* [*https://www.niehs.nih.gov/health/topics/agents/air-pollution*](https://www.niehs.nih.gov/health/topics/agents/air-pollution)

Chen, L., Wang, Y., Zhang, Z., & Li, H. (2024). Seasonal distribution patterns of key pollutants and meteorological influences. *Atmosphere*, *15*(5), 553. <https://doi.org/10.3390/atmos15050553>

*City of London Air Quality Strategy Delivering healthy air in the City of London Draft for Consultation*. (2019). <https://www.cityoflondon.gov.uk/assets/Services-Environment/city-of-london-air-quality-strategy-2019-2024.pdf>

Ellis-Petersen, H. (2024, November 22). “The air is killing us”: why Delhi’s pollution problem runs deeper than smog season. The Guardian; The Guardian. <https://www.theguardian.com/world/2024/nov/22/the-air-is-killing-us-why-delhi-india-pollution-problem-runs-deeper-than-smog-season>

ET Online. (2024, May 17). Delhi’s pollution worsens: This May is already more polluted than the previous 2 years. The Economic Times; Economic Times. <https://economictimes.indiatimes.com/news/india/delhis-pollution-worsens-this-may-is-already-more-polluted-than-the-previous-2-years/articleshow/110202588.cms>

Khan, M. Z. (2024, November 15). *Explained: Why Delhi-NCR struggles with severe air pollution every winter*. @Bsindia; Business Standard. <https://www.business-standard.com/india-news/explained-why-delhi-ncr-struggles-with-severe-air-pollution-every-winter-124111501328_1.html>

Liu, S., Liu, O., & Chen, J. (2023). A Review on Business Analytics: Definitions, Techniques, Applications and Challenges. Mathematics, 11(4), 899. <https://doi.org/10.3390/math11040899>

Liu, Y., Zhang, H., & Wang, X. (2024). A time-based-spatial forecasting framework for urban air quality prediction. *Scientific Reports*, *14*(1), 83248. <https://doi.org/10.1038/s41598-024-83248-z>

Patra, S., Reddy, P. R., & Rao, D. R. (2023). Machine learning-based air quality prediction in Visakhapatnam, India. *Chemosphere*, *342*, 138900. <https://doi.org/10.1016/j.chemosphere.2023.138900>

Smith, A., Johnson, T., & Lee, K. (2022). Statistical and machine learning methods for evaluating air quality trends under changing meteorological conditions. *Atmospheric Chemistry and Physics*, *22*(15), 10551–10568. <https://doi.org/10.5194/acp-22-10551-2022>

Wang, J., Liu, H., & Li, F. (2023). Assessing the impact of clean heating policies on winter air quality in China: A causal inference approach. *Environmental Research*, *223*, 115378. <https://doi.org/10.1016/j.envres.2023.115378>

Williams, M., Kim, S., & Turner, C. (2019). Creating national daily maps of ambient PM2.5 concentrations using statistical and machine learning methods. *arXiv Preprint*. <https://arxiv.org/abs/1904.08931>

*World Air Quality Historical Database.* (n.d.). Aqicn.org. <https://aqicn.org/historical/>