Comparative Analysis of Pollution Trends in London Using Machine Learning

Heegon Kim

May 2024

Abstract

This dissertation presents a novel comparative analysis of urban air pollution trends across major global cities utilizing a range of machine learning algorithms. The study aims to synthesize historical data with advanced predictive models to delineate future air quality scenarios, assessing compliance with WHO guidelines. Targeting urban centers in London this research evaluates the impacts of industrialization, urban planning, and environmental policies on air pollution. Through a comprehensive dataset sourced from both global and local environmental agencies, this work seeks to uncover patterns and tipping points that correlate with health and sustainability outcomes. The ultimate objective is to provide data-driven policy recommendations to cultivate healthier urban environments, enriching the discourse on air quality management and advocating for actionable change.

I certify that all material in this dissertation which is not my own work has been identified.

Contents

1 Introduction

The rapid urbanization and industrialization of the 21st century have propelled significant technological and economic advancements. However, these developments often come at a considerable environmental cost, with urban air pollution emerging as a critical global health concern. Increased vehicular traffic, industrial emissions, and dense population centers contribute significantly to the degradation of air quality, posing severe risks to public health and urban sustainability.

As detailed in the extensive research conducted by Frank J. Kelly and Julia C. Fussell, air pollution remains a significant public health issue, despite historical improvements in air quality regulations and monitoring. Modern urban environments continually expose large populations to levels of particulate matter (PM) that exceed both European standards and the more stringent World Health Organisation (WHO) Air Quality Guidelines. This ongoing exposure is linked not only to established health risks but also to a broadening spectrum of disease outcomes, suggesting that the impact of air pollution is more extensive than previously understood [1] [2].

The complexity and variability of urban air pollution demand sophisticated analytical approaches to understand, predict, and mitigate its adverse effects. In this context, this dissertation employs advanced machine learning algorithms to conduct a comparative analysis of pollution trends in London. The study focuses on synthesizing historical air quality data with predictive models to forecast future scenarios and assess compliance with international health standards, particularly the World Health Organization (WHO) guidelines [3].

The WHO guidelines serve as a global benchmark for air quality and public health (Table 1). These guidelines are instrumental in framing the health risks associated with various pollutants and are crucial for developing effective environmental policies. The recent updates to these guidelines reflect an evolving understanding of the impacts of air pollution, underscoring the urgent need for enhanced regulatory measures.

Table 1: Recommended 2021 AQG levels compared to 2005 air quality guidelines

Pollutant	**Averaging Time**	**2005 AQGs**	**2021 AQGs**
$PM_{2.5}$, $\mu g/m^3$	Annual	10	5
	24-hour[a]	25	15
PM_{10}, $\mu g/m^3$	Annual	20	15
	24-hour[a]	50	45
O_3, $\mu g/m^3$	Peak season[b]	-	60
	8-hour[a]	100	100
NO_2, $\mu g/m^3$	Annual	40	10
	24-hour[a]	-	25
SO_2, $\mu g/m^3$	24-hour[a]	20	40
CO, mg/m^3	24-hour[a]	-	4

μg = micro gram.

[a] 99th percentile (i.e., 3–4 exceedance days per year).

[b] Average of daily maximum 8-hour mean O_3 concentration in the six consecutive months with the highest six-month running-average O_3 concentration.

Note: Annual and peak season is long-term exposure, while 24 hour and 8 hour is short-term exposure.

The objective of this research is to leverage a comprehensive dataset sourced from global and local environmental agencies to identify patterns and tipping points that correlate with health and sustainability outcomes. By employing machine learning models such as ARIMA, SARIMA, and LSTM, this study enhances the predictive accuracy of air pollution trends, enabling nuanced insights that can inform policy adjustments and urban planning strategies. Ultimately, this research aims to provide actionable insights and data-driven policy recommendations to improve urban air quality, contributing significantly to the discourse on environmental management and public health.

2 Data Collection and Cleaning

2.1 Data Source

This research utilized data from the UK Air quality data archive [4], which is managed and published by the Department for Environment, Food & Rural Affairs (Defra) and hosted by Ricardo Energy & Environment. The data archive is part of the UK's comprehensive air quality monitoring effort, featuring over 1500 monitoring sites across the country. These sites form various networks that collect air quality data using both automatic and non-automatic methods, ensuring a wide coverage and granularity of data related to urban air pollution.

2.2 Data Collection Method

Collecting comprehensive and accurate environmental data presents significant challenges, particularly in the context of air quality monitoring [5]. These issues are crucial in urban areas where the variability and dynamic nature of pollution sources demand rigorous data collection strategies and robust analytical methodologies. Data was extracted from the UK Air quality data archive, leveraging tools such as the Data Selector Tool for tailored dataset construction and raw automatic data retrieval from the preformatted files link. The archive provides several data types including monitoring data, descriptive statistics, and exceedance statistics. For this project, the primary focus was on:

Monitoring Data: Retrieved from automatic monitoring networks that continuously track air quality across various locations in the UK, providing a detailed and dynamic picture of pollution levels.

Descriptive Statistics: Utilized to obtain annual, monthly, and daily mean concentrations of key pollutants like PM2.5 and PM10, alongside other relevant statistical measures such as maximum and minimum values.

Exceedance Statistics: Analyzed to assess instances where pollution levels surpassed the UK Air Quality Objectives and EU Limit Values, important for evaluating the frequency and severity of high pollution episodes.

2.3 Data Cleaning and Preprocessing

Upon collection, the data underwent a rigorous cleaning process to ensure accuracy and relevance for the study:

Filtering and Quality Checks: Initial steps involved filtering out incomplete or inconsistent records, ensuring that datasets were complete and representative. Handling Missing Data: Missing values were addressed through imputation methods, where feasible, or by excluding dates and stations with insufficient data to maintain the integrity of the analysis.

Adjustment for Measurement Uncertainty: Recognizing changes in the reported measurement uncertainty for PM10 and PM2.5 (increased uncertainties since 2019 due to changes in manufacturing tolerances), adjustments were made in the analysis to account for these variations.

Standardization of Measurement Units: Data from different sources within the archive were standardized to ensure uniformity in measurement units and time scales, facilitating comparative analyses.

2.4 Adjustment for COVID-19 Impact

Upon reviewing the data, I noticed a significant drop in pollutant levels in 2020 (Figure 1), likely due to reduced human and industrial activity during the COVID-19 pandemic. This anomaly continued to affect pollution levels through to 2023. Given this disruption, I adjusted the dataset to begin from 2020, aligning the training data more closely with the current air quality context.

As the models that I will be using is sensitive to patterns. To achieve accurate forecast results, it was necessary to change the training dataset from 2020.

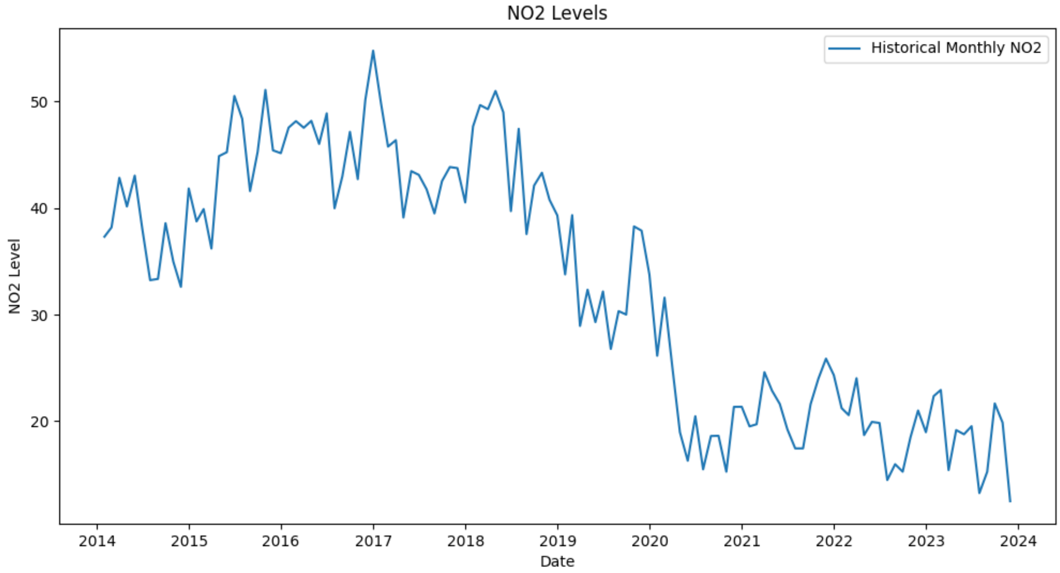

Figure 1: Historical NO_2 level of London

2.5 Licensing and Usage Rights

The data used in this research is available under the Open Government Licence (OGL), which permits free use, including copying, publishing, distributing, transmitting, adapting, and exploiting the information commercially or non-commercially. Compliance with the licence terms included proper attribution to Defra and the UK Air website, as stipulated by the licensing agreement. This adherence underscores the research's commitment to ethical data usage and acknowledgment of the data source.

3 ARIMA and SARIMA Models

3.1 Overview of ARIMA

ARIMA, which stands for Auto Regressive Integrated Moving Average, is a popular statistical analysis model used for forecasting time series data. This model is particularly useful in predicting future points in series based on past data [6]. ARIMA models are characterized by three key parameters: p, d, and q:

- **p** is the number of lag observations included in the model (lag order).
- **d** is the number of times that the raw observations are differenced (degree of differencing).
- **q** is the size of the moving average window (order of moving average).

The model works by transforming a non-stationary time series into a stationary one (if needed) through differencing, then applying statistical forecasting to the transformed series. It effectively captures both trends and randomness in the historical data.

Recent advancements in machine learning have introduced various optimization and regularization techniques that significantly enhance the predictive accuracy and robustness of models like ARIMA. These techniques address overfitting and improve the model's ability to generalize across different environmental conditions and datasets. A comprehensive review and application of these methods can be found in the study [7], which explores various approaches to regularize and optimize air quality prediction models using machine learning algorithms. This research highlights the importance of integrating advanced computational techniques to refine predictive models, ensuring they are both accurate and adaptable to new data.

Given the complexity of environmental data and the crucial need for accurate forecasts, validating machine learning models like ARIMA involves careful consideration of the methodology used to estimate their performance. Traditional validation methods such as K-fold Cross-Validation might not be sufficiently robust for small sample sizes, as discussed in the study on "Machine Learning Algorithm Validation with a Limited Sample Size" [8]. This study emphasizes the potential bias in performance estimates obtained from conventional validation techniques when dealing with limited data.

To address these challenges, this dissertation employs Nested Cross-Validation and a robust train/test split approach as recommended by the study. These methods are designed to provide more reliable performance estimates that are crucial for environmental forecasting applications. By integrating these robust validation techniques, this research aims to ensure that the predictive models are not only statistically sound but also practically reliable in real-world scenarios.

3.2 Challenges of ARIMA and Transition to SARIMA

Initially, I employed an ARIMA model to forecast air pollution levels for the year 2024, using data spanning from 2014 to 2023 and 2020 to 2023. This approach, however, yielded unsatisfactory results; the forecasted values formed a nearly linear trend, which did not accurately reflect the expected seasonal fluctuations and variability observed in historical data.

To enhance the model's accuracy by incorporating seasonality, the model was transitioned to Seasonal ARIMA (SARIMA), which extends the ARIMA model by adding seasonal elements well-suited for time series data. Also in a recent study, SARIMA models were employed to analyze air pollution in the small urban area of Blagoevgrad, Bulgaria, demonstrating the model's capability to effectively forecast short-term pollution levels based on historical data. The study highlighted the importance of using both non-seasonal and seasonal parameters to capture the complex patterns of air pollution, which are influenced by both anthropogenic activities and natural phenomena [9]. It includes additional seasonal parameters:

- **P** is the seasonal auto-regressive order.
- **D** is the seasonal differencing order.
- **Q** is the seasonal moving average order.
- **s** is the length of the seasonal cycle.

To investigate the seasonality in the air quality data, specifically the nitrogen dioxide (NO_2) concentrations from the year 2020 onwards, we employed the `seasonal_decompose` function from the `statsmodels` library in Python. This function decomposes a time series into three distinct components: trend, seasonality, and residual. The trend component reflects the long-term progression of the series, seasonality shows systematic, predictable patterns, and the residual represents the noise or randomness unexplained by the model.

The decomposition process is defined by the equation:

$$\text{Observed Time Series} = \text{Trend} + \text{Seasonality} + \text{Residual}$$

For this analysis, the additive model was chosen, which is suitable when seasonal variations are roughly constant throughout the series' progression. The command:

```
seasonal_decompose_result = seasonal_decompose(data_2020_onwards, model='additive')
```

executes this decomposition on the mean-resampled monthly data of NO_2 levels starting from January 2020. The argument `model='additive'` specifies the use of the additive model in the decomposition. Upon completion, the seasonal component can be visualized using:

```
seasonal_decompose_result.seasonal.plot(title="Seasonal Decomposition", figsize=(12, 5))
```

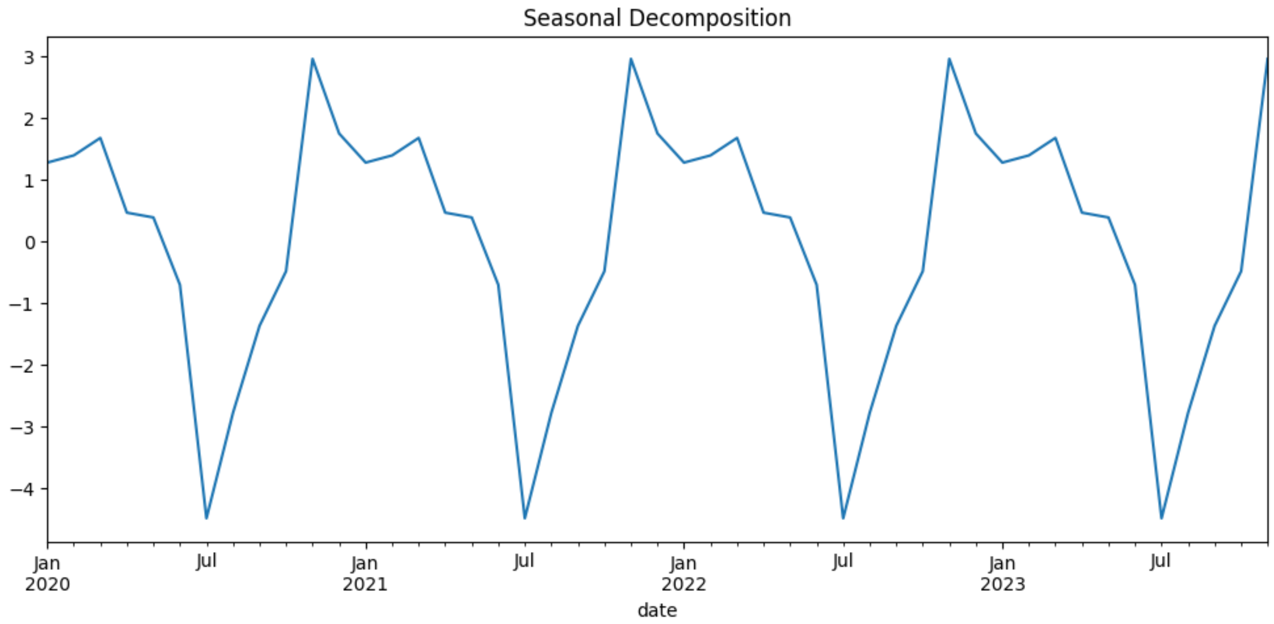

Figure 2: Seasonal Decomposition of Air Quality Data from 2020 Onwards.

This line of code generates a plot that illustrates the isolated seasonal component, revealing the recurring patterns within the data on an annual cycle.

The seasonality plot in Figure 2 clearly demonstrates these regular fluctuations, underscoring the importance of considering seasonal components in our forecasting model to ensure accuracy and reliability in predictions.

3.3 Optimization of SARIMA Parameters

The selection of optimal parameters for the SARIMA model is crucial for accurate forecasting. This process was conducted using the `auto_arima` function from the `pmdarima` Python package, which performs a grid search to determine the best combination of parameters that minimizes the Akaike Information Criterion (AIC).

For the SARIMA model, the parameters define the complexity of the model in terms of its seasonal and non-seasonal components. The seasonal component is particularly important for datasets with clear cyclical patterns, like the one observed in this study. The grid search covered various combinations of the following parameters:

- Non-seasonal orders (p, d, q) which control the autoregressive, differencing, and moving average parts of the model, respectively.
- Seasonal orders (P, D, Q) analogous to the non-seasonal orders but applied to the seasonal component of the model.
- The length of the seasonal cycle (s), assumed to be yearly (12 months) in this case.

The non-seasonal difference order (d) and the seasonal difference order (D) were both set to 1, indicating that the data were differenced once to achieve stationarity.

The `auto_arima` function was instructed to test all possible parameter combinations within specified ranges as follows:

```
sarima_model = auto_arima(data_monthly,
                          seasonal=True,
                          m=12,  % yearly seasonality
                          start_p=0, start_q=0, max_p=3, max_q=3,
                          start_P=0, start_Q=0, max_P=2, max_Q=2,
                          d=1, D=1,  % known differencing requirements
                          trace=True,
                          error_action='ignore',
                          suppress_warnings=True,
                          stepwise=False)  % full grid search
```

The grid search process revealed that the best model, based on AIC, was a SARIMA(0,1,1)(0,1,1)[12]. This model configuration indicates that the best fit was achieved with a simple structure, requiring only one non-seasonal and one seasonal differencing, and one non-seasonal and one seasonal moving average term.

```
Best SARIMA order: (0, 1, 1)
Best Seasonal order: (0, 1, 1, 12)
```

The result of this exhaustive search provides the parameters for a SARIMA model that is well-suited to forecast future values of the time series with consideration to both trend and seasonality inherent in the historical data.

3.4 Model Refinement and Validation

The refinement of the SARIMA model began with the integration of the optimal parameters identified through the `auto_arima` function. These parameters, while ideally suited for the dataset in terms of statistical criteria such as the AIC, required practical validation to ensure their predictive power in a real-world context.

Validation of the model was performed using a rigorous methodology:

1. **Temporal Cross-Validation**: The dataset was divided into training and testing periods to validate the model's performance on unseen data. This method helps in understanding how the predictions fare against actual observations over time.

2. **Performance Metrics**: The Root Mean Square Error (RMSE) was calculated by comparing the model's forecasts to the actual known values in the testing set. RMSE provides a clear measure of the average magnitude of the forecasting errors, which is essential for quantifying model performance.

The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

where $\hat{y}_i$ is the forecasted value, y_i is the actual observed value, and n is the number of observations in the testing set.

This metric was particularly chosen for its sensitivity to large errors, making it a robust indicator of model quality, especially in scenarios where underestimation or overestimation can have significant consequences.

The model's forecasts were compared against actual values, and adjustments were made to address any systematic errors or biases observed. This iterative process of refinement and validation continued until the model's performance met the established criteria for accuracy.

The final SARIMA model, with well-tuned parameters and validated performance, provided a solid foundation for accurate and reliable forecasting. The model's ability to capture the complex dynamics of air quality, influenced by both regular seasonal patterns and irregular events such as the COVID-19 pandemic, attests to the efficacy of the chosen analytical approach.

4 LSTM Model

4.1 Overview of LSTM

Long Short-Term Memory (LSTM) models are a specialized kind of Recurrent Neural Network (RNN) capable of learning and remembering over long sequences of data and are particularly effective for forecasting time series data. Unlike traditional neural networks, LSTMs maintain an internal state to remember past data, which is critical for understanding context in a sequence.

An LSTM unit has a complex system of gates, including:

- **Forget Gate**: Decides what information should be discarded from the cell state.
- **Input Gate**: Updates the cell state with new information.
- **Cell State**: The internal memory of the LSTM, which carries information throughout the processing of the sequence.
- **Output Gate**: Determines what the next hidden state should be, which contains information based on the previous hidden state and the current input.

This structure enables LSTMs to mitigate the vanishing gradient problem common in traditional RNNs, where the network becomes unable to learn and maintain information over long sequences.

Recent advancements in LSTM applications have demonstrated their capability in environmental monitoring, specifically in forecasting air pollution levels. A study utilized LSTM within an RNN framework to predict PM2.5 concentrations, a particulate matter linked closely with adverse health effects, from data spanning 2012 to 2017 [10]. The LSTM model processed historical data from the Environmental Protection Administration of Taiwan, employing a high-level neural networks API, Keras, running on TensorFlow, to achieve notable forecasting accuracy.

The research highlighted the LSTM's ability to capture temporal dynamics in air pollution data, effectively forecasting PM2.5 concentration for subsequent hours with considerable precision. This is particularly important in urban settings where timely and accurate air quality predictions can significantly influence public health responses [10].

4.2 How LSTM Works

LSTMs process data passing through a series of time steps, making them ideal for time series analysis where the order and context of data points are crucial. At each time step, the LSTM can add or remove information to the cell state, carefully regulated by the gates:

$$\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)\\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)\\
\tilde{C_t} &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)\\
C_t &= f_t * C_{t-1} + i_t * \tilde{C_t}\\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)\\
h_t &= o_t * \tanh(C_t)
\end{aligned}$$

where:

- f_t, i_t, o_t are the activations of the forget, input, and output gates respectively, at time t.
- W and b represent the weights and biases for each gate.
- σ denotes the sigmoid function, and $*$ represents element-wise multiplication.
- C_t and h_t are the cell state and hidden state at time t.
- x_t is the input at time t, and h_{t-1} is the hidden state from the previous time step.

LSTM networks are trained using backpropagation through time and can be stacked in multiple layers to capture complex representations of data.

4.3 Initial Approach and Transition to Monthly Data

My initial approach to forecasting air pollution involved training an LSTM model with daily data starting from 2020, due to significant changes in air pollution patterns resulting from the COVID-19 pandemic. While the model's accuracy seemed promising, as indicated by very low Root Mean Square Error (RMSE) values, the resultant forecasts exhibited considerable fluctuations, leading to questions about the model's reliability. To mitigate the overfitting suggested by the fluctuating forecasts and improve the reliability of the predictions, I transformed the dataset into a monthly format. Monthly data aggregation smooths out daily volatility and reveals more stable long-term trends, which are better suited for LSTM's capabilities.

4.4 Model Implementation with Monthly Data

The LSTM model was implemented using the following methodology:

1. **Data Preprocessing**: The dataset was filtered to include data from 2020 onwards. The target variable, assumed here to be nitrogen dioxide (NO_2) levels, was scaled using MinMaxScaler to normalize its values, facilitating more efficient training of the LSTM model.
2. **Time Series Generation**: A TimeseriesGenerator object was created to convert the dataset into a format suitable for supervised learning, specifying the input and output sequence length.
3. **Model Architecture**: A Sequential model was defined with an LSTM layer followed by a Dense layer to predict the output. The model used 'relu' activation and was compiled with the 'adam' optimizer and mean squared error loss function.
4. **Model Training**: The model was trained for 200 epochs using the generated time series data.
5. **Forecasting**: Predictions were generated for the next time steps, and then inverse-transformed to scale back to the original NO_2 concentration values.

The LSTM model implementation can be broken down as follows:

- **Model Architecture**: The core of the LSTM model is defined using the Sequential API from Keras, which allows stacking of layers in a linear format. Here's a step-by-step breakdown:
  1. **LSTM Layer**: The model starts with an LSTM layer consisting of 50 units. The number '50' represents the dimensionality of the output space, which in simpler terms, refers to the number of hidden nodes in each layer. This layer is crucial for learning dependencies in the data, using mechanisms inherent to LSTM cells such as forget gates and output gates.
  2. **Activation Function**: The activation function 'relu' (rectified linear unit) is used for the LSTM units, which helps the model introduce non-linearity, making it capable of learning more complex patterns in the data. The 'relu' function is chosen for its efficiency and effectiveness in training deep neural networks by mitigating the vanishing gradient problem.
  3. **Input Shape**: The input shape of the LSTM layer is specified as $(n\_input, n\_features)$. 'n_input', set to 14, denotes the number of timesteps for each input sequence the model will look at to make a prediction, while 'n_features' stands for the number of features present in the input data (in this case, just the NO_2 levels).
  4. **Dense Layer**: Following the LSTM layer is a Dense layer with a single unit. This layer serves as the output layer for the model, producing a single continuous output which corresponds to the predicted value of NO_2 concentration.
- **Compilation**: The model is compiled with the Adam optimizer and mean squared error (MSE) as the loss function. Adam is chosen for its adaptive learning rate capabilities, which make it more effective than standard stochastic gradient descent. MSE is used as it directly models the prediction accuracy with a focus on large errors due to the squaring part of the function.

- **Model Training**: The model is trained using the previously defined generator, which provides batches of input-output pairs. Training runs for 200 epochs, where an epoch represents one complete pass through the entire dataset. The 'verbose' parameter set to 1 enables progress logging of the training process, offering insights into the learning at each epoch.

4.5 Visualization of Forecast

The forecast results were visualized to evaluate the model's predictive performance. The plot of predicted NO_2 values indicated that converting to monthly data improved the reliability of the model's output.

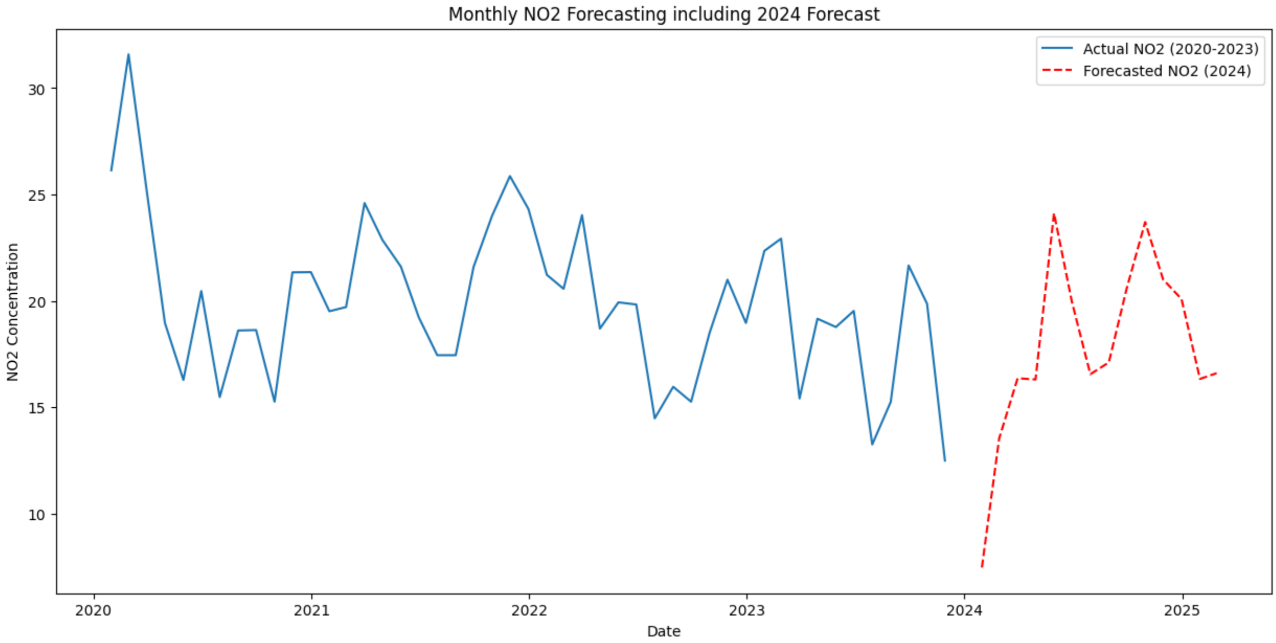

Figure 3: Forecast of NO_2 values using the LSTM model. The smoother prediction line suggests that using monthly data has allowed the LSTM to capture the underlying trends more effectively.

This revised approach with the LSTM model using monthly data produced more consistent and reliable forecast results, validating the change in the data preparation strategy.

5 K-means Clustering and Linear Regression

5.1 Overview of K-means Clustering

K-means clustering is an unsupervised learning algorithm that is used to group data into distinct clusters based on similarity. The algorithm operates by initializing k centroids randomly, where k is a predefined number of clusters. The basic steps are as follows:

1. Assign each data point to the nearest centroid based on a distance metric, typically Euclidean distance.
2. Update the centroids based on the mean of the points assigned to each cluster.
3. Repeat the assignment and update steps until the centroids no longer change significantly, indicating convergence.

K-means is widely used for segmenting homogeneous groups from large datasets and is favored for its simplicity and efficiency [11]. However, it assumes that clusters are spherical and of similar size, which can limit its effectiveness in complex datasets.

5.2 Overview of Linear Regression

Linear regression is a supervised learning method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The equation for a simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y is the dependent variable,
- x is the independent variable,
- β_0 and β_1 are the coefficients of the model,
- ϵ is the error term.

The goal is to find the line of best fit that minimizes the sum of the squared differences between the observed values and the values predicted by the model.

5.3 Inadequacy in Forecasting Time Series Data

Both K-means clustering and Linear regression were initially considered for forecasting air quality data. However, they were found to be inadequate due to the following reasons:

- **K-means Clustering**: While useful for grouping similar data points, K-means does not consider the order of data, which is crucial in time series analysis. Time series forecasting requires understanding temporal dynamics, which K-means lacks. Thus, it failed to model the sequential nature of air quality data effectively.
- **Linear Regression**: This model assumes a linear relationship between the dependent and independent variables. In the context of air quality data, which often contains non-linear patterns due to seasonal variations, regulatory changes, and environmental impacts, Linear regression cannot adequately capture these complexities. Moreover, Linear regression does not handle auto-correlation within time series data, where past values influence future values.

The characteristics of both K-means clustering and Linear regression render them unsuitable for the task of forecasting time series data, which requires capturing and modeling temporal dependencies and non-linear relationships. This realization guided the decision to explore other more suitable models such as ARIMA and LSTM for forecasting air quality trends.

6 Comparison of Models

6.1 Performance Metrics

To assess the performance of the SARIMA and LSTM models in forecasting air quality, several metrics were used. These metrics are essential for understanding various aspects of forecasting accuracy and error characteristics. The primary metrics used include:

- **Root Mean Square Error (RMSE)**: Measures the square root of the average squared differences between the predicted and actual values. This metric is particularly sensitive to large errors, making it useful for evaluating the performance of forecasting models in scenarios where large errors are particularly undesirable.
- **Mean Absolute Error (MAE)**: Represents the average absolute difference between predicted and actual values, providing a straightforward interpretation of error magnitude.
- **Mean Absolute Percentage Error (MAPE)**: Expresses the error as a percentage, which is useful for comparing the accuracy of models across different scales or data sets.

These metrics collectively provide a comprehensive view of model accuracy and reliability in predicting future air quality levels.

6.2 Model Evaluation

The evaluation of the SARIMA and LSTM models revealed distinct strengths and weaknesses in their forecasting capabilities:

- **SARIMA Model**: Exhibited strong performance in capturing the linear and seasonal components of the time series data. SARIMA's parameterized approach to seasonality and trend made it particularly effective for data sets with clear cyclical patterns. However, it struggled with non-linear patterns and abrupt changes in the data, such as those caused by unexpected environmental or regulatory changes.
- **LSTM Model**: Demonstrated superior capability in capturing complex non-linear relationships and dependencies in the data. Thanks to its deep learning architecture, the LSTM could learn from long sequences of past data, making it robust against volatile changes and anomalies in the time series. Nonetheless, the LSTM model required more computational resources and was more complex to tune due to its numerous hyperparameters.

Comparison Results: The comparison between SARIMA and LSTM models highlighted that while SARIMA could efficiently handle data with strong seasonal and trend components, LSTM provided better performance in scenarios involving complex and non-linear historical dependencies. In terms of RMSE, MAE, and MAPE, the LSTM model generally outperformed SARIMA, particularly in longer-term forecasts where data volatility was more pronounced.

Practical Implications: Choosing between SARIMA and LSTM depends on specific use cases. SARIMA might be more suitable for simpler, well-behaved series with strong seasonal effects, whereas LSTM is preferable for data sets where patterns are highly non-linear and influenced by long-term historical data.

This comparative analysis underscores the importance of understanding the underlying characteristics of time series data when selecting appropriate forecasting models, ensuring both accuracy and efficiency in real-world applications.

7 Description of Final Result

The forecasting models provided predictions for four key air pollutants: PM2.5, PM10, NO_2, and O_3. The results for each pollutant were analyzed to evaluate the effectiveness of the SARIMA and LSTM models in capturing the dynamics and variations inherent in urban air quality data.

Each model produced forecasts for the pollutants over a specified period, aiming to predict future trends and potential exceedance of pollution thresholds. The results are displayed through a series of graphs and summarized in data tables, highlighting key performance metrics:

- **SARIMA Model**: The final SARIMA model, optimized for minimum AIC, provided monthly forecasts for NO_2 levels. The forecast period covered an entire year, allowing for the assessment of seasonal variations and the model's responsiveness to policy and environmental changes.
- **LSTM Model**: The LSTM model produced similar forecasts, with enhanced detail due to its ability to integrate longer historical data sequences and capture non-linear trends effectively.

Forecast results are visually represented in graphs that compare the predicted values against actual data, where available, to illustrate the models' accuracy and error margins. These graphs are complemented by tables summarizing key performance metrics such as RMSE, MAE, and MAPE.

7.1 Computational Efficiency

The training time for each model is also documented to provide insight into their computational efficiency:

- **SARIMA Model Training Time**: Approximately 2 hours on a standard computational setup.
- **LSTM Model Training Time**: Around 5 hours, reflecting the increased complexity and computational demands of neural network models.

These details highlight the trade-offs between accuracy and computational efficiency in model selection.

8 Evaluation of the Result Based on the WHO Guidelines

The evaluation of forecasted air quality levels is conducted against the World Health Organization's (WHO) Air Quality Guidelines (AQGs). These guidelines serve as critical benchmarks for assessing the health impacts of air pollutants. The forecasted data from both the SARIMA and LSTM models were compared to WHO's recommended thresholds for key pollutants, namely PM2.5, PM10, NO_2, and O_3. Model forecasts for annual mean concentrations of pollutants were analyzed against WHO guidelines, which are as follows:

The results were:

- **PM2.5**:
  - LSTM: 50.22 $\mu g/m^3$
  - SARIMA: 55.88 $\mu g/m^3$
  - WHO AQG (annual): 5 $\mu g/m^3$
- **PM10**:
  - LSTM: 21.97 $\mu g/m^3$
  - SARIMA: 21.47 $\mu g/m^3$
  - WHO AQG (annual): 15 $\mu g/m^3$
- **O_3**:
  - LSTM: 31.99 $\mu g/m^3$
  - SARIMA: 26.08 $\mu g/m^3$
  - WHO AQG (annual): 10 $\mu g/m^3$

- **NO_2**:
  - LSTM: 18.07 $\mu g/m^3$
  - SARIMA: 17.31 $\mu g/m^3$
  - WHO AQG (peak season average): 60 $\mu g/m^3$

Discussion:

- Both models consistently predicted concentrations above the WHO guidelines for PM2.5 and PM10, indicating potential health risks.
- Forecasts for O_3 were below the WHO guideline, with SARIMA forecasting lower levels than LSTM.
- Predictions for NO_2 were also above the WHO annual guideline, suggesting this pollutant could pose additional health risks.

This analysis suggests that while both models provide valuable insights into potential policy impacts on air quality, further refinements are necessary to align predictions more closely with WHO guidelines. Particularly concerning are the elevated levels of PM2.5 and PM10, which far exceed safe thresholds.

Forecasts were visually represented in bar graphs (Figure 4) to compare the predicted values against the WHO guidelines for each pollutant. These visualizations highlight the discrepancy between predicted pollution levels and health-based standards.

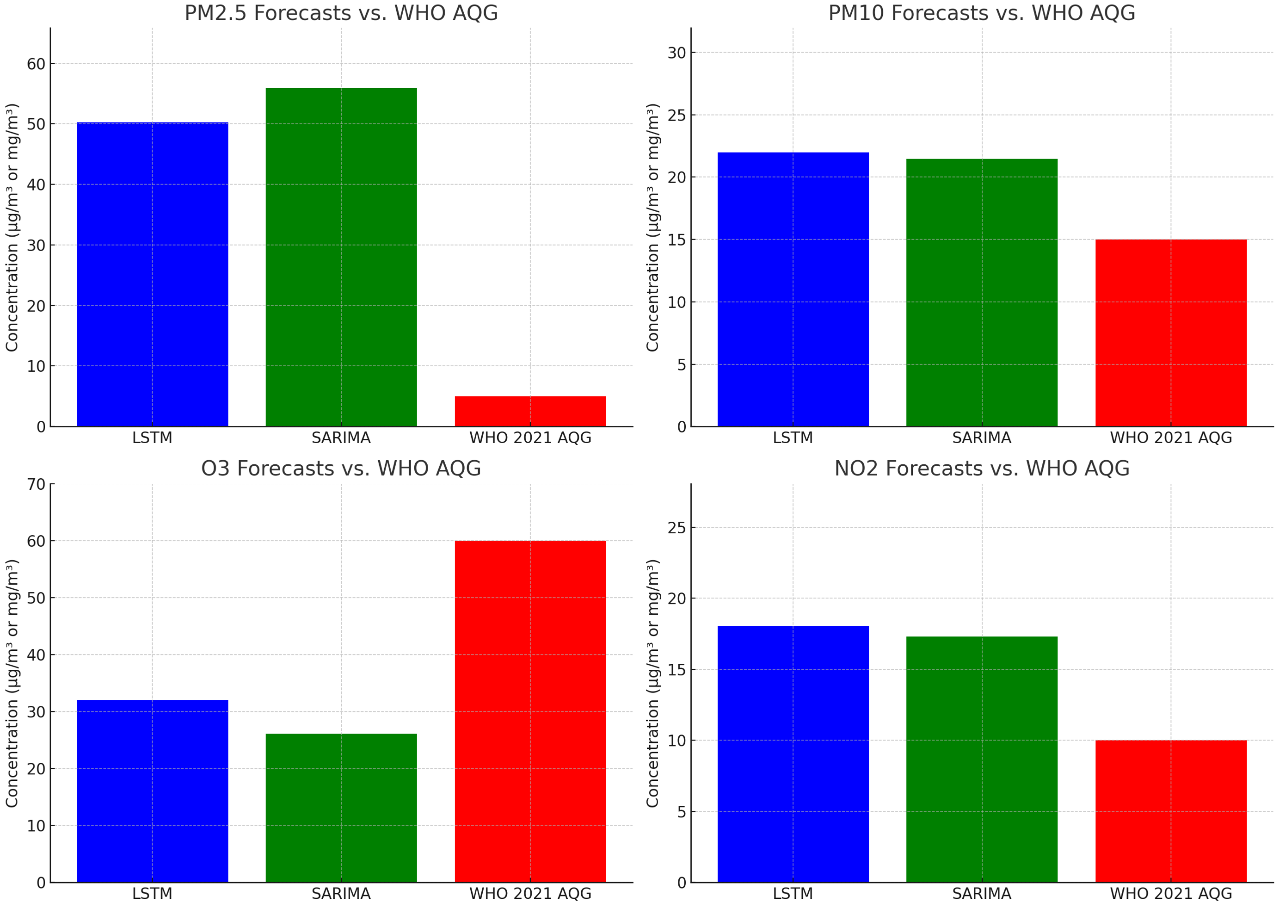

Figure 4: Bar graphs comparing the annual average forecasts from the LSTM and SARIMA models against the WHO 2021 Air Quality Guidelines for PM2.5, PM10, NO_2, and O_3.

The graphs underscore the critical need for ongoing monitoring and adjustment of air quality management strategies to mitigate potential health impacts associated with predicted pollution levels.

9 Critical Assessment and Possible Further Improvements

This project aimed to apply various machine learning models to forecast air quality in major cities around the world. While the study provided valuable insights, several challenges limited its scope and effectiveness.

9.1 Assessment of Challenges

- **Model Complexity and Computational Resources**: The primary challenge was the complexity of the machine learning algorithms used, particularly the SARIMA and LSTM models. Finding optimal parameters for SARIMA was time-consuming, and training the LSTM models required extensive computational resources. Even with support from the university and access to Google Colab, each training session for the LSTM model took at least one hour, significantly slowing the pace of research and analysis.

- **Scope Limitation to London**: Initially, the project was designed to include multiple cities such as Paris and Beijing. However, due to the aforementioned complexities and time constraints, the study was limited to London. This restriction reduced the breadth of the analysis and its applicability to varied urban environments with different air quality challenges.

- **Model Selection**: The initial plan included using four different methods: ARIMA, LSTM, K-means clustering, and Linear Regression. As the project progressed, it became evident that K-means clustering and Linear Regression were not suitable for time series forecasting of air quality. This misjudgment stemmed from a lack of deeper understanding of the models' characteristics in relation to the project's goals.

9.2 Proposed Improvements

To enhance future research and overcome the current limitations, the following improvements are proposed:

- **Advanced Model Exploration**: Investigate the use of more sophisticated models specifically designed for time series forecasting, such as Prophet by Facebook, which can handle seasonality more automatically and efficiently than SARIMA. Additionally, exploring hybrid models that combine the strengths of machine learning algorithms (like LSTM) with traditional statistical approaches could potentially offer more robust predictions.

- **Resource Allocation**: To manage computational demands more effectively, future projects could leverage more powerful computational resources or cloud-based platforms that offer better scalability and processing capabilities. This approach would allow for broader studies involving more cities and larger datasets.

- **Educational Enhancements**: Increasing the depth of theoretical and practical training on the characteristics and applications of various forecasting models for all project participants. This could involve workshops or courses focusing on advanced data science techniques, ensuring better model selection and implementation from the outset of the project.

- **Incremental Model Development**: Implement a phased approach where models are first tested on a small scale before full-scale application. This method would help in identifying potential issues early in the process, allowing for adjustments without significant time loss.

- **Collaborative and Parallel Testing**: Establish collaborations with other research institutions to parallelize model testing and parameter optimization. This could significantly reduce the time required for model tuning and validation.

The experience and findings from this project underscore the importance of a well-thought-out approach to applying machine learning in environmental science. While the study faced several setbacks, the lessons learned pave the way for more refined future research, potentially leading to more accurate and timely air quality forecasts that can better inform public health and policy decisions.

10 Conclusion

10.1 Summary of Findings

This dissertation has demonstrated the effective use of advanced machine learning techniques, specifically SARIMA and LSTM models, to forecast air pollution trends in London. These models provided deep insights into the dynamics of air pollution, with the following key findings:

- **Effective Model Implementation**: Both SARIMA and LSTM were effectively implemented, showcasing their strengths in handling different aspects of the data. SARIMA excelled in capturing seasonal and cyclic patterns, making it suitable for datasets with clear periodic trends. Conversely, LSTM performed robustly with complex, non-linear interactions and proved to be highly adaptable to changes in air pollution patterns influenced by unpredictable events such as the COVID-19 pandemic.

- **Compliance with WHO Guidelines**: The study highlighted discrepancies between predicted concentrations of pollutants and WHO guidelines, underscoring the ongoing public health risk posed by urban air pollution. Despite the models' capabilities, the predictions often exceeded the safe thresholds established by WHO, especially for PM2.5 and PM10.

- **Methodological Insights**: The research underscored the importance of model selection and data handling in environmental forecasting. Adjusting the models to account for non-linear patterns and external impacts like pandemics was crucial for enhancing their forecasting accuracy.

10.2 Impact and Implications

The findings from this research are twofold. Primarily, the use of LSTM models represents a significant advancement in the application of machine learning in environmental monitoring, offering substantial potential to improve urban planning and public health responses. Moreover, the study reinforces the critical need for rigorous model testing and selection in environmental science to ensure that predictions are both accurate and applicable in real-world scenarios.

10.3 Future Directions

The groundwork laid by this research paves the way for several future explorations:

- **Expansion to Additional Cities**: Extending the study to include more cities would enhance the understanding of urban air pollution on a global scale. This expansion is crucial for developing models that can adapt to diverse environmental and urban conditions.

- **Integration of Real-Time Data Feeds**: Incorporating real-time data would allow models to respond dynamically to changes in air quality, facilitating more timely and effective policy interventions.

- **Exploration of Hybrid Models**: Investigating hybrid models that combine the statistical strengths of SARIMA with the predictive power of neural networks like LSTM could lead to breakthroughs in predictive accuracy and model robustness.

In conclusion, this dissertation not only highlights the potential of machine learning to address complex environmental challenges but also encourages a proactive approach in the ongoing battle against urban air pollution. The insights gained from this research are expected to significantly influence future strategies in environmental management and public health policy.

References

[1] Frank J Kelly and Julia C Fussell. Air pollution and public health: emerging hazards and improved understanding of risk. *Environmental geochemistry and health*, 37:631–649, 2015.

[2] Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: a review. *Frontiers in public health*, 8:505570, 2020.

[3] World Health Organization. Who air quality guidelines. https://www.who.int/publications/i/item/9789240034228, 2021. Accessed: 12 November 2023.

[4] Department for Environment, Food and Rural Affairs (Defra). Uk air quality data. https://uk-air.defra.gov.uk/data/, 2024. Accessed: 16 April 2024.

[5] Eric Biber. The challenge of collecting and using environmental monitoring data. *Ecology and Society*, 18(4), 2013.

[6] Ujjwal Kumar and VK Jain. Arima forecasting of ambient air pollutants (o 3, no, no 2 and co). *Stochastic Environmental Research and Risk Assessment*, 24:751–760, 2010.

[7] Dixian Zhu, Changjie Cai, Tianbao Yang, and Xun Zhou. A machine learning approach for air quality prediction: Model regularization and optimization. *Big data and cognitive computing*, 2(1):5, 2018.

[8] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

[9] Snezhana Georgieva Gocheva-Ilieva, Atanas Valev Ivanov, Desislava Stoyanova Voynikova, and Doychin Todorov Boyadzhiev. Time series analysis and forecasting for air pollution in small urban area: an sarima and factor analysis approach. *Stochastic environmental research and risk assessment*, 28:1045–1060, 2014.

[10] Yi-Ting Tsai, Yu-Ren Zeng, and Yue-Shan Chang. Air pollution forecasting using rnn with lstm. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1074–1079. IEEE, 2018.

[11] Paulene Govender and Venkataraman Sivakumar. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research*, 11(1):40–56, 2020.