Anomaly Detection in Air Quality Data using Gaussian HMM

Chengjun Wu
@ucsd.edu

Lindsey Gu
@ucsd.edu

Yuchen Song
yus121@ucsd.edu

Yishan Cai
@ucsd.edu

Abstract

1 Problem Description

Accurate air quality monitoring is fundamental for understanding pollution dynamics, supporting regulatory decision-making, and protecting public health. Modern multisensor measurement systems enable continuous collection of pollutant data such as CO, NOX, and Benzene, producing long multivariate time series that reflect underlying environmental conditions. Effectively modeling these temporal patterns is essential for interpreting pollution levels, identifying regime changes, and improving the robustness of monitoring frameworks.

Hidden Markov Models (HMMs) provide a principled probabilistic approach for characterizing time-dependent environmental processes. By representing observations as emissions from latent states, HMMs can capture underlying pollution regimes, temporal structure, and transitions that are not directly observable. In addition, by learning normal pollution patterns through latent states, HMMs naturally identify atypical periods—such as sensor malfunctions, abrupt spikes, or unexpected pollution events—providing a statistically grounded approach for anomaly detection and quality control.

In this project, we apply HMM-based inference and learning methods to a year-long multivariate air quality dataset (9358 hourly instances from March 2004 to February 2005) from the UCI Machine Learning Repository (1) to model pollution regimes and detect measurement anomalies. Specifically, we [1] train Gaussian HMMs to discover latent pollution states, [2] analyze temporal regime transitions and their relationship to environmental conditions, and [3] develop likelihood-based anomaly scores to identify irregular or unreliable measurements. Our goal is to evaluate whether latent-state modeling can improve both interpretability and reliability in real-world air quality monitoring systems.

2 Data Sourcing and Processing

2.1 Dataset Description

We use the *Air Quality* dataset collected by De Vito et al. and hosted on the UCI Machine Learning Repository (1). The dataset contains 9,358 hourly-averaged measurements recorded from March 2004 to February 2005 in an urban road-level monitoring station in Italy. Data were obtained from an array of five metal-oxide (MOX) chemical sensors embedded in an air-quality multisensor device, together with a set of co-located reference analyzers that provide certified ground-truth pollutant concentrations.

The dataset includes both high-quality *ground truth* (GT) pollutant concentrations and raw sensor outputs. GT variables consist of carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene (C_6H_6), total nitrogen oxides (NO_x), and nitrogen dioxide (NO_2), each measured by a reference-grade analytical instrument. In parallel, the multisensor device provides five MOX sensor

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

readings (PT08.S1–S5), each nominally sensitive to a particular pollutant but known to exhibit cross-sensitivities and drift effects, as discussed in De Vito et al. (2008). Additional meteorological features include temperature (T), relative humidity (RH), and absolute humidity (AH), all of which influence pollutant dispersion and sensor response characteristics.

Missing values are encoded with a placeholder value of -200, reflecting moments of sensor malfunction or unavailable GT readings. Because no official preprocessing or imputation protocol is provided with the dataset, we develop a tailored missing-value handling strategy detailed in the next subsection.

2.2 Data Integrity and Missing Value Processing

To design an appropriate imputation strategy, we first computed the missing-value ratio for each variable. The results reveal substantial heterogeneity across features: NMHC(GT) has an extremely high missing rate of 90.23%, CO(GT), NO_2(GT), and NO_x(GT) have moderate missing rates around 17–18%, whereas most sensor and meteorological variables (PT08.S1–S5, T, RH, AH) exhibit only 3.91% missing entries.

Based on these ratios, we categorize variables into three tiers:

- **High missing ratio ($>$80%)**: Variables in this tier, such as NMHC(GT), contain insufficient information to permit reliable reconstruction and are therefore removed.
- **Low missing ratio ($<$10%)**: For variables with only sparse, isolated missing entries, we apply linear interpolation combined with forward and backward filling. This preserves temporal continuity while avoiding unnecessary model complexity.
- **Mid missing ratio (10–80%)**: GT pollutant variables with moderate missing rates are imputed using a regression-based approach. Each GT pollutant is predicted from its corresponding MOS sensor reading (e.g., NO_2(GT) from PT08.S4, CO(GT) from PT08.S1), leveraging the strong physical and statistical coupling between sensors and certified analyzers. This yields more realistic estimates than purely time-based interpolation.

This tiered imputation strategy provides a principled and data-driven method for handling missing values, ensuring that the downstream HMM modeling operates on clean and physically meaningful input features.

2.3 Trend Analysis

To understand the baseline temporal structure of the pollutant concentrations before applying HMM modeling, we examined long-term, diurnal, and weekly patterns in the ground-truth (GT) variables and relevant meteorological features.

2.3.1 Long-Term Temporal Trends

Figure 1 presents the 24-hour rolling mean concentrations of the major pollutants and meteorological variables over the 13-month monitoring period (March 2004–April 2005). Carbon monoxide (CO) remained relatively stable throughout the year, with mean values of approximately 2.1 ± 0.8 mg/m^3. Occasional episodic peaks exceeding 4 mg/m^3 were observed, likely associated with unfavorable dispersion conditions or localized emission events.

Nitrogen oxides (NO_x and NO_2) exhibited substantially higher temporal variability. NO_x ranged from 50 to 600 µg/m^3, with a marked increasing trend beginning around September 2004, potentially linked to reduced atmospheric mixing heights and increased heating-related emissions during colder months. A sharp decline in NO_2 concentrations around October 2004 may indicate a potential data quality issue and warrants further inspection.

Meteorological variables display clear seasonal cycles: temperature fluctuates between 5–30°C, while relative humidity varies between 25–80%. The strong inverse relationship between temperature and humidity aligns with typical continental climate behavior.

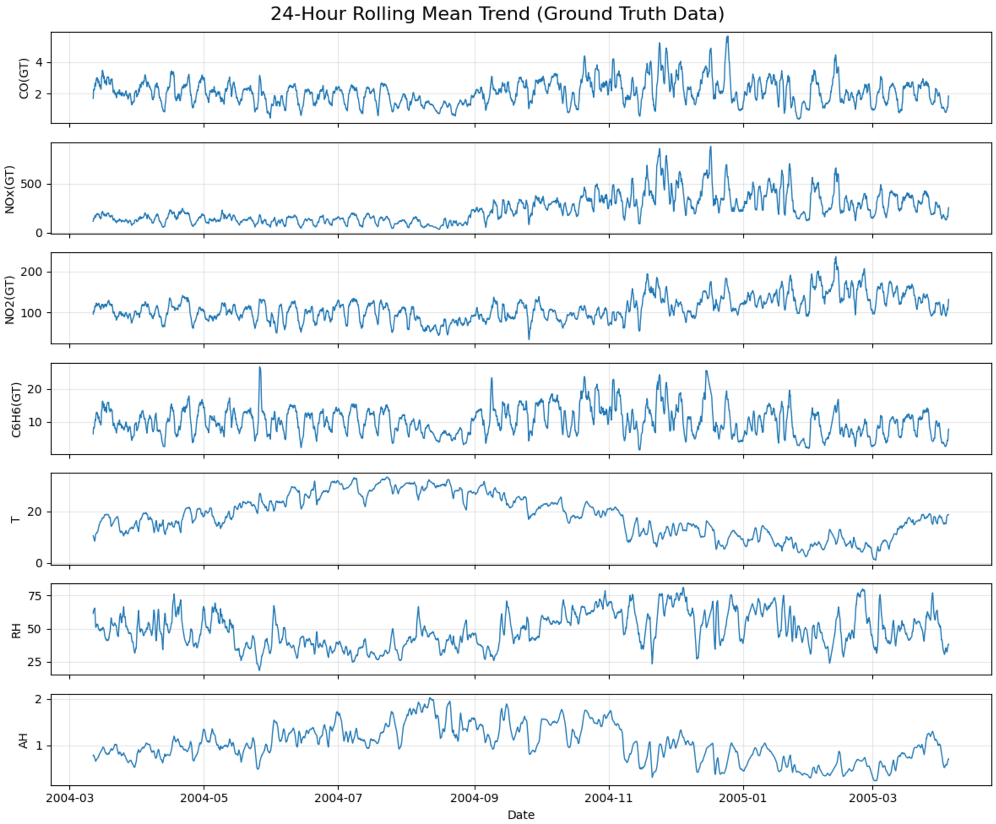

Figure 1: 24-hour rolling mean concentrations of major pollutants and meteorological variables from March 2004 to April 2005.

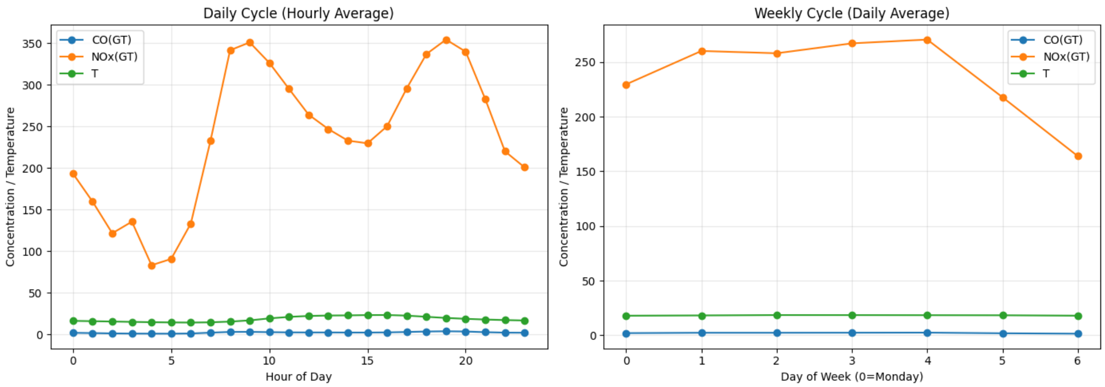

Figure 2: Diurnal and weekly pollutant patterns. NO_x shows strong bimodal traffic peaks and a clear weekend effect, while CO exhibits weaker weekly variation.

2.3.2 Diurnal and Weekly Cycles

Figure 2 illustrates diurnal and weekly patterns in pollutant concentrations. NO_x exhibits a pronounced bimodal diurnal structure, with peaks during morning (08:00–10:00) and evening (18:00–20:00) rush hours, characteristic of traffic-dominated urban environments. The peak-to-trough ratio of roughly 4:1 highlights the strong influence of vehicular emissions.

Weekly-cycle analysis reveals a clear "weekend effect" in NO_x: Sunday concentrations are approximately 37% lower than the weekday average, reflecting reduced traffic volume. In contrast, CO shows minimal weekly variation, suggesting contributions from more spatially distributed and temporally stable emission sources such as residential heating and regional atmospheric transport.

These patterns suggest the presence of underlying pollution "regimes" that HMMs may effectively capture.

3 Modeling and Inference

3.1 HMM Modeling

To capture the latent pollution regimes underlying the observed air quality measurements, we employ a Gaussian Hidden Markov Model (HMM) with four hidden states. The model is trained on five key environmental features: CO(GT), NO$_x$(GT), NO$_2$(GT), temperature (T), and relative humidity (RH). All features are standardized using z-score normalization to ensure comparability across variables with different scales and physical units.

3.2 Gaussian HMM Formulation

Let $z_t \in \{1, \ldots, K\}$ denote the latent state at time t, and let $x_t \in \mathbb{R}^d$ be the observed feature vector. A Hidden Markov Model assumes the following generative structure:

$$p(z_t \mid z_{t-1}) = A_{z_{t-1}, z_t}, \qquad p(x_t \mid z_t = k) = \mathcal{N}(\mu_k, \Sigma_k).$$

The joint distribution factorizes as

$$p(x_{1:T}, z_{1:T}) = \pi_{z_1} \prod_{t=2}^{T} A_{z_{t-1}, z_t} \prod_{t=1}^{T} \mathcal{N}(x_t \mid \mu_{z_t}, \Sigma_{z_t}).$$

This defines a K-state Gaussian HMM governing transitions between latent pollution regimes.

3.3 Model Specification

We configure a Gaussian HMM with the following settings:

- **Number of hidden states**: 4, representing distinct latent pollution regimes.
- **Covariance type**: Full covariance matrices, allowing the model to capture complex dependencies among pollutants and meteorological factors.
- **Training algorithm**: Baum–Welch expectation–maximization (EM) with a maximum of 100 iterations.
- **Convergence tolerance**: 0.01.

Temporal features such as hour-of-day, day-of-week, and month are extracted from timestamps for interpretability analyses in Section 4, but are not used directly as HMM inputs to ensure the model focuses on latent emission patterns rather than deterministic periodic trends.

3.4 Training Procedure and Model Selection

To mitigate convergence to poor local optima—a known issue in EM-based estimation—we adopt a multi-start training strategy with five independent random initializations. Each model instance is trained to convergence, and the log-likelihood of the observed sequence is evaluated. The final model is chosen as the one achieving the highest log-likelihood.

The selected model attains a log-likelihood of $-37{,}118.75$. Although negative, this value is expected because it corresponds to the sum of log-probabilities over a long multivariate time series. Training curves exhibit smooth, monotonic improvement in log-likelihood before stabilizing, indicating successful EM convergence.

3.5 Inference

Given the learned HMM parameters $\{\pi, A, \mu_k, \Sigma_k\}$, we compute posterior distributions over latent states using the forward–backward algorithm. The marginal posterior is obtained via

$$p(z_t = k \mid x_{1:T}) = \frac{\alpha_t(k)\,\beta_t(k)}{\sum_j \alpha_t(j)\,\beta_t(j)},$$

where $\alpha_t(k)$ and $\beta_t(k)$ denote the forward and backward messages.

For regime interpretation, we also compute the most likely latent-state sequence using the Viterbi algorithm:

$$\hat{z}_{1:T} = \arg\max_{z_{1:T}} p(z_{1:T} \mid x_{1:T}).$$

These inferred state trajectories form the basis for the regime analysis in Section 4.

3.6 Anomaly Detection Framework

While the Gaussian HMM provides a compact representation of latent pollution regimes, our primary objective is to leverage this model for anomaly detection in the air quality time series. In this subsection, we formalize how anomalies are defined, quantified, and detected based on the probabilistic structure learned by the HMM.

3.6.1 Problem Definition

Air quality measurements may exhibit irregular behaviors that fall into two main categories:

- **Point anomalies** are short-lived and localized deviations caused by sudden pollution spikes, sensor malfunctions, or extreme environmental conditions.
- **Distributional drift**: long-term shifts in underlying pollution patterns, such as seasonal transitions or structural changes in pollutant dynamics.

Let $x_t \in \mathbb{R}^d$ denote the multivariate observation at time t, and let $z_t \in \{1, \dots, K\}$ be the latent state inferred by the HMM. Our aim is to quantify how unlikely an observation or a window of observations is under the learned emission and transition structure of the model.

3.6.2 Point Anomaly Detection via Mahalanobis Distance

Each latent state z in the Gaussian HMM is characterized by an emission distribution $\mathcal{N}(\mu_z, \Sigma_z)$. For a time point assigned to state z_t, we compute its state-conditional deviation using the Mahalanobis distance:

$$D_M(x_t) = \sqrt{(x_t - \mu_{z_t})^\top \Sigma_{z_t}^{-1} (x_t - \mu_{z_t})}. \tag{1}$$

A large $D_M(x_t)$ indicates that x_t is improbable under the emission model of its latent state. We identify point anomalies using a data-driven threshold:

$$x_t is anomalous if D_M(x_t) > \text{Quantile}_{0.99}(D_M(x_{1:T}))\,, \tag{2}$$

where the quantile is computed over the training period. This approach is robust and state-aware, effectively normalizing deviations by the covariance structure of each regime.

3.6.3 Likelihood-Based Scoring (Alternative)

An alternative anomaly score uses the emission log-likelihood:

$$\ell_t = \log p(x_t \mid z_t), \tag{3}$$

with low values indicating unlikely observations. While effective, we find that the Mahalanobis formulation in eq:mahalanobis is numerically more stable for our data and yields cleaner separation between normal and anomalous behavior.

3.6.4 Drift Detection via Sliding-Window State Distributions

Point-wise methods do not capture slow structural changes in the pollution regime. To detect long-term drift, we analyze how the empirical latent-state distribution evolves over time. For a sliding window of size W, we compute:

$$p_k^{(t)} = \frac{1}{W} \sum_{i=t-W+1}^{t} \mathbf{1}(z_i = k), \tag{4}$$

where $p^{(t)} \in \Delta^{K-1}$ is the frequency of each latent state in the window.

We define a baseline distribution $p^{(0)}$ from an initial stable period. Drift at time t is quantified using the Total Variation (TV) distance:

$$\mathrm{Drift}(t) = \frac{1}{2} \sum_{k=1}^{K} \left| p_k^{(t)} - p_k^{(0)} \right| . \tag{5}$$

A window is flagged as drift when:

$$\mathrm{Drift}(t) > \mathrm{Quantile}_{0.99}(\mathrm{Drift}(1:T)) . \tag{6}$$

This procedure captures regime-level changes that are common in air quality data, such as the transition from summer to winter, which often presents markedly different pollution patterns.

3.6.5 Summary

The proposed framework integrates two complementary detection mechanisms:

- **Point anomaly detection** using Mahalanobis distance and likelihood-based scoring to capture abrupt deviations in individual observations.
- **Drift detection** using sliding-window latent-state distributions to identify long-term structural shifts in pollution regimes.

Together, these approaches leverage both the emission geometry and temporal dynamics learned by the HMM, enabling robust detection of a broad spectrum of anomalous behaviors in air quality data.

4 Results and Discussion

4.1 Training Behavior and Convergence

The Gaussian HMM was trained using the Baum–Welch EM algorithm with multiple random initializations. Across runs, the log-likelihood exhibited the expected monotonic improvement characteristic of EM-based methods. The multi-start procedure was essential; different random initializations occasionally converged to distinct local optima, and we selected the model with the highest final log-likelihood. Convergence was reached within the allotted iteration budget, indicating that the four-state specification effectively captures the dominant temporal and multivariate structure present in the data.

4.2 Latent State Summary and Emission Characteristics

The inferred latent states correspond to interpretable pollution regimes. State 0 acts as a baseline condition with moderate pollutant concentrations and balanced meteorological variables. State 1 captures a winter-dominated high-$\mathrm{NO_x}$ regime, marked by elevated NOx and $\mathrm{NO_2}$ levels and lower temperatures—consistent with reduced dispersion and increased heating-related emissions during the colder months. State 2 represents clean summer conditions with lower pollutant levels, higher temperatures, and higher relative humidity. State 3 corresponds to short-lived pollution spikes, typically aligned with abrupt increases in CO or NOx. These patterns align well with known seasonal and environmental behaviors observed in urban air quality systems.

4.3 Temporal Structure of Inferred Pollution Regimes

The temporal evolution of the inferred states reveals clear seasonal organization. State 1 dominates extensive portions of the winter months, reflecting persistent high-NO_x pollution episodes. State 2 becomes more prevalent during the summer, capturing extended periods of cleaner atmospheric conditions. State 0 appears throughout the year as a background or intermediate regime. In contrast, State 3 appears only sporadically and for extremely short durations, consistent with the interpretation of sudden pollution spikes or sensor irregularities. Overall, the temporal structure learned by the HMM successfully encodes long-range seasonal variability in pollutant behavior.

4.4 State Duration and Transition Dynamics

Analysis of state duration reveals clear differences in temporal persistence. State 1 exhibits long durations, aligning with sustained winter pollution patterns. State 2 also forms multi-hour segments during summer periods. State 0 displays intermediate durations, reflecting its role as a background regime. State 3 exhibits extremely short durations— often lasting only a single time step—consistent with transient pollution spikes. The transition behavior mirrors these observations: transitions into State 3 are rare and typically followed by an immediate return to a non-spike state, while transitions among States 0, 1, and 2 reflect slower, seasonally driven dynamics. These duration and transition characteristics validate the ability of the HMM to distinguish persistent pollution regimes from short-lived anomalies.

4.5 Anomaly Detection via Likelihood-Based Scoring

(Will be implemented after the experiment step)

4.6 Comparison with Other Models

(Will be implemented after the experiment step)

5 Conclusion

6 Reflections and Contributions

References

[1] S. Vito, “Air Quality.” UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C59K5F.