Strategic Analysis Report

Executive Summary

This report presents an analysis of hourly air quality measurements collected in an Italian city over a one-year period from March 2004 to February 2005. The dataset contains 9358 instances of hourly averaged responses across 15 variables capturing concentrations of multiple pollutants. These variables include Carbon Monoxide (CO), Benzene (C6H6), Nitrogen Oxides (NOx, NO2), and auxiliary environmental factors such as temperature, humidity, and sensor readings.

Temporal Analysis

The foundation visualization performed on this data mainly compared the effectiveness of the sensors in detecting the chemical concentrations as compared to the ground truth data. The monthly-year trends of this analysis for ground measurement generally show a constant level across all chemicals, except for NOx (GT) and NO2 (GT), which are demonstrated to fluctuate more, especially in 2025, compared to 2024. In contrast, the corresponding sensor readings for the same year-month periods show substantial volatility across all measurements, suggesting that the sensors may not reliably reflect the true chemical concentrations. There is also a discrepancy in the year analysis between ground truth measurements and sensor readings, which further highlights the ineffectiveness of sensor readings across both types of analysis.

Cyclic Trends

Daily cyclic trends were analyzed by taking the average per hour across the days. NOx (GT) and NO2 (GT) ground truth measurements were more volatile, as expected, during the day as well. However, an interesting observation is that the sensor readings were not only significantly more fluctuating but also significantly higher than the ground truth measurements, which could suggest that there might be a measurement error in these instruments.

To analyze the days as a whole, weekly cyclic analysis was also plotted by taking the average across the different weeks to see if the day makes a difference. It was surprising to note that the

lines were most stable in this graph, showing that there is minimal difference across days, even for the fluctuating chemicals like NOx (GT) and NO2 (GT).

Correlation and Statistical Significance

To understand the data more from a statistical perspective, correlation and p-values were calculated. The sensors measuring CO and NHMC seemed to be highly correlated with many attributes on the table, including variables such as C6H6, which shows that there are dependencies. This analysis, therefore, gives insight into which pollutants or sensor readings influence each other. All the p-values are nearing 0 or are 0, which shows that the observed effect or relationship is highly statistically significant.

Autocorrelation and Partial Autocorrelation

The Autocorrelation Function shows the correlation between a time series and its lagged versions, while the Partial Autocorrelation Function shows the correlation between the time series and its lag, after removing the effects of shorter lags. For the ground truth measurements, there is a constant positive lag in their Autocorrelation Function that is more significant in the beginning. However, for their Partial Autocorrelation Function, the graph remains constant around the 0 axis with minimal deviations above and below, which highlights that there are no significant time lag differences. This holds true for the sensor measurements as well.

Decomposition Analysis

Decomposition analysis separates a time series into trend, seasonal, and residual components to better understand its underlying patterns. Seasonal and residual trends seemed to be constant with no change except for NMHC(GT), which had a residual pattern. This could be because NMHC(GT) experiences irregular fluctuations or short-term variations not explained by the overall trend or seasonality. In terms of the general trend, the chemicals had fluctuations across time, and this could be attributed to changes in environmental conditions, human activities, or other activities contributing to the pollution. For sensor measurements, Seasonal and Residual trends were not visible, and similar to the ground truth measurements, there were fluctuations in the general trend across time through peaks and troughs.

Anomaly Detection

For ground truth measurements, the anomalies detected were all on the higher extreme, and this could be because of sudden spikes in pollutant levels, which could have pushed the readings well above typical values. Interestingly, while for all chemicals the anomalies were spread out across time, for NO2(GT), the anomalies arose only after February 2025. This suggests that for NO2(GT), a specific significant change or event starting around February 2025 affected the readings.

For sensor measurements, the trend of anomalies being significantly higher continued. However, interestingly, for the NO2 sensor, the anomalies showed up before the period of February 2025. This suggests that sensor-specific factors or local events may have caused anomalies ahead of the trend observed in the ground truth measurements.

Business Intelligence

Identifying daily, weekly, and seasonal cycles allows environmental authorities to implement targeted interventions and could result in more proactive approaches being implemented as compared to reactive approaches, which could be too late for the environment. These types of analysis over a longer period of time help understand the fluctuations even better. The anomalies presented through this analysis, while they may not be reasoned currently due to limited knowledge, provide critical insights that can guide further investigation and inform more robust predictive models for air quality management.

Cross-pollutant correlations inform which sensor readings provide the most value for forecasting. More than just analyzing the effectiveness of the sensor and its chemical, it is supposed to detect, and it also provides insight into dependencies and correlations between factors that we might not have accounted for. This helps to address any multicollinearity concerns one might have.

Modeling Foundation

The predictive modeling approach, informed by this analysis, will leverage Random Forest, baseline models, and LSTM neural networks to forecast ground truth chemical concentrations. The target variable for each model will be the corresponding ground truth measurement, while

the features will include all sensor readings alongside engineered temporal variables (hour, day, month, season, and cyclical encodings). Given the high correlation among sensor measurements, all sensor readings are included as features to maximize predictive power and capture interdependencies between pollutants. Finally, these models would be compared against common metrics such as RMSE to evaluate the best model for future predictions in this use case.