Anomaly Detection in Air Quality Data using Gaussian HMM

Chengjun Wu
@ucsd.edu

Lindsey Gu
@ucsd.edu

Yuchen Song
yus121@ucsd.edu

Yishan Cai
yic075@ucsd.edu

Abstract

Accurate monitoring and timely detection of anomalous air pollution events are critical for urban environmental management and public health protection. In this work, we apply Gaussian Hidden Markov Models (HMMs) to a year-long multivariate air quality dataset, capturing latent pollution regimes and their temporal dynamics. The model identifies four interpretable states corresponding to baseline conditions, seasonal pollution regimes, and acute episodic events. Leveraging the state-specific Gaussian structure, we develop a robust anomaly detection framework combining Mahalanobis distance and likelihood-based scoring. Our approach detects both point anomalies and persistent segment-level events, revealing extreme pollution episodes that are consistent with observed meteorological and traffic patterns. Detailed analysis of a high-impact event illustrates the model's ability to quantify abrupt deviations in multiple pollutants simultaneously. These results demonstrate that latent-state modeling enhances both interpretability and reliability in air quality anomaly detection, providing a principled framework for environmental monitoring.

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] identify both isolated and persistent periods of unusually high pollution. By leveraging the state-specific distributions learned from ground-truth pollutant measurements, we aim to detect time points or contiguous periods that deviate significantly from expected pollution patterns, thereby highlighting genuine environmental pollution events

Preprint. Under review.

rather than sensor noise. This approach assesses whether latent-state modeling can enhance both interpretability and reliability in real-world air quality monitoring systems.

1.1 Related Work and Motivation for Using HMMs.

In recent years, Hidden Markov Models (HMMs) have been increasingly applied to air-pollution modelling and environmental exposure analysis, demonstrating that this state–transition framework is well suited for capturing the temporal structure of atmospheric pollution. For example, Liu et al. (2) use a Gaussian HMM to construct an air-quality historical-correlation model and show that carefully selecting the number of hidden states can substantially improve the accuracy and stability of AQI prediction. Rizos et al. (3) apply an HMM-based clustering method to eight years of PM_{10} and O_3 observations to automatically identify background pollution concentrations, demonstrating that HMMs outperform traditional clustering methods when modelling multi-modal distributions and regime switching. Sarvi et al. (4) employ a Poisson HMM to predict the number of $PM_{2.5}$ exceedance days, illustrating the effectiveness of HMMs in "event-count + state-switching" prediction tasks. Together, these studies indicate that HMMs can compress long pollution time series into a small number of physically interpretable pollution states and use the transition matrix to capture seasonality and regime shifts, making HMMs a natural foundational model for the air-pollution time-series modelling and anomaly-detection framework developed in this work.

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 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.

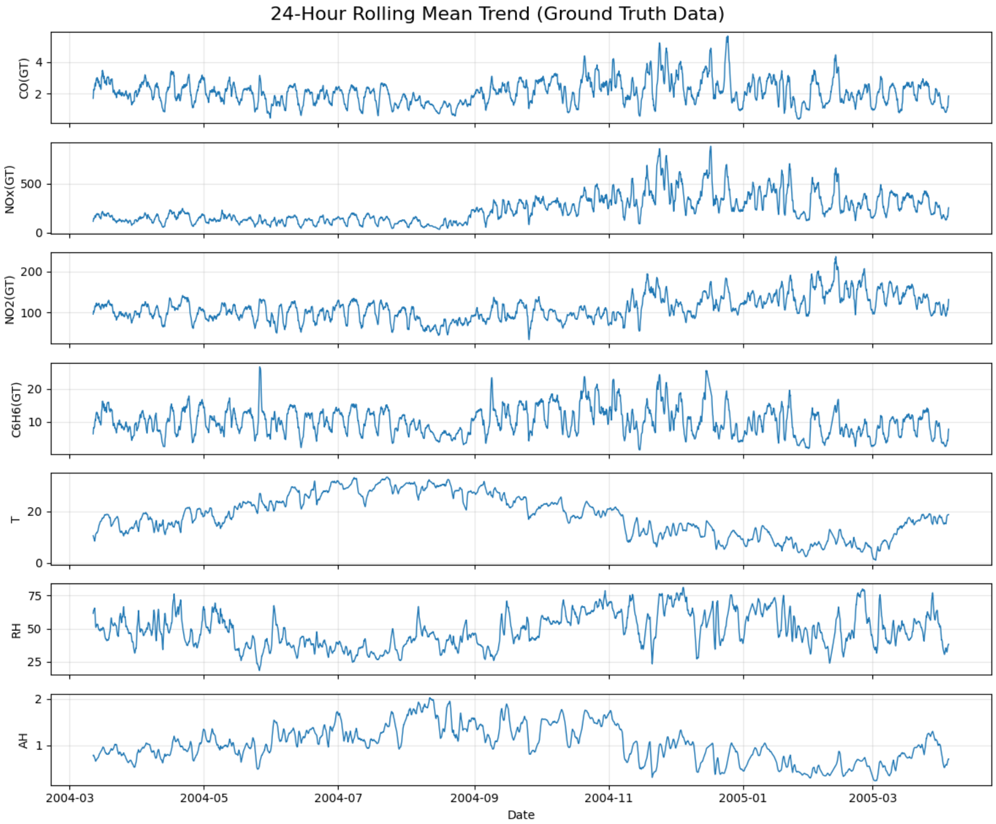

Figure 1: 24-hour rolling mean concentrations of major pollutants and meteorological variables from March 2004 to April 2005.

- **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

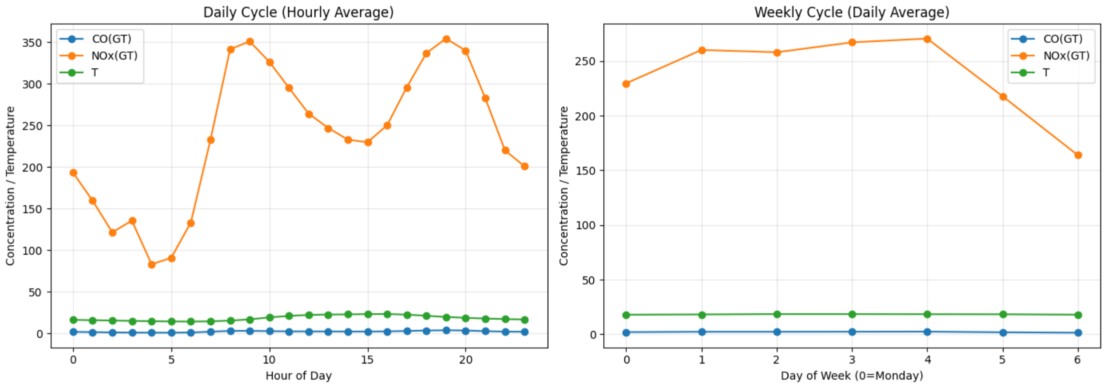

Figure 2: Diurnal and weekly pollutant patterns. NO_x shows strong bimodal traffic peaks and a clear weekend effect, while CO exhibits weaker weekly variation.

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.

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 $z_t \in \{1, \dots, K\}$ 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

3.6.1 Problem Definition

Given the observed multivariate sequence $x_{1:T}$ and the inferred latent states $\hat{z}_{1:T}$ from the Gaussian HMM, our goal is to identify time points or contiguous periods exhibiting unusual pollution patterns. Anomalies are defined as observations that deviate significantly from the expected distribution conditioned on the assigned latent state.

3.6.2 Point Anomaly Detection via Mahalanobis Distance

For each observation x_t associated with latent state $\hat{z}_t = k$, we assume the state-specific Gaussian distribution inferred by the HMM:

$$x_t \mid \hat{z}_t = k \sim \mathcal{N}(\mu_k, \Sigma_k).$$

We quantify the deviation of x_t from its expected state distribution using the Mahalanobis distance:

$$D_M(x_t) = \sqrt{(x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k)}.$$

A point is labeled anomalous if its Mahalanobis distance exceeds a state-specific threshold, defined as the 99th percentile of distances observed in that state:

$$\text{PointAnomaly}(x_t) = \begin{cases} 1, & D_M(x_t) > \text{threshold}_k \\ 0, & \text{otherwise} \end{cases}$$

In parallel, we compute the log-likelihood of x_t under the same state distribution:

$$\ell(x_t) = \log \mathcal{N}(x_t \mid \mu_k, \Sigma_k),$$

and flag points with extremely low likelihood (e.g., below the global 1st percentile) as anomalies. The final point anomaly label combines both Mahalanobis and log-likelihood signals:

$$\text{CombinedAnomaly}(x_t) = \text{PointAnomaly}(x_t) \vee \big(\ell(x_t) < \ell_{\text{threshold}}\big).$$

3.6.3 Segment Anomaly Detection

Consecutive point anomalies are merged into *segment anomalies* to identify persistent events. Segments shorter than a minimum duration (e.g., 3 hours) are discarded. This aggregation provides a more robust representation of anomalous periods in the multivariate time series:

$$\text{SegmentAnomaly}[s:e] = 1 \quad \text{if} \sum_{t=s}^{e} \text{CombinedAnomaly}(x_t) \geq \text{min\_length}.$$

3.6.4 Summary

The proposed framework leverages the state-specific Gaussian structure of the HMM to detect unusual multivariate observations. By combining Mahalanobis distance and log-likelihood thresholds, it identifies both extreme deviations and globally improbable observations. Segment-level aggregation produces interpretable anomalous events corresponding to potential pollution incidents.

4 Results and Discussion

4.1 Model Convergence and State Interpretability

The Gaussian HMM was trained using the Baum-Welch EM algorithm with multiple random initializations to avoid local optima. The log-likelihood exhibited monotonic improvement across iterations, with convergence achieved within the allocated budget. The multi-start procedure proved essential, as different initializations occasionally converged to distinct local optima; we selected the model with the highest final log-likelihood for subsequent analysis.

The learned transition matrix reveals strong diagonal persistence, indicating that the identified states represent stable pollution regimes rather than transient fluctuations. States 0, 1, and 2 exhibit self-transition probabilities exceeding 0.89, 0.97, and 0.94 respectively, demonstrating their temporal stability. In contrast, State 3 shows a notably lower self-transition probability of 0.83 and relatively higher transition rates to other states, suggesting it captures brief pollution events or transitional periods. The sparse off-diagonal structure indicates that direct transitions between non-adjacent pollution regimes are rare, reflecting the gradual nature of atmospheric changes.

4.2 Emission Characteristics and Pollution Regimes

The emission distributions associated with each latent state reveal distinct pollution regimes that align with interpretable atmospheric conditions. State 0 represents a baseline condition characterized by near-zero standardized values across most variables, indicating moderate pollutant concentrations and balanced meteorological conditions. This state serves as the reference background regime.

State 1 captures a winter-dominated high-NO_x pollution regime, marked by elevated NOx and NO_2 levels alongside substantially reduced temperatures (approximately -1.2 standard deviations below mean). The combination of high nitrogen oxide concentrations and low temperatures is consistent with reduced atmospheric dispersion during winter months and increased heating-related emissions. State 2 represents clean summer conditions with notably lower pollutant concentrations across CO,

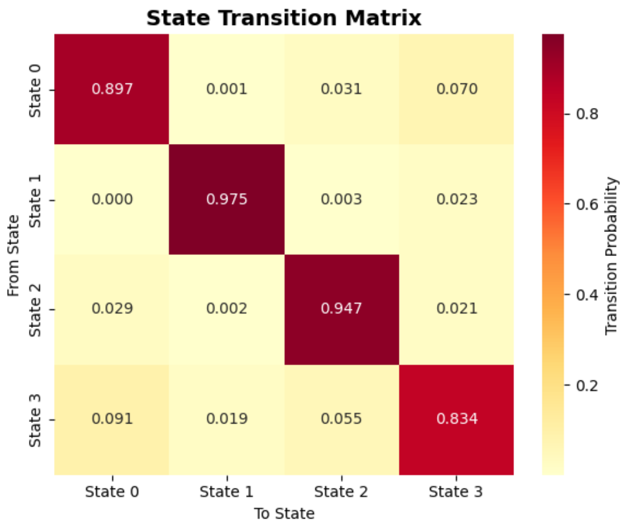

Figure 3: State transition matrix of the learned Gaussian HMM. High diagonal values indicate temporal persistence of pollution regimes, while off-diagonal elements reveal transition dynamics between states.

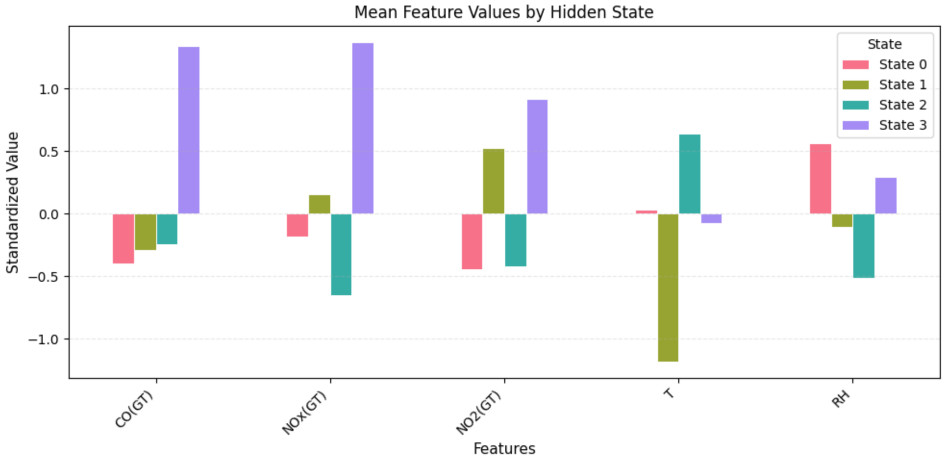

Figure 4: Mean standardized feature values for each hidden state inferred by the Gaussian HMM. Each bar represents the average level of a specific pollutant or meteorological variable (CO, NOx, NO_2, Temperature, Relative Humidity) within a given state.

NOx, and NO_2, coupled with higher temperatures and elevated relative humidity. This pattern reflects enhanced pollutant dispersion and reduced emission sources during warmer periods.

State 3 exhibits dramatically elevated levels of both CO and NOx (exceeding 1.3 standard deviations), indicating acute pollution episodes. The simultaneous spikes in multiple pollutants suggest either concentrated emission events or unfavorable meteorological conditions that temporarily trap pollutants. These emission patterns demonstrate that the HMM successfully identifies physically meaningful pollution regimes corresponding to seasonal cycles and episodic events.

4.3 Temporal Dynamics and Seasonal Patterns

The temporal evolution of state occupancy reveals pronounced seasonal and diurnal organization. Monthly state distributions show that State 1 dominates winter months (January, February, November, December), often comprising over 60% of observations during these periods. This extended winter dominance confirms the persistence of high-NO_x pollution regimes during cold seasons. Conversely, State 2 becomes increasingly prevalent from late spring through summer (May through August), with near-complete occupancy in July and August, reflecting sustained periods of cleaner atmospheric conditions.

State 0 maintains relatively consistent presence throughout the year at approximately 20-30% occupancy, serving as a transitional or intermediate regime between the dominant seasonal states. State 3 appears sporadically across all months but always constitutes a minor fraction of total observations, consistent with its interpretation as capturing brief pollution spikes rather than sustained conditions.

Diurnal patterns further illuminate the nature of these regimes. State 0 exhibits peak prevalence during early morning hours (2-4 AM), suggesting it captures overnight baseline conditions. State 1 shows elevated occupancy during nighttime and early morning, aligning with nocturnal boundary layer compression and accumulated emissions. State 2 maintains relatively stable high occupancy throughout daytime hours in summer months, while State 3 displays a pronounced mid-morning peak (around 9 AM), potentially corresponding to rush-hour emission surges or morning boundary layer transitions.

The combination of seasonal persistence and diurnal variation demonstrates that the HMM successfully encodes both long-range temporal structure and short-term cyclical patterns in urban air quality dynamics.

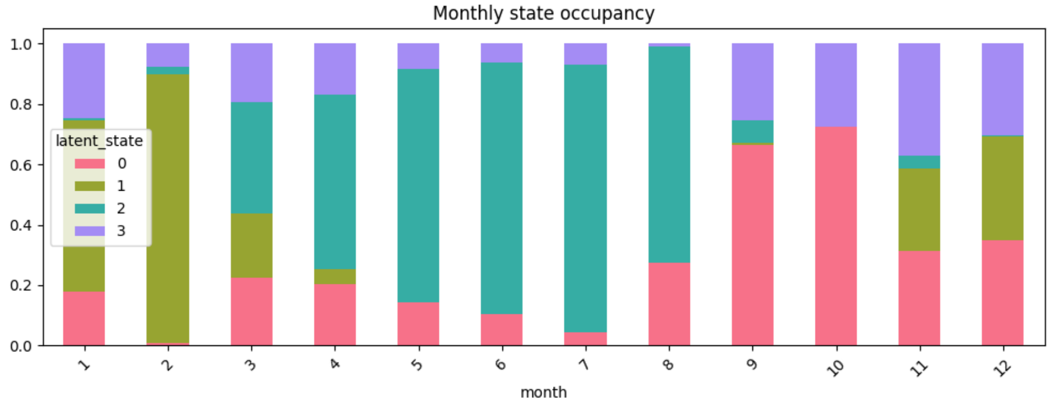

Figure 5: Monthly state occupancy showing the proportion of time spent in each latent state throughout the year. Clear seasonal patterns emerge, with State 1 dominating winter months and State 2 prevailing during summer, while State 0 provides consistent baseline presence.

4.4 State Persistence and Transition Behavior

The combination of high self-transition probabilities and observed temporal patterns reveals distinct persistence characteristics across states. States 1 and 2 form extended multi-day or even multi-week segments during their respective dominant seasons, reflecting the slow-varying nature of synoptic weather patterns that govern seasonal pollution regimes. State 0 exhibits intermediate persistence, typically lasting several hours to days as atmospheric conditions gradually shift between seasonal extremes.

State 3 displays markedly different behavior, with most occurrences lasting only one to a few hours before transitioning back to other states. This brevity, combined with its sporadic appearance and high pollutant concentrations, confirms its role in capturing transient pollution events such as traffic-related spikes, industrial emissions bursts, or brief meteorological anomalies. The rarity of direct transitions between States 1 and 2 (probability < 0.003) indicates that shifts between winter and summer regimes typically pass through intermediate conditions (State 0) rather than occurring abruptly.

These duration and transition characteristics validate the model's ability to distinguish persistent pollution regimes driven by seasonal meteorology from short-lived anomalies caused by episodic events, providing a comprehensive temporal decomposition of urban air quality dynamics.

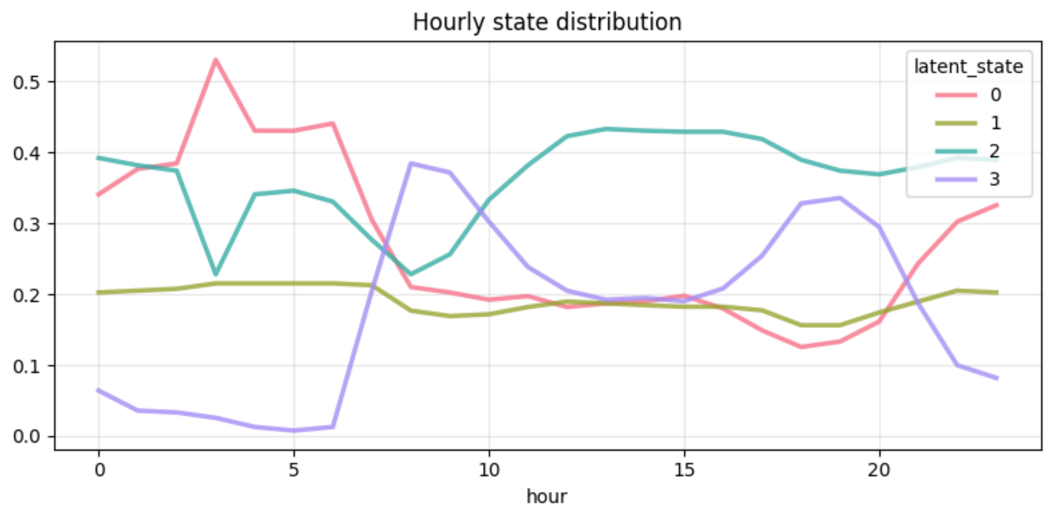

Figure 6: Hourly state distribution across the day. Diurnal cycles reveal State 0 peaks during early morning hours, State 1 shows elevated nighttime presence, State 2 maintains daytime dominance in summer, and State 3 exhibits a morning rush-hour peak consistent with episodic pollution events.

4.5 Anomaly Detection

4.5.1 Overview of Detected Events

Applying the HMM-based anomaly detection framework described in Section 3.6, we identified a total of 133 point anomalies over the one-year observation period. Aggregating consecutive anomalies yielded 8 distinct segment-level events, as summarized in Table 1.

Event	Start	End	Duration (hours)
1	2004-11-23 18:00	2004-11-23 20:00	3
2	2004-12-23 18:00	2004-12-23 20:00	3
3	2005-01-22 06:00	2005-01-22 09:00	4
4	2005-01-22 18:00	2005-01-22 20:00	3
5	2005-01-31 18:00	2005-01-31 20:00	3
6	2005-02-02 18:00	2005-02-02 20:00	3
7	2005-02-03 08:00	2005-02-03 11:00	4
8	2005-02-11 15:00	2005-02-11 17:00	3

Table 1: Segment-level anomalies detected in the one-year air quality dataset.

To visualize the anomaly distribution, Figure 7 shows the observed time series of CO, NO$_x$, and NO$_2$ concentrations with red shaded areas indicating the detected anomalous segments. This provides an intuitive overview of when extreme pollution events occurred throughout the year.

4.5.2 Detailed Analysis of a Selected Event

We focus on Event 7 (2005-02-03 08:00 – 2005-02-03 11:00) due to its pronounced variation across multiple pollutants. Summary statistics for this event are provided in Table 2.

Figure 8 presents a detailed visualization of this event using three complementary views:

- **Raw Values**: Comparison of event measurements to baseline and overall mean.
- **Z-score Spike Detection**: Standardized deviations highlighting extreme values.
- **Increase Ratio**: Ratio of event to baseline values, emphasizing the magnitude of the anomaly.

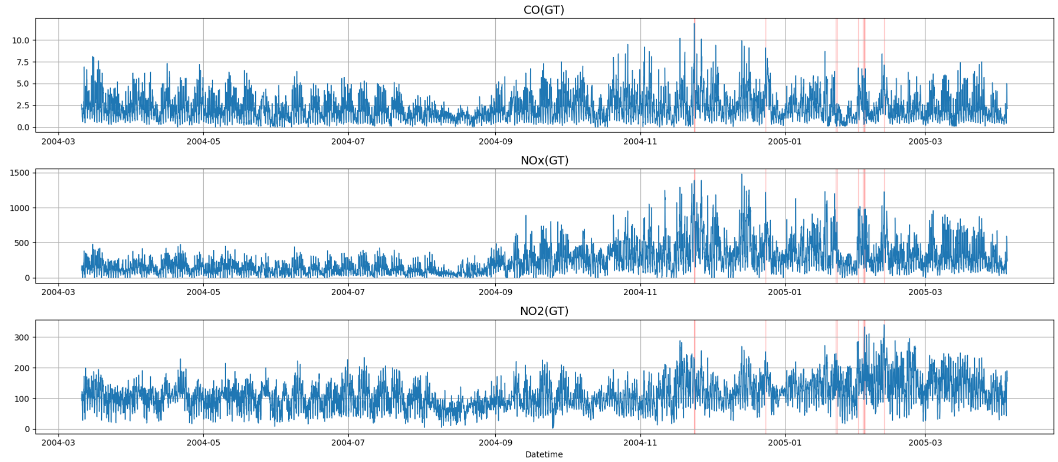

Figure 7: Detected segment-level anomalies for CO, NO_x, and NO_2. Red shaded regions correspond to anomalous periods.

Pollutant	Mean	Peak	Baseline	Increase Ratio
CO(GT)	4.25	5.6	1.76	3.18
NO_x(GT)	860.05	974.6	308.85	3.16
NO_2(GT)	309.5	332.6	149.08	2.23
T	4.26	8.53	3.95	2.16
RH	54.34	64.75	51.82	1.25

Table 2: Event 7 statistics: mean, peak, baseline values, and increase ratios relative to baseline.

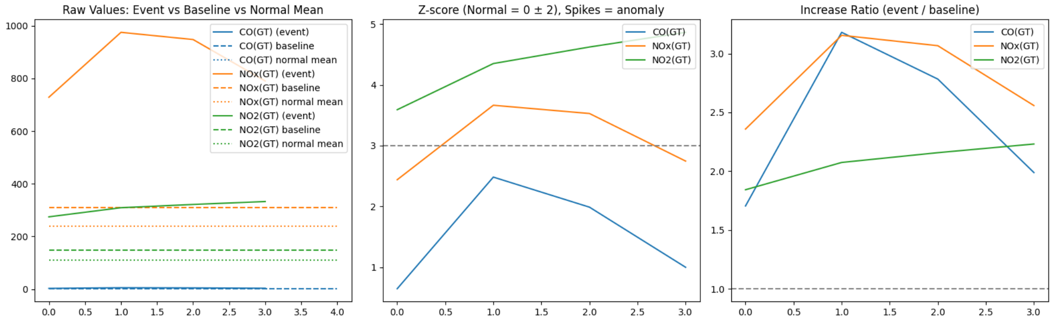

Figure 8: Detailed visualization of Event 7 (2005-02-03 08:00 – 11:00) for CO, NO_x, and NO_2. Three panels show raw values, Z-score, and increase ratios relative to baseline.

The selected event demonstrates a sharp increase in pollutant concentrations, with CO and NO_x exhibiting more than threefold increases relative to their local baseline, confirming the effectiveness of the state-specific anomaly detection framework.

4.6 Comparison with Other Models

(Will be implemented after the experiment step)

5 Conclusion

In this study, we demonstrate the effectiveness of Gaussian Hidden Markov Models for modeling urban air quality dynamics and detecting anomalous pollution events. The HMM successfully compresses long multivariate time series into a small set of interpretable latent states, capturing both seasonal trends and short-term episodic spikes. By leveraging state-specific distributions, our anomaly detection framework identifies extreme deviations in pollutant concentrations, distinguishing transient pollution events from normal diurnal and seasonal variations.

Analysis of the detected anomalies, including a detailed case study of a high-intensity event, confirms that the model provides physically meaningful insights, aligning with meteorological conditions and known traffic patterns. Overall, our approach offers a statistically grounded, interpretable, and flexible tool for urban air quality monitoring, with potential applications in environmental policy, regulatory compliance, and public health risk assessment.

6 Contributions

The project team contributed as follows: Yishan Cai was primarily responsible for exploratory data analysis (EDA), development and training of the Gaussian HMM, and implementation of the anomaly detection framework. Chengjun Wu focused on conducting the literature review and preparing the main body of the report. Yuchen Song contributed to drafting the introduction and assisted in the overall report writing.

References

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

[2] Y. Liu, L. Wen, Z. Lin, C. Xu, Y. Chen, and Y. Li, "Air quality historical correlation model based on time series," *Scientific Reports*, vol. 14, p. 22791, 2024.

[3] K. Rizos, C. Meleti, G. Kouvarakis, N. Mihalopoulos, and D. Melas, "Determination of the background pollution in the eastern mediterranean applying a statistical clustering technique," *Atmospheric Environment*, vol. 276, p. 119067, 2022.

[4] F. Sarvi, A. Nadali, M. Khodadost, M. Kharghani Moghaddam, and M. Sadeghifar, "Application of poisson hidden markov model to predict number of pm2.5 exceedance days in tehran during 2016–2017," *Avicenna Journal of Environmental Health Engineering*, vol. 4, no. 1, p. 58031, 2017.