Visualisation of High Dimensional Time Series Pollution Data

Abstract: Pollution continues to grow in tandem with the rapid urbanisation of cities. Combating this trend remains a significant challenge; there is an increasing need for analytical techniques that provide methods for analysis by detecting trends, correlations and presenting these insights in a manner easily understood by domain users. This research consumes a voluminous dataset containing 26,305 hourly readings for multiple pollutants(CO, O_3, SO_2, NO_2, PM2.5, PM10) and air quality for the city of London from 2023 to 2025. Relational seasonal behaviour and seasonal spearman rank are used to evaluate the effect of pollutants against themselves and other gases on a seasonal pattern. The result is fairly conclusive with observed atmospheric behaviour. Primary combustion-related pollutants, including CO, NO_2, PM10, PM2.5, and SO_2, are strongly concentrated during winter due to increased emissions, reduced photochemical removal, and poor atmospheric dispersion. In contrast, ozone shows strong summer dominance driven by enhanced photochemical production under increased solar radiation. The opposing seasonal behaviour of ozone and nitrogen dioxide reflects their coupled atmospheric chemistry, while particulate matter shows cold-season enhancement linked with secondary aerosol formation and atmospheric stability. Furthermore, a jupyter notebook is developed showing multiple visualization techniques, where the data is grouped into distinct categories to extract trends and correlations within and across these groups.

1. Introduction

With rapid urbanisation and the expansion of cities air pollution has emerged as one of the most persistent environmental challenges. Pollutant emissions from various sources such as vehicular traffic, domestic heating, industrial activities, and secondary atmospheric processes have increased in both magnitude and complexity proportionally to this expansion. To monitor and mitigate these effects, there is a heavy reliance on a dense network of IoT sensors that continuously record air quality indicators within a programmable range. While these sensing infrastructures provide unprecedented volumes of data, they also introduce a different challenge, namely how to effectively explore, interpret, and extract insight from these generated data. Air quality datasets are inherently complex, they consist of multiple interfacing pollutants, non-linear temporal dependencies, cyclically occurring events(seasonal, hour, daily cycles), and strong correlation with meteorological and photochemical processes.Traditional tabular analysis or univariate plots are insufficient to capture such structure, particularly when datasets span multiple years with hourly or sub-hourly resolution. Visualization attempts to solve this problem, by allowing domain users to identify trends, correlations and anomalies.

Previous studies have demonstrated the usefulness of traditional visualization techniques for air pollution time-series analysis. Alsultanny[1] highlights the effectiveness of one-, two-, and three-dimensional visualisations for identifying time-based patterns and correlations between pollutant gases, particularly inverse relationships driven by atmospheric chemistry. Similarly, Rahim and Masseran[2] address the challenge of high-dimensional pollution data through a combination of individual time-series plots, correlation matrices, and smoothing techniques to reduce noise and emphasise long-term trends. While these approaches successfully reveal descriptive relationships and temporal patterns, visualization is largely treated as an endpoint and not as an integrated component of a broader analytical pipeline. In particular, the lack of integration with machine learning models limits their applicability for predictive modelling and automated pattern discovery.

2. Result

Behaviour of these six pollutants(CO, O_3, SO_2, NO_2, $PM_{2.5}$, PM_{10}) are observed under three temporal categories; Hourly, Monthly, and Seasonally. The hourly analysis is plotted on a strip plot to show temporal distribution within an interval of four hours and reveals diurnal variations across all pollutants. CO and NO_2 show higher concentrations and greater variability during the morning and late evening hours between 8–12 and 20–24, while exhibiting more compact and stable distributions around midday (12–16). SO_2 remains consistently low across all intervals, and suggests more episodic emissions as it experiences sporadic spikes in the early mornings and late evening periods. O_3 displays a strong daytime preference, with low concentrations during nighttime and early morning and increases during midday and afternoon hours, reflecting photochemical formation under sunlight. Both $PM_{2.5}$ and PM_{10} are elevated and more variable during morning and late evening periods, coinciding with increased emissions and limited atmospheric dispersion. During midday, concentrations decrease and become more stable, likely due to enhanced atmospheric mixing and dispersion under daylight conditions.

Monthly analysis is plotted on a scatter plot and faceted by year. This analysis reveals recurring monthly behaviour rather than abrupt long-term changes in pollutant levels. CO exhibits greater variability across months and years while NO_2 and particulate matter($PM_{2.5}$, and PM_{10}) show more moderate fluctuations. While SO_2 remains consistently low overall.

Seasonal analysis is performed using two techniques and plotted using a heatmap; relational seasonal behaviour, z-score normalisation is performed on each pollutant to extract its seasonal variation relative to its mean. Secondly, Spearman correlation is used to assess relationships between pollutants.

The relative behaviour heatmap reveals chemically consistent behaviour across all pollutants. Primary combustion-related pollutants - CO, NO_2, PM10, PM2.5, and SO_2 - are strongly concentrated during winter due to increased emissions, poor atmospheric dispersion and reduced photochemical removal. In contrast, ozone shows strong summer dominance driven by increased photochemical production due to increased solar radiation. The opposing seasonal behaviour of ozone and nitrogen dioxide reflects their coupled atmospheric chemistry, while particulate matter shows cold-season enhancement linked with secondary aerosol formation and atmospheric stability. Overall, the observed seasonal patterns are closely correlated with established atmospheric chemical processes typically found in urban air pollution.

The spearman seasonal correlation heatmap reveals winter remains characterized by a strong positive correlation among primary combustion-related pollutants - CO, NO_2, PM10, PM2.5, and SO_2 -, reflecting shared emission sources and limited atmospheric dispersion, while ozone shows strong negative correlations due to suppressed photochemical production and enhanced NO_x titration. In spring, these positive correlations persist but weaken, indicating improved atmospheric mixing and the gradual onset of photochemical activity, with ozone remaining inversely related to nitrogen dioxide and particulate matter. During the summer season, the correlations among most pollutants weaken greatly as enhanced atmospheric dispersion and photochemical processes dominate, leading to Ozone becoming largely decoupled from primary combustion-related pollutants, indicating that photochemical production rather than local emissions governs its variability.. Although moderate correlations remain among traffic-related pollutants, their reduced strength reflects shorter chemical lifetimes and effective mixing. In autumn, positive correlations among primary pollutants strengthen again as emission-driven behavior re-emerges and photochemical activity declines, with ozone reverting to strong negative associations with nitrogen dioxide and particulate matter. Across all seasons, the consistently strong coupling between $PM_{2.5}$ and PM_{10} highlights shared particulate formation and accumulation processes.

Two machine learning models are performed on this data: supervised time-series regression and supervised time-series classification. Both work on the Air Quality Index(AQI) property performing prediction and classification respectively. Air Quality Index is a composite indicator derived from the concentration of multiple pollutants and is inherently non-linear, multivariate and temporal. Air quality is temporally autocorrelated, meaning current air depends heavily on recent historical conditions. Given this, feature lag is implemented capturing both short term persistence and daily cyclical patterns. Before these models are used, chronology in the data is explicitly enforced, training and test data are split chronologically with the test data containing future data.

Random forest regression is used as the prediction model; the resulting Root Mean Squared Error indicates that the model's AQI predictions deviate from the true values by approximately ±6 AQI units. An R^2 value of 0.829 further confirms that the selected features explain the majority of AQI variability and capture the dominant drivers of air quality.The feature importance analysis reveals that AQI predictions are primarily driven by ozone levels and lagged particulate matter concentrations, particularly PM2.5. The prominence of 24-hour lag features indicates strong temporal persistence and daily cyclic behaviour in the nature of air pollution. Temporal context variables, such as hour of day, further contribute by capturing emission and dispersion patterns.

Random forest classification is used on discretized AQI values using the 33rd and 66th percentile splitting data into three categories: good, moderate, bad. The Random Forest classifier achieved an overall accuracy of 92%, with consistently high precision, recall, and F1-scores across the board. The model demonstrated excellent recall for good air-quality conditions and strong detection of poor air-quality periods, while moderate AQI exhibited slightly lower recall. The feature importance analysis reveals that AQI classification is primarily driven by ozone concentration and its short-term temporal persistence, highlighting the dominant role of photochemical pollution in determining air-quality. Lagged particulate matter features further contribute by capturing temporal inertia and daily recurrence patterns, while current PM and NO_2 levels refine class boundaries.

3. Implementation

Jupyter notebook is used alongside several python frameworks; Pandas, Plotly, Seaborn, Matlib, and Sklearn. Air quality data for the city of London was retrieved from 3 publicly available datasets. The CO_2 column was dropped because it contained more than 80% missing values. We convert the Date column to a date type, then create a copy of the created dataframe, leaving the original in its initial state.

Multiple grouping strategies are used for visualisation, each aiming to either extract new insight from the data or affirm a previous insight. The copied dataframe is converted from a wide format to long format to support faceted visualisation; the AQI (Air Quality Index) variable is excluded from this stage of analysis. The data are then aggregated temporally into four-hour intervals and plotted using an interactive strip plot faceted by pollutant with Plotly.

Aggregation is also performed to produce coarser temporal representations. Pollutants are aggregated by month to reduce high-frequency variability and provide a lower temporal resolution, after which the aggregated values are visualised using a scatter plot faceted by year.

Seasonal behaviours of pollutants are visualised using two methods: a seasonal profile heatmap and a seasonal Spearman correlation heatmap. To generate seasonal profile data, pollutant values are aggregated by season and normalised using the z-score technique. The resulting data are then plotted as a heatmap, revealing each pollutant's seasonal behaviour in relation to itself. Spearman's rank correlation is also used to examine relationships between pollutants. Pollutant data are grouped by season, and correlation matrices are computed separately for each season. These are visualised as heatmaps faceted by season to compare pollutant relationships across seasonal periods.

Using the initial long format dataframe explicitly sorted by date, Random Forest regressor with a strictly enforced chronological data split is implemented to predict Air Quality Index (AQI) from pollutant concentrations. Air quality is temporally autocorrelated, meaning current air depends heavily on recent historical conditions. Given this, feature lag is implemented capturing both short term persistence and daily cyclical patterns. As previously mentioned data is trained on chronologically split data, 80% of data is used for training and the remaining 20% are used for testing. Given the chronologically sorted nature of this data, the test is resultantly future data. This helps the model to test its prediction accuracy. Feature importance is then extracted to determine which features greatly affect AQI data.

Air quality classification is performed using Random Forest Classifier with a strictly enforced chronological data split on discretized data using the 33rd and 66th percentile split into three categories(Good, Moderate, Bad). Similar data preparation techniques to the regressor plot is used .We then extract the classification score and plot a confusion matrix. Feature importance is then extracted to determine which features greatly affect AQI classification.

4. Conclusion

This research implements five visualization methods for discovering patterns, trends and correlation between multiple pollutants(CO, O_3, SO_2, NO_2, PM2.5, PM10). Data is grouped across multiple temporal resolutions, namely, hourly, monthly, and seasonally to visualise the behavior of these pollutants in relation to itself, each other and temporal patterns. Additionally machine learning models are utilized for air quality prediction and classification. The observed visual and statistical patterns are consistent with established atmospheric behaviour, demonstrating that the proposed visualisation-driven workflow provides a reliable foundation for analysing high-dimensional air quality time-series data.

References

[1]. Alsultanny, Yas A. "Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases." *Journal Port Science Research* 4, no. 2 (2021): 69-76.

[2]. Rahim, Ulya Abdul, and Nurulkamal Masseran. "Visualization of Multivariate Time Series pollutant variables in Malaysia." In *Journal of Physics: Conference Series*, vol. 1988, no. 1, p. 012089. IOP Publishing, 2021.

Appendix

Github: https://github.com/knightfall22/time-series-visualization
Jupyter Notebook: https://knightfall22.github.io/time-series-visualization/