Phase 2: Advanced Environmental Data Intelligence and Pattern Analysis

Executive Summary

This report analyzes one year of environmental sensor data (March 2004 – April 2005, 9,357 hourly records) for three pollutants: Carbon Monoxide (CO), Nitrogen Oxides (NOx), and Benzene. The analysis highlights key temporal patterns, pollutant correlations, and anomalies to provide a foundation for predictive modeling in Phase 3. Some key insights include:

- CO and NOx exhibit moderate missingness, while Benzene is more reliable.
- Pollutant concentrations exhibit two daily peaks, in the morning (08:00–09:00) and evening (18:00–20:00), both strongly tied to commuter traffic, with a daily low around 04:00–05:00 when emissions are minimal.
- Weekday pollutant levels are consistently higher than weekend levels..
- Correlations between pollutants are strong (CO–Benzene r = 0.93, CO–NOx r = 0.80, NOx–Benzene r = 0.72; all statistically significant), suggesting common emission sources.
- Seasonal decomposition reveals elevated NOx and Benzene levels in winter, linked to heating demand.
- Autocorrelation confirms strong persistence, with dependencies extending 24–48 hours.
- No major anomalies were detected in the cleaned dataset, indicating preprocessing successfully removed sensor outliers and data quality issues.

Visualizations for the discussions below can be found in the attached python notebook and in the final report.

Data Quality and Missingness

The dataset spans from March 2004 to April 2005, and includes 9,357 hourly observations. Missingness was non-trivial for CO and NOx (18.0% and 17.5%, respectively), while Benzene had only 3.9% missing values.

A business implication of this is that sensor reliability directly impacts forecasting accuracy. For modeling, imputation or lag-based filling methods will be required, especially for CO and NOx. Benzene provides a relatively stable series and can serve as a predictive proxy when other sensors fail.

Temporal Patterns

Daily Cycles

All three pollutants exhibit strong diurnal patterns with two distinct peaks:

- A morning peak around 08:00–09:00, coinciding with the rush-hour commute to work and school.
- A larger evening peak around 18:00–20:00, as people return home.

Concentrations are at their lowest between 04:00–05:00, when traffic activity is minimal and emissions are at their daily trough.

This bimodal cycle closely reflects expected urban traffic behavior and highlights the strong link between human mobility patterns and air quality.

Weekly Cycles

Pollutant levels are consistently higher Monday through Friday, with reductions over the weekend. The highest concentrations occur on Thursdays (DOW = 4), with the lowest on Sundays (DOW = 6). This implies that human activity, particularly commuting and industrial schedules, is the dominant driver.

Correlation and Dependencies

Pairwise correlations show strong interdependence:

- CO vs NOx: r = 0.80
- CO vs Benzene: r = 0.93
- NOx vs Benzene: r = 0.72

All correlations are statistically significant ($p < 0.001$). This suggests that pollutants share common combustion-related emission sources such as vehicles and heating systems.

Seasonal and Trend Analysis

STL decomposition highlights three key elements:

1. **Trend:** The long-term trend for NOx, and to a lesser extent CO and Benzene, shows elevated concentrations during the colder months. This is consistent with increased heating demand in winter. Outside of winter, the trend is relatively stable.

2. **Seasonal Component:** A strong daily cycle is visible across all pollutants, corresponding to traffic-related emissions. This cyclical component captures the regular morning and evening rush-hour peaks.

3. **Residuals:** The residual series contains short-lived spikes not explained by trend or seasonality. These may represent brief pollution episodes or sensor fluctuations.

Autocorrelation and Partial Autocorrelation

Autocorrelation (ACF) and partial autocorrelation (PACF) confirm significant persistence:

- Short lags (1–3 hours): Strong positive autocorrelation, meaning pollutant levels at one hour are strongly predictive of the next few hours.
- Daily cycle (24 hours): Pronounced secondary peaks reflect the repeating diurnal traffic-driven cycle.
- Extended influence: Dependencies remain detectable up to 48 hours, indicating multi-day carryover effects.

Anomaly Detection

Following preprocessing, no major anomalies were identified in the cleaned dataset. This outcome suggests that the data quality procedures, particularly the handling of sentinel values such as -200, were effective in filtering out faulty sensor readings and outliers.

Even though anomalies were not present in the historical dataset, anomaly detection remains a valuable tool in a real-time deployment context. It can provide early-warning signals for extreme pollution episodes, enabling timely public health advisories, and sensor malfunctions or drift, ensuring system reliability and data integrity.

Feature Engineering

The Phase 2 findings highlight several features that should be incorporated into predictive models. Temporal features such as hour-of-day and day-of-week, which capture the strong diurnal and weekly cycles are tied to human mobility. Lagged pollutant levels can be used to leverage persistence and autocorrelation as revealed by ACF/PACF analysis. Rolling averages and smoothing windows can account for short-term memory effects and help mitigate noise.Cross-pollutant predictors, as strong correlations between CO, NOx, and Benzene allow each pollutant to provide predictive power when others have missing or unreliable readings.

Operational Implications

The analysis also yields actionable insights for air quality management:

- Traffic-related interventions during peak hours could significantly reduce exposure to pollutants.
- Seasonal monitoring strategies are essential, as NOx and Benzene concentrations rise during winter months due to heating demand and atmospheric stagnation.
- Sensor redundancy and data quality monitoring are critical, particularly for CO and NOx, which show higher rates of missingness. Benzene's relative reliability suggests it could serve as a fallback predictor when other sensors underperform.

Predictive Modeling Strategy

The Phase 2 analysis shows strong autocorrelation, clear daily and weekly cycles, and high cross-pollutant correlations, which directly inform the modeling strategy for Phase 3. Traditional time-series models such as SARIMA can serve as interpretable baselines by capturing seasonality and persistence, while machine learning approaches like Random Forests can leverage lagged, temporal, and cross-pollutant features to model more complex dynamics. For longer-term dependencies, LSTM architectures may also be considered, but even with simpler models, the strong interdependencies among CO, NOx, and Benzene suggest that a multi-pollutant forecasting framework will be especially valuable for robustness against sensor gaps or failures.