Seasonal Impact on Air Quality Levels in London: Analysis of Air Pollution Trends and Sources

Riya Dodthi
220266251
Dr. Miles Hansard
Msc. Big Data Science

***Abstract*—This research project delves into the intricate interplay between seasonal variations and air quality levels within the bustling metropolis of London. The central hypothesis underpinning this study is that the cyclical changes in weather and atmospheric conditions throughout the year can significantly influence the concentrations of various air pollutants. Through a meticulous analysis of extensive historical air quality data sourced from diverse boroughs across London, an array of pollutants including PM2.5, PM10, Nitrogen oxides and SO2 are scrutinized. Concurrently, meteorological parameters encompassing temperature, humidity, global radiation and other parameters are incorporated to unravel potential underlying correlations and discernible trends.**

By unravelling the nuanced relationship between London's fluctuating seasons and its air quality dynamics, this research aspires to furnish crucial insights that inform targeted and contextually relevant strategies for effective pollution management. The implications of this work resonate profoundly within the spheres of urban planning, policy-making, and public health. Ultimately, a comprehensive understanding of how London's air quality responds to seasonal shifts can empower stakeholders with the knowledge necessary to cultivate a healthier, more sustainable, and environmentally conscious urban landscape.

I. Introduction

Air quality is a critical aspect of urban living, directly influencing public health, environmental sustainability, and overall quality of life. In rapidly growing cities like London, where a dense population coexists with a diverse range of economic and industrial activities, maintaining a healthy atmosphere is of paramount importance. The presence of air pollutants, such as particulate matter (**PM_{10}** and **$PM_{2.5}$**), nitrogen oxides (**NO_x**), and sulfur dioxide (**SO_2**), poses significant challenges to achieving optimal air quality standards.

Previous studies have reported increased mortality directly related to the modification of temperature patterns induced by climate change (Ren et al., 2011; Vardoulakis et al., 2014; Carmona et al., 2016; Lee et al., 2018) In addition, climate change influences air quality (Jacob and Winner, 2009) as ambient air pollutants are very sensitive to meteorological conditions (Elminir, 2005; de la Paz et al., 2016; Westervelt et al., 2016; Chen et al., 2018a) and exposure to airborne pollutants is also a leading contributor to global disease burden (Lelieveld et al., 2015; Cohen et al., 2017). Recent studies have suggested that two-way interactions between weather variables (e.g., temperature) and air pollution should be carefully considered to characterize synergistic effects on health (Stafoggia et al., 2008; Chen et al., 2018b)

One of the lesser-explored dimensions in the realm of air quality research pertains to the potential impact of seasonal variations on pollutant concentrations in London. Climate, meteorological conditions, and varying sources of emissions can all contribute to shifts in air quality levels across different seasons. Investigating the relationship between seasons and air quality can provide invaluable insights into understanding the dynamics of pollution in urban environments and subsequently aid in developing targeted strategies for mitigation.

A. Particulate Matter (PM_{10} and $PM_{2.5}$)

Particulate matter (PM) consists of tiny solid particles and liquid droplets suspended in the air. It is categorized based on size, with PM_{10} referring to particles with a diameter of 10 micrometres or smaller, and $PM_{2.5}$ referring to particles with a diameter of 2.5 micrometres or smaller. These particles can originate from various sources, including vehicular emissions, industrial processes, construction activities, and natural sources. Both PM_{10} and $PM_{2.5}$ have been associated with adverse health effects, including respiratory and cardiovascular diseases, due to their ability to penetrate deep into the respiratory system.

B. Nitrogen oxides (NO_x)

Oxides of nitrogen (NOx) encompass a group of compounds comprising nitrogen and oxygen molecules. The primary oxides include nitric oxide (NO), formed during combustion processes; nitrogen dioxide (NO2), a reddish-brown gas originating from the oxidation of nitric oxide; and Nitrous Oxide (N2O), a greenhouse gas contributing to climate change. Notably, NO2 is a significant component of urban air pollution, with implications for health and environmental quality. These compounds are byproducts of combustion, emitted from sources like vehicles and industrial activities. Given their adverse effects on air quality, the environment, and human health, understanding and mitigating NOx emissions remain critical.

C. Sulfur Dioxide (SO_2)

Sulfur dioxide (SO_2) is produced by the burning of fossil fuels containing sulfur, such as coal and oil, in power plants and industrial facilities. SO_2 is known to irritate the respiratory system and exacerbate asthma and other respiratory conditions. It can react in the atmosphere to form fine particulate matter, contributing to the overall pollution load.

This study aims to investigate the potential impact of seasonal variations on air quality levels in London, focusing on pollutants like PM_{10}, $PM_{2.5}$, NO_x, and SO_2. Utilizing historical air quality data from monitoring stations across diverse city environments, including Roadside and Background locations, the analysis will encompass various pollutants and meteorological parameters such as temperature, humidity, and global radiation. The findings of this research could provide valuable insights for urban planners, policymakers, and public health officials, shedding light on the effectiveness of current strategies and the need for tailored interventions to address air quality challenges in the city, ultimately contributing to the overall well-being of its residents.

II. Related Work

The literature review for this project reveals a substantial body of research examining the complex interplay between seasonal variations and air quality levels in urban environments. Numerous studies have contributed to our understanding of how changing weather conditions can significantly influence pollution dynamics.

In their study, Chen et al. (2013) investigated the seasonality in the association between particulate matter with an aerodynamic diameter of less than 10 μg (PM10) and daily mortality in 17 Chinese cities. The researchers fitted the "main" time-series model after adjusting for time-varying confounders using smooth functions with natural splines. Additionally, a "seasonal" model was established to obtain the season-specific effect estimates of PM10. The observed seasonal pattern remained relatively consistent across various model specifications. The findings from their analysis suggested that the acute effect of particulate air pollution on mortality could vary by season, with the most pronounced effects occurring during winter and summer in China (Chen et al., 2013).

In another study by Liu et al. (2020), the authors explored the impact of intense human activities and adverse meteorological conditions on air quality and human health. The concentration of air pollutants at most monitoring stations exhibited significant negative correlations with wind speed, precipitation, and relative humidity while displaying a positive correlation with atmospheric pressure. Moreover, as latitude increases, the influence of temperature on air pollutant concentration becomes more prominent. The study underscores the urgency of revealing the relationship between air pollution and meteorological conditions using long-term daily or real-time data for effective pollution control.

According to Ailish M et al.(2020) daily mean PM2.5 observations from 42 UK background sites indicate that easterly, south-easterly and southerly wind directions and anticyclonic circulation patterns enhance background concentrations of PM2.5 at all UK sites by up to 12 μg/m$^-3$. Results from back trajectory analysis and the European Monitoring and Evaluation Programme for UK model (EMEP4UK) show this is due to the transboundary transport of pollutants from continental Europe.

Barmpadimos et al(2012) examined PM10, PM2.5, and PM coarse concentration trends in European urban and rural areas from 1998 to 2010. Generalized Additive Models link meteorological variables to PM levels, emphasizing wind speed, wind direction, boundary layer depth, precipitation, temperature, and weather patterns. Complex relationships between temperature and PM2.5, PM coarse, and wind speed were observed. Meteorologically adjusted PM time series displayed decreasing trends for PM10 and PM2.5 concentrations. Quantile regression revealed significant reductions in very large PM concentrations. Overall, meteorological factors play a vital role in shaping PM concentrations and trends.

In conclusion, the extensive body of literature underscores the intricate connections between weather conditions and air quality in urban environments. This serves as the bedrock for the present research, which aims to untangle the nuanced relationships between seasonal variations and pollution levels in the dynamic urban landscape of London.

III. Methodology

Visualization: Visualization is vital for effective data communication. It presents insights clearly and appealingly, aiding decision-making and information dissemination. This system offers effective visuals, facilitating data exploration and conveying air pollution data, analyses, and recommendations to stakeholders and policymakers.

Predictive modeling:
This project involves utilizing advanced algorithms to predict air quality levels using historical data and meteorological variables. By establishing intricate relationships between pollutants and environmental factors, predictive models offer insights into seasonal variations and air quality dynamics. This approach enhances pollution pattern understanding and equips stakeholders with predictive tools for decision-making, urban planning, and policy formulation. Through time-series forecasting and data-driven predictions, this project aims to contribute to a sustainable and health-conscious urban environment in London.

By meeting these requirements, the system will enable effective air pollution analysis, monitoring, and decision support, empowering stakeholders to implement targeted interventions and policies to improve air quality and protect public health in London.

Data Collection:
In the data collection stage of this project, comprehensive data on air pollution in London is gathered through various sources. Air quality monitoring stations strategically placed

across the city provide real-time measurements of pollutants. These stations enable the assessment of pollution levels and spatial variations across different areas of London. Additionally, satellite imagery is employed to complement ground-based measurements, allowing for a broader understanding of pollution patterns and the identification of hotspots.

Meteorological data, including information on pressure, temperature inversions, and precipitation, is collected to examine the influence of weather conditions on air pollution. This data helps analyze the dispersion and accumulation of pollutants in different atmospheric conditions.

Supplementary data sources, such as traffic data, industrial emissions records, and population density information, are also considered to gain a comprehensive understanding of pollution sources and their contributions.

By incorporating data from multiple sources, the data collection stage ensures a robust dataset for subsequent analysis. This dataset serves as the basis for informed decision-making, policy formulation, and the development of effective strategies to address air pollution in London.

Dataset: The dataset used for this analysis includes emissions estimates of key pollutants (NOx, PM10, PM2.5, and SO2) by source type for the base year 2019, along with detailed road transport emissions data. It also provides modelled 2019 ground-level concentrations of annual mean NOx, NO2, PM10, and PM2.5 at a 20m grid resolution, including the number of daily means exceeding 50 μg/m^3 for PM10. Population exposure estimations for NO2 and PM2.5 concentrations above specific thresholds are included (GLA, 2019).

Additionally, London's Average Air Quality Levels dataset offers monthly averages for pollutant concentrations at roadside and background locations (King's College London, 2019).

The weather dataset compiles various attributes, such as date, cloud cover, temperature, precipitation, and more, recorded at a weather station near Heathrow Airport (EMMANUEL F. WERR, 2021). These datasets collectively contribute to a comprehensive analysis of air quality and meteorological conditions in London.

A. *Data Visualization:*

Temporal analysis

The visualization presented showcases the temporal variations of air quality in both roadside and background environments in London. The data utilized in this visualization encompasses the summation of mean concentrations of pollutants, including Nitrogen Oxides, PM10 and PM2.5 particles, and Sulphur Dioxide, measured in micrograms per cubic meter μg/m^3. The measurements were captured at different time intervals throughout the day, month and year from 2008 to 2019.

By comparing the air quality data from roadside and background environments, the visualization provides insights into the differences and similarities in pollutant levels at different locations within the city. The temporal aspect allows for the examination of pollution patterns and fluctuations throughout months and years providing valuable information on peak pollution periods and potential sources of pollution.

This visualization serves as a valuable tool for understanding the temporal dynamics of air pollution in London, aiding in the identification of pollution patterns, the assessment of pollutant sources, and the formulation of targeted mitigation strategies.

Diagnostic/Exploratory analysis

This type of analysis provides insights into the contributions of different sources to pollutant levels and the total pollution attributed to specific sectors, all measured in µg/m³. This type of analysis aims to uncover the causes and factors that contribute to the problem at hand, shedding light on the key drivers of pollution in the given context.

By analyzing the contribution of each source category to pollutant levels, the visualizations provide valuable insights into the relative significance of different emission sources. This information can help identify the most influential contributors to pollution and guide targeted mitigation strategies.

Model Development

Temporal visualizations reveal noticeable patterns in average pollutant levels across months, indicating a potential link between seasonal variations and pollution. To delve deeper and forecast pollution levels, a dual approach involving time-series analysis and predictive modelling is adopted. Through time-series analysis, underlying temporal trends are unravelled. Additionally, a predictive model incorporating weather attributes and pollutant concentrations is developed. This model aims not only to illuminate seasonal pollution impacts but also to foresee upcoming pollution levels. This combined effort enriches the comprehension of pollution dynamics, contributing significantly to the formulation of informed strategies for pollution management and mitigation.

Dataset:

The analysis will draw from two primary datasets: meteorological data encompassing various weather parameters, and London's average pollution dataset, which spans both roadside and background areas. The objective is to investigate potential relationships between distinct weather attributes such as sunshine, cloud cover, and temperature, and the corresponding levels of pollutants, representing the primary focus of inquiry.

However, preceding the commencement of the actual analysis, a preliminary step demands careful consideration. Ensuring meticulous data preparation and systematic organization is essential. This preparatory phase assumes significance as it establishes a robust foundation for the ensuing analytical procedures.

This necessitates a series of steps aimed at refining and structuring the data:

1) **Data Cleaning:** Commencing the investigative journey, the refinement process shall be initiated by judi-

ciously eliminating extraneous columns from the dataset. Among these, the time column adjudged non-essential for immediate analysis shall be excluded to enhance the dataset's simplicity and facilitate subsequent data management. Addressing the challenge of missing values, a fundamental concern in maintaining data integrity, a proactive stance shall be adopted by imputing absent values with zeros. This proactive approach is indispensable to ensure the thoroughness and robustness of the dataset, thereby reinforcing the foundation for analytical endeavours.

Further enhancing the dataset's consistency and compatibility by making the representation of temporal information more uniform. This involves converting date columns into the DateTime format, which creates a consistent timeline. This streamlined approach makes it easier to compare data across different time periods, enhancing our understanding of the information.

In summation, these preliminary measures, encompassing the elimination of redundancies, the imputation of missing values, and the standardization of temporal data, serve as the cornerstone for an insightful and methodologically rigorous analytical exploration. Collectively, they contribute to the scholarly robustness of the research endeavour, reinforcing its reliability and scholarly significance.

2) **Data Transformation:** To attain a comprehensive overview on a monthly scale, the computation of average pollutant concentrations for each month becomes imperative. This step holds particular significance owing to the existence of multiple entries for the same month within the dataset, corresponding to diverse time intervals. In parallel, the calculation of average values for meteorological attributes on a monthly basis is also of paramount importance.

   Subsequent to this, the aggregation of total pollution concentration data shall be executed independently for both roadside and background locations. The amalgamation of this data with the weather dataset will be facilitated, employing the date column as a linking mechanism.

   Such an approach is instrumental in ensuring the holistic encapsulation of pollutant trends concomitant with the associated weather conditions. The synthesis of these datasets endeavours to unveil potential correlations existing between pollutants and meteorological elements, thereby illuminating the reciprocal influences these factors exert upon one another.

3) **Data Integration:** Data integration is pivotal in analysis, merging diverse sources for comprehensive insights. Combining weather and pollution data enables correlated exploration, unveiling latent connections. This approach is crucial for comprehensive insights that might be missed when studying datasets individually.

By traversing through these procedural stages, a robust foundation for a successful analysis is established. While it may not carry the allure of more captivating segments, this preparatory phase is pivotal in ensuring the data's integrity before drawing deductions from it.

Time-Series Analysis - SARIMA Model:

Time-series analysis of air pollution environmental levels involves the identification of long-term variation in the mean (trend) and of cyclical or periodic components (Salecedo et al.,1999). The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a time series forecasting technique designed to capture and model the complex patterns and trends present in time-dependent data. SARIMA extends the standard ARIMA (Autoregressive Integrated Moving Average) model by incorporating seasonality components.

Components of the SARIMA Model:

1. AR Component: $AR(p)$ past observations' influence.
2. I Component: $I(d)$ differencing for stationarity.
3. MA Component: $MA(q)$ past errors' impact.
4. Seasonal AR Component: $SAR(P)$ lagged values over seasons.
5. Seasonal I Component: $SI(D)$ seasonal differencing.
6. Seasonal MA Component: $SMA(Q)$ past errors over seasons.

Parameters: p, d, q, P, D, Q, s.

Effective for complex time series with seasonality.

Multicollinearity:

To systematically detect and mitigate the potential challenges arising from multicollinearity, a preliminary step involves the construction of a correlation matrix. This matrix serves as a diagnostic tool to unveil the interrelationships among features within the dataset. the focus lies particularly on identifying features that exhibit substantial correlations exceeding the threshold of 85%.

Subsequent to this identification process, a judicious curation of features ensues. Those exhibiting strong interdependence, surpassing the established correlation threshold, undergo thoughtful elimination. This discerning culling mitigates multicollinearity concerns, enhancing the integrity of subsequent analysis.

For instance, Figure 1 highlights the presence of collinearity exceeding 0.85 among features such as sunshine, global radiation, and the trio of max, mean, and min temperature. This redundancy suggests that their collective impact on the label is redundant. Thus, a logical approach involves removing the least influential feature within this subset, optimizing the model's efficiency while maintaining its efficacy.

Additionally, attention is directed towards attributes characterized by weak correlations with the label, falling within the range of -0.1 to 0.1. These attributes contribute minimally to the analytical landscape. A purposeful elimination of such attributes is warranted to enhance the analysis's coherence, reduce noise, and refine insights without compromising the overall perspective.

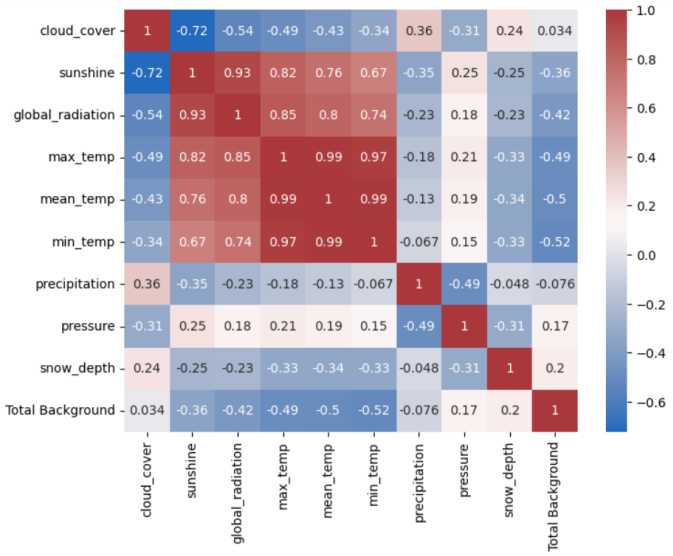

Fig. 1. Correlation matrix between weather features and background level where red and blue denote high positive and negative correlation respectively.

Outlier Detection:

Moreover, scatter plots are employed to methodically scrutinize the distribution of data points concerning both features and their corresponding labels within the training dataset. This procedural step aids in the identification of potential outliers.

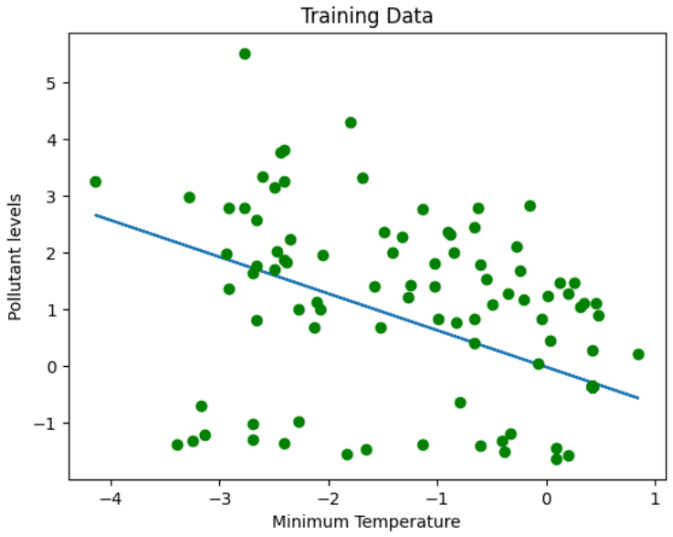

Fig. 2. Linear regression on the data using the least squares method to fit a linear curve to the relationship between the "Minimum Temperature" (x) and "Roadside Pollutant levels" (y)

Figure 2 applies linear regression using least squares to model the relationship between "Minimum Temperature" (x) and "Roadside Pollutant levels" (y). Outliers can impact slope and intercept calculations, as the line minimizes squared residuals. Outliers with large residuals can distort the line's fit by exerting a pulling effect.

To mitigate this, it is suggested to address these outliers. Identifying them enables informed decisions, including their potential removal or adjustment. Such actions enhance model robustness, aiding effective generalization across the dataset and fortifying the analytical process.

Fundamentally, this process centres on achieving precision and clarity. Through meticulous feature selection, the objective is to disentangle intricate relationships, uphold analytical precision, and underscore significant patterns that align harmoniously with the primary objective.

By starting with these initial steps, the data is readied for analysis. This readiness enables a thorough exploration of how weather variables and pollution levels connect, uncovering insights that enhance the understanding of air quality in London.

Furthermore, instead of manually handling feature selection and multicollinearity, an alternative approach is available: considering the ElasticNet Regression model.

Elastic Net regression combines techniques from Lasso and Ridge regularization. This mix provides a well-rounded approach to linear regression, making the model more stable, helping with feature selection, and managing multicollinearity issues.

The core of Elastic Net regression is minimizing this cost function:

$$\text{Cost}(w) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \cdot \left(\frac{1-\rho}{2}\sum_{j=1}^{p} w_j^2 + \rho \sum_{j=1}^{p} |w_j| \right) \quad (1)$$

Where:

- n represents the number of data points,
- p signifies the number of features (predictors),
- y_i denotes the actual target value for the i-th data point,
- $\hat{y}_i$ signifies the predicted value for the i-th data point using the Elastic Net model,
- w_j signifies the weight (coefficient) of the j-th feature,
- α characterizes the regularization parameter, which determines the overall strength of regularization,
- ρ controls the mixing parameter, striking the balance between the L1 and L2 penalties.

The Elastic Net's cost function encompasses two essential elements: the Mean Squared Error (MSE), gauging model prediction accuracy, and the regularization term merging L2 (squared magnitude) and L1 (absolute magnitude) penalties. The mixing parameter ρ significantly shapes the interplay between these penalties.

Through iterative optimization, Elastic Net determines coefficients w_j that minimize the cost function, achieving dual objectives: accurate data fitting and counteracting overfitting via regularization. Fine-tuning hyperparameters α and ρ empowers practitioners to precisely control regularization and L1-L2 balance, facilitating customized model behaviour tailored to data traits and research objectives.

In summary, Elastic Net regression offers a comprehensive and versatile framework for linear regression, adeptly navigating the challenges of multicollinearity, feature selection,

and model stability. Its mathematical foundation, characterized by the interplay of L1 and L2 regularization, empowers researchers to craft models that strike a harmonious balance between complexity and simplicity.

Modelling- The initial step encompasses the division of the dataset into two distinct portions: a training set comprising 70% of the data and a validation set containing the remaining 30%. In order to enhance the model's efficacy and foster stable training, a pipeline is constructed, incorporating three primary stages. The GridSearch() technique is then harnessed to systematically explore a range of hyperparameters for the models employed within this pipeline:

- **Standardization:**
  For a regression model with features on very different scales, standardization using the StandardScaler() is generally a better choice than normalization. Standardization ensures that features have a mean of 0 and a standard deviation of 1, which is achieved using the following equation:

$$x_{\text{standardized}} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

  This approach helps optimization algorithms converge faster, provides robustness to outliers, maintains interpretability, and ensures consistent algorithm performance across features with varying scales.

- **Polynomial regression:**
  Polynomial Regression() is then applied which fits a curved line to data by adding polynomial terms to the linear regression equation. It's useful for modelling nonlinear relationships between variables. The equation becomes:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

  Where n is the degree of the polynomial. It captures bends and curves in the data but can overfit if not controlled.
- **Regression Model:**
  The final stage in the pipeline would be the ElasticNet Regression model which would take the the input data that has passed through previous stages of the pipeline and provide an output of predicted label in this case the pollutant levels.

Residual Outlier Detection:
Following the model's training, the identical training features are employed for the prediction of the corresponding training labels. This procedural stride endeavours to discern and mitigate any outliers within the data, preempting the model from learning from data instances that are unreliable or erroneous. This entails the computation of residuals through the subtraction of the actual training labels from the predicted training labels. Subsequently, the z-scores of these residuals are computed. Through the establishment of a predetermined threshold, residual records that surpass this threshold are identified and subsequently eliminated, effectively eradicating outliers from the training dataset.

Following the elimination of outliers (as depicted in Figure 3), the subsequent course of action involved the retraining of the model, utilizing the refined training dataset. This iterative training phase is instituted to safeguard the model's performance from the detrimental impact of outliers, ensuring its efficacy in subsequent predictions. Ultimately, the model's performance is assessed on the validation dataset, gauging its capacity for generalization and its aptness in making predictions on unseen data instances.

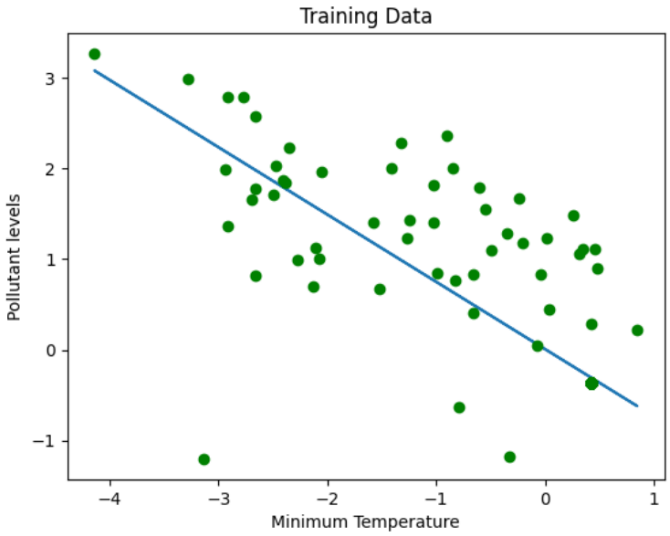

Fig. 3. After eliminating outliers, the regression line tends to better represent the underlying relationship between variables. It becomes less influenced by extreme data points, resulting in a more accurate depiction of the overall trend.

IV. Results

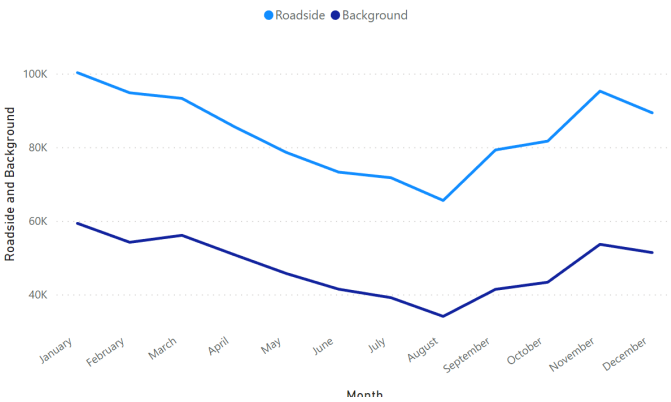

Fig. 4. Average monthly pollutant trends near roadside and background locations.

Visualization:
In Figure 4, the investigation into average pollutant levels

across both Roadside and Background environments reveals a consistent and significant trend. Across the analysis, the concentrations of pollutants exhibit a propensity to rise during colder months and diminish during warmer periods.

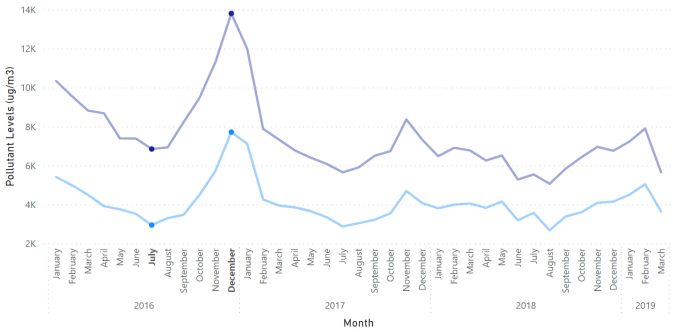

Fig. 5. A more in-depth monthly trend of pollutant concentrations over the years shows a cyclic pattern near roadside and background locations, with concentrations being lower during warmer months and higher during colder ones.

Moreover, this recurring pattern extends from 2008 to 2019, as depicted in Figure 5. For instance, in July 2016, the average pollutant levels measured 2,948 µg/m³ for the Background environment as compared to December when it increased to 7,719µg/m³. This consistent trend persists across subsequent years, including 2017 and 2018, underscoring the potential influence of seasonal dynamics on pollution levels.

These findings underscore the importance of studying the interplay between weather changes and air quality fluctuations. This understanding provides insights into pollution variations, aiding in the development of strategies to manage pollution effectively. By deciphering seasonal impacts, cities can enhance environmental quality and public health.

In the subsequent phase, attention is directed towards the identification of primary contributors to pollutant concentrations across distinct boroughs. For this purpose, Python's GeoPandas library is employed, streamlining the management of geographic data within the familiar panda's framework. This library furnishes tools for handling shapes and executing spatial operations such as data combination and overlay. Augmented by visualization capabilities, GeoPandas enriches the efficacy of the geospatial data exploration and analysis.

In the analysis conducted, it becomes evident that Hillingdon portrays the highest levels of pollutants for PM10, PM2.5, NOx, and SO2, as illustrated in Figure 6.

Subsequently, a focused inquiry is made into the predominant sources of pollutants within this borough. This endeavour facilitates an understanding of the fundamental factors and contributes to the informed formulation of policy decisions.

In Hillingdon (Fig. 7), the primary sources of PM10 and PM2.5 are linked to industrial and commercial activities, notably construction and demolition dust, along with Part A1/B installations encompassing industrial processes like refineries and factories and Aviation being the primary source for SO2. Similarly, for NOx (Fig. 8), transport sources—especially Aviation, diesel cars and industrial sources like Oil/Coal combustion are the significant contributors to pollution levels.

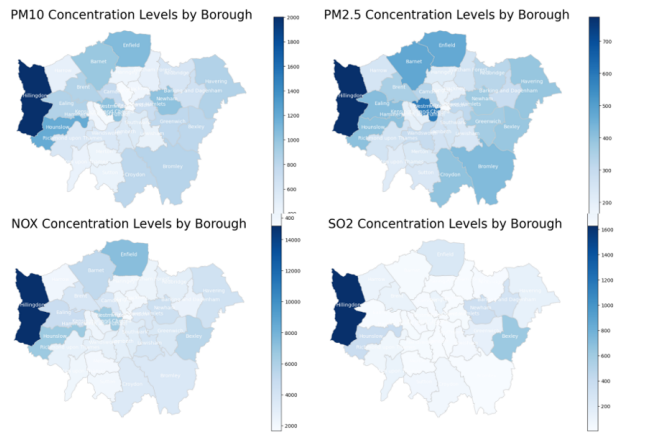

Fig. 6. GeoSpatial map highlighting the concentration of all 4 pollutants. Hillingdon (dark blue, extreme left) displays the highest concentration in all categories.

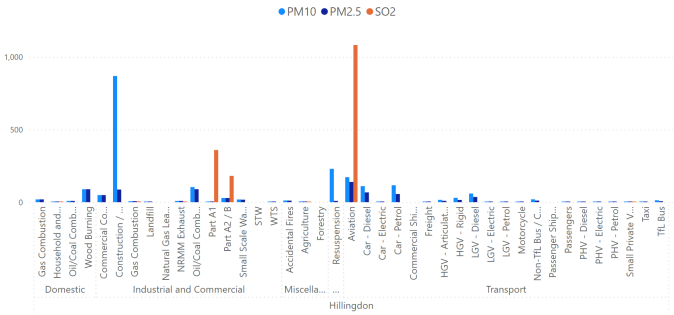

Fig. 7. Hillingdon pollutant sources for PM10, PM2.5 and SO2. Highest source for PM10-Construction dust(Blue), PM2.5 distributed over various sources and Aviation majorly contributing towards SO2

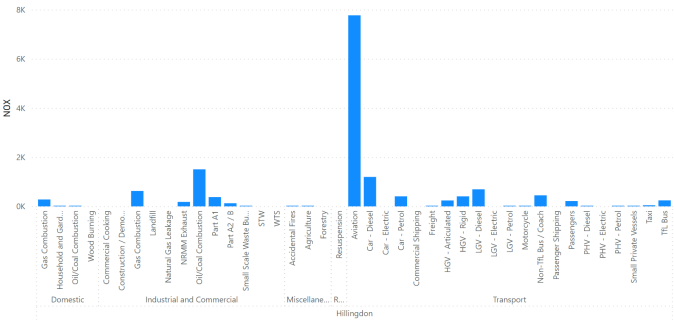

Fig. 8. Hillingdon pollutant sources for NOx where transport sources—especially Aviation, diesel cars and industrial sources like Oil/Coal combustion are the significant contributors to pollution levels

Model:

Time Series:

The time-series analysis reveals a relationship between the months and the pollutant level. A cyclic pattern can be

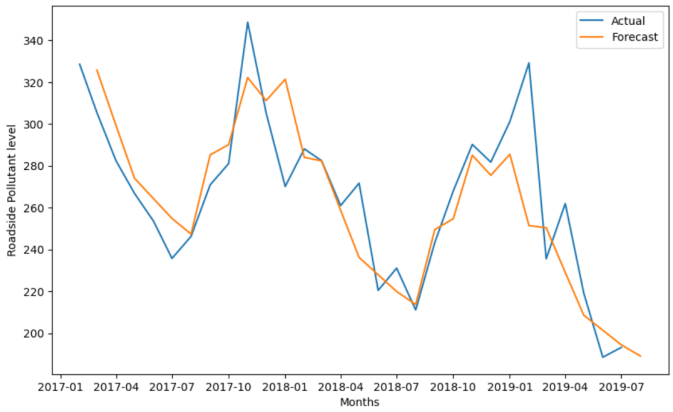

Fig. 9. Time series analysis between roadside pollutant data and months with actual and forecasted data.

observed where the pollutant levels are at their lowest during the warmer months and high during the colder months. The best hyperparameters (p, d, q)(P, D, Q, S) are obtained through manual grid-search. The performance of both the models is evaluated using Root Mean Squared Error which measures the average magnitude of the errors between predicted and actual values and is displayed in Table 1.

	Roadside	Background
RMSE	27.81	18.67

TABLE I
TIME-SERIES MODEL PERFORMANCE WITH SARIMA (1, 0, 0) (1, 0, 1, 12) FOR ROADSIDE AND SARIMA (0, 1, 0) (1, 0, 1, 12) FOR BACKGROUND

ElasticNet Model:

The utilized methodology integrated the ElasticNet Regression model alongside Polynomial regression, yielding the ensuing results:

		Roadside	Background
R^2 before Outlier Elimination	Train	45%	60%
	Test	65%	67%
	MSE	0.42	0.27
R^2 after Outlier Elimination	Train	82%	83%
	Test	70%	69%
	MSE	0.36	0.26

TABLE II
MODEL PERFORMANCE FOR BOTH ROADSIDE AND BACKGROUND DATA

Mean Squared Error is a measure of the difference between the predicted and actual values. Lower MSE indicates more accurate predictions. Whereas R^2, or the coefficient of determination is a measure of the goodness of fit of the model. Higher R² reflects better fit. Optimal results were achieved by fine-tuning essential hyperparameters: an alpha value of 0.1 for the ElasticNet model and a polynomial degree of 1, underscoring a linear interrelation between features and the label.

Furthermore, the process of outlier detection and elimination played a pivotal role in enhancing model performance. By identifying and removing unreliable or erroneous instances from the training data, the model's ability to learn from accurate and representative examples was reinforced. As a result, both training and test performance showed marked improvement after this preprocessing step.

V. CONCLUSION

In conclusion, the "Seasonal Impact on Air Quality Levels in London" has delved into the intricate dynamics of air quality and its interaction with seasonal variations within the bustling metropolis of London. Through meticulous analysis of extensive historical air quality data and meteorological parameters, this research has shed light on the multifaceted relationships between pollutants and atmospheric conditions.

The findings of this project underscore the significant impact of seasonal changes on air quality levels. The central hypothesis, that cyclical shifts in weather and atmospheric factors play a pivotal role in influencing pollutant concentrations, has been substantiated. The study has revealed patterns where certain pollutants around certain environments exhibit fluctuations in tandem with specific seasons, driven by factors like temperature, humidity, and more.

The research's implications in urban planning, policy, and public health are significant. Insights drive strategies for season-aligned air quality management, shaping a sustainable London. However, complexities underline the interplay of air quality factors. Future exploration includes advanced machine learning and more parameters. These endeavours deepen understanding and contribute to improving air quality for residents' well-being.

VI. FUTURE WORK

The current research project opens avenues for future work that can further enrich the understanding of air quality dynamics in London. Building upon this foundation, several directions emerge for extended investigations.

1) **Temporal Granularity:** Expanding the analysis to finer temporal granularity, such as daily or hourly data, could reveal intricate patterns in pollutant concentrations and their interactions with meteorological variables. This could shed light on short-term fluctuations and aid in identifying pollution sources during specific times of the day.
2) **Alternate Modeling Techniques:** While ElasticNet offers valuable insights, exploring other advanced machine learning techniques like Random Forests or neural networks could potentially extract deeper patterns and non-linear relationships within the data.
3) **Scenario Simulations:** Employing the developed model for scenario simulations could help evaluate the potential effects of policy interventions or urban development plans on air quality under different seasonal conditions.

VII. References

A. *Web reference:*

GLA and TFL Air Quality (2019), London Atmospheric Emissions Inventory (LAEI) 2019 Available from - https://data.london.gov.uk/dataset/london-atmospheric-emissions-inventory–laei–2019

King's College London (2019), London Average Air Quality Levels Available from - https://data.london.gov.uk/dataset/london-average-air-quality-levels

EMMANUEL F. WERR (2021), London Weather Data Available from - https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

B. *Periodical (journal) reference:*

Carmona, R., Díaz, J., Mirón, I.J., Ortiz, C., Luna, M.Y. and Linares, C., 2016. Mortality attributable to extreme temperatures in Spain: A comparative analysis by city. Environment International, 91, pp.22-28.

Ren, C., O'Neill, M.S., Park, S.K., Sparrow, D., Vokonas, P. and Schwartz, J., 2011. Ambient temperature, air pollution, and heart rate variability in an ageing population. American journal of epidemiology, 173(9), pp.1013-1021.

Vardoulakis, S., Dear, K., Hajat, S., Heaviside, C., Eggen, B. and McMichael, A.J., 2014. Comparative assessment of the effects of climate change on heat-and cold-related mortality in the United Kingdom and Australia. Environmental health perspectives, 122(12), pp.1285-1292.

Chen R, Peng RD, Meng X, Zhou Z, Chen B, Kan H. Seasonal variation in the acute effect of particulate air pollution on mortality in the China Air Pollution and Health Effects Study (CAPES). Sci Total Environ. 2013 Apr 15;450-451:259-65. doi: 10.1016/j.scitotenv.2013.02.040. PMID: 23500824; PMCID: PMC3885864.

Liu Y, Zhou Y, Lu J. Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. Sci Rep. 2020 Sep 3;10(1):14518. doi: 10.1038/s41598-020-71338-7. PMID: 32883992; PMCID: PMC7471117.

Ailish M. Graham, Kirsty J. Pringle, Stephen R. Arnold, Richard J. Pope, Massimo Vieno, Edward W. Butt, Luke Conibear, Ellen L. Stirling, James B. McQuaid, Impact of weather types on UK ambient particulate matter concentrations, Atmospheric Environment: X, Volume 5, 2020, 100061, ISSN 2590-1621,

Barmpadimos, I., Keller, J., Oderbolz, D., Hueglin, C., and Prévôt, A. S. H.: One decade of parallel fine (PM2.5) and coarse (PM10–PM2.5) particulate matter measurements in Europe: trends and variability, Atmos. Chem. Phys., 12, 3189–3203, 2012.

Lee, W., Bell, M.L., Gasparrini, A., Armstrong, B.G., Sera, F., Hwang, S., Lavigne, E., Zanobetti, A., Coelho, M.D.S.Z.S., Saldiva, P.H.N. and Osorio, S., 2018. Mortality burden of diurnal temperature range and its temporal changes: a multi-country study. Environment International, 110, pp.123-130.

Jacob, D.J. and Winner, D.A., 2009. Effect of climate change on air quality. Atmospheric environment, 43(1), pp.51-63.

Salcedo, R., Alvim Ferraz, M., Alves, C., & Martins, F. (1999). Time-series analysis of air pollution data. Atmospheric Environment, 33(15), 2361-2372.

Elminir, H.K., 2005. Dependence of urban air pollutants on meteorology. Science of the total environment, 350(1-3), pp.225-237.

de la Paz, D., Borge, R. and Martilli, A., 2016. Assessment of a high-resolution annual WRF-BEP/CMAQ simulation for the urban area of Madrid (Spain). Atmospheric Environment, 144, pp.282-296.

Westervelt, D.M., Horowitz, L.W., Naik, V., Tai, A.P.K., Fiore, A.M. and Mauzerall, D.L., 2016. Quantifying PM2. 5-meteorology sensitivities in a global climate model. Atmospheric Environment, 142, pp.43-56.