Air Pollution Dynamics Modeling

Gelsomino Ludovico

Introduction

Air pollution is a significant public health issue, particularly due to its impact on respiratory diseases. This project aims to model air pollution dynamics using hourly air quality data from the U.S. Environmental Protection Agency (EPA), focusing on ten stations along the U.S. West Coast during the summer of 2020. We will explore how to identify and estimate different pollution levels and their stability, predict high pollution persistence, and provide real-time estimation using streaming data. Additionally, we will examine whether using aggregated data (e.g., daily averages) can reduce noise and if incorporating spatial dimensions into our analysis can provide further insights. The dataset includes spatial coordinates, timestamps, PM2.5 concentrations, temperature, wind speed, and station IDs. Our analysis will involve describing the time series data, applying Gaussian Hidden Markov Models (HMM), providing online estimation and prediction, and incorporating spatial dimensions using multivariate dynamic linear models (DLM).

Part I

Hidden Markov Models (HMMs) are exceptionally well-suited for modeling PM2.5 levels, offering distinct advantages due to their ability to capture the underlying state-dependent processes that are not directly observable. These models operate by assuming that the observed data are generated by a system transitioning between a number of hidden states, each representing a different regime of air quality conditions. The first part of the project will involve the estimation of several Hidden Markov Models (HMMs). The initial set of models will be univariate, assuming that the process has two and three underlying states. Subsequently, the second set of models will be multivariate, incorporating wind speed and temperature, and will still be estimated under two and three hidden states. For this part of the project, we will focus on station 47.

The Data

As illustrated in Figure 1, which shows PM2.5 levels and temperature, we observe distinct patterns that are pertinent to environmental studies. The PM2.5 graph displays sharp, isolated spikes indicating episodes of high particulate matter concentration, likely due to specific transient environmental or anthropogenic activities. Conversely, the temperature graph shows regular diurnal fluctuations without a significant long-term trend, characteristic of daily thermal cycles. Additionally, wind speed exhibits large variations between the minimum and maximum ranges.

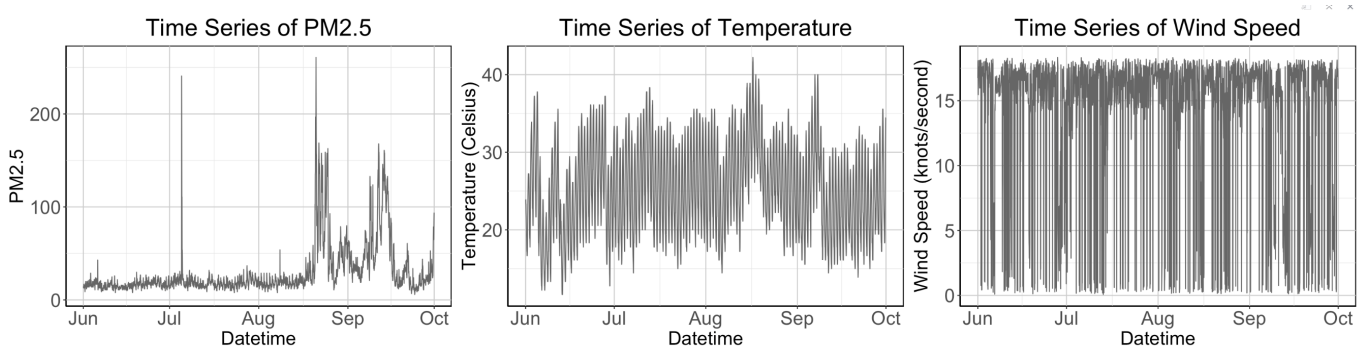

Figure 1: Time Series Plots

Univariate Hidden Markov Models

In this section we will estimate two univariate models. The first model will assume two hidden states, while the second model will assume three hidden states. The estimated Transition probabilities of the Hidden Markov Models are:

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \left\| \begin{array}{cc} 0.992 & 0.008 \\ 0.021 & 0.979 \end{array} \right\| \end{array} \quad \mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left\| \begin{array}{ccc} 0.968 & 0.000 & 0.032 \\ 0.000 & 0.938 & 0.062 \\ 0.023 & 0.082 & 0.895 \end{array} \right\| \end{array}$$

The estimated means and standard deviation of the hidden states are reported in the Table 1.

Model	State	Re1.(Intercept)	Re1.sd
Univariate HMM (Two States)	St1	17.533	4.286
	St2	63.279	38.832
Univariate HMM (Three States)	St1	70.299	39.023
	St2	15.011	2.580
	St3	22.495	4.294

Table 1: Response Parameters for Univariate HMMs

The univariate Hidden Markov Model (HMM) [Figure 2] with two states effectively captures the temporal dynamics of PM2.5 levels. In this model, State 1 represented by orange dots, captures lower PM2.5 values and dominates throughout most of the period, indicating consistent lower pollution levels. State 2, represented by red dots, corresponds to higher PM2.5 values and appears intermittently, often during significant pollution peaks. The observed PM2.5 values, represented by grey lines and points, show significant variability with occasional high spikes. The two-state model successfully differentiates between periods of high and low PM2.5 levels, with State 1 capturing critical pollution events and State 2 representing the more frequent, cleaner periods.

In contrast, the univariate HMM with three states [Figure 2] offers a more nuanced differentiation between pollution levels. State 1, represented by orange dots, captures the highest PM2.5 values and appears during extreme pollution events. State 2, represented by red dots, captures the lowest PM2.5 values and dominates during most periods, indicating low pollution levels. State 3, represented by blue dots, corresponds to intermediate PM2.5 values, indicating moderate pollution levels. The observed PM2.5 data shows variability with peaks and troughs, where high spikes are captured by State 1, baseline lower values by State 2 and moderate values by State 3.

Overall, the two-state model simplifies the analysis by categorizing data into high and low pollution levels, suitable for a basic understanding of pollution dynamics. In contrast, the three-state model offers a detailed analysis by adding an intermediate state, providing a more refined understanding of pollution severity and its temporal dynamics.

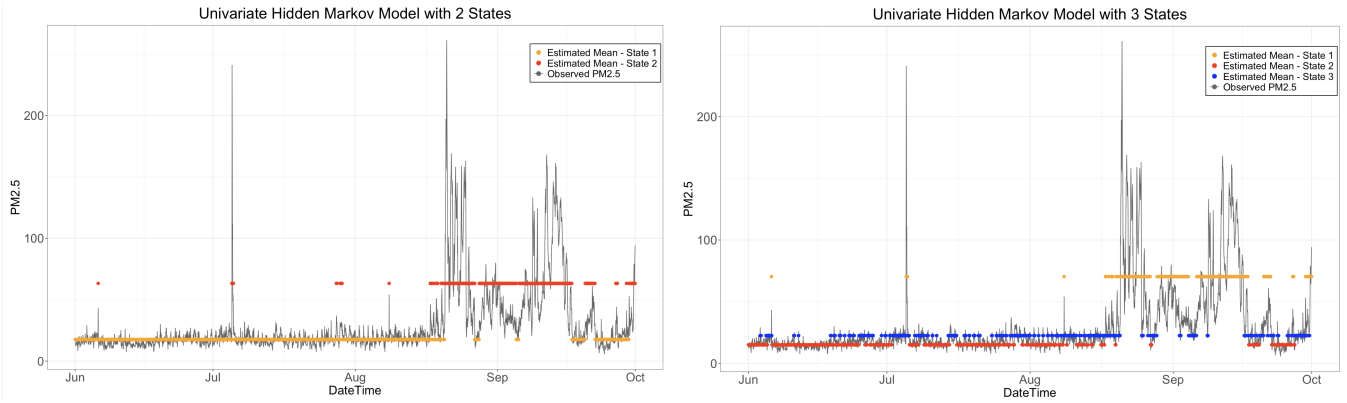

Figure 2: Univariate Hidden Markov Models

Multivariate Hidden Markov Models

In this section we will estimate two multivariate models. The first model will assume two hidden states, while the second model will assume three hidden states. The estimated Transition probabilities of the Hidden Markov Models are:

$$\mathbf{P} = \begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 0.980 & 0.020 \\ 2 & 0.008 & 0.992 \end{array} \qquad \mathbf{P} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0.971 & 0.029 & 0.000 \\ 2 & 0.023 & 0.916 & 0.061 \\ 3 & 0.000 & 0.017 & 0.983 \end{array}$$

The estimated means and standard deviation of the hidden states are reported in Table 2.

Model	State	Re1.(Intercept)	Re1.temp	Re1.wind	Re1.sd
Multivariate HMM (Two States)	St1	69.351	0.164	-0.767	38.615
	St2	12.554	0.197	0.000	4.122
Multivariate HMM (Three States)	St1	65.076	1.173	-0.274	37.631
	St2	21.728	0.435	-0.016	8.165
	St3	11.625	0.203	0.015	3.597

Table 2: Response Parameters for Multivariate HMMs

The multivariate Hidden Markov Model (HMM) with two states [Figure 3], incorporating covariates such as temperature and wind, effectively captures the dynamics of PM2.5 levels. In this model, State 1, represented by orange dots, captures higher PM2.5 values, identifying significant pollution events and more frequent spikes in the data. Conversely, State 2, represented by red dots, corresponds to lower PM2.5 values and predominantly occurs during periods of lower pollution levels. The observed PM2.5 values, depicted by grey lines and points, exhibit considerable variability with occasional high spikes. This two-state model successfully distinguishes between periods of high and low PM2.5 levels, providing a nuanced understanding of air quality dynamics by incorporating influential covariates.

In contrast, the multivariate HMM with three states [Figure 3] offers an even finer differentiation of pollution levels by including the same covariates. State 1, represented by orange dots, identifies the highest PM2.5 values, capturing extreme pollution events and significant spikes. State 2, represented by red dots, corresponds to intermediate PM2.5 values, reflecting moderate pollution levels that occur more frequently. State 3, represented by blue dots, captures the lowest PM2.5 values, indicating the cleanest periods with minimal pollution. The observed PM2.5 data, shown by grey lines and points, reveal variability with peaks and troughs, where high spikes are captured by State 3, moderate values by State 2, and baseline lower values by State 1.

Overall, the two-state model, with its integration of covariates, categorizes data into high and low pollution levels, improving the comprehension of pollution dynamics. The three-state model further refines this analysis by adding an intermediate state, allowing for a more accurate and detailed understanding of pollution severity and its temporal fluctuations.

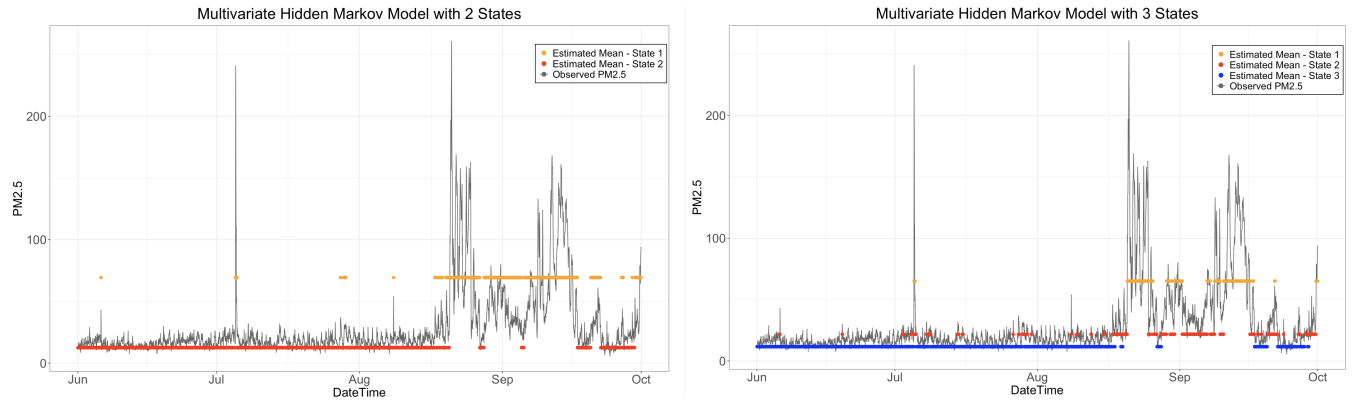

Figure 3: Multivariate Hidden Markov Models

Remarks

In this first part we successfully used Hidden Markov Models (HMMs) to identify and estimate different levels of pollution from the data, with univariate HMMs classifying based on PM2.5 time series data and multivariate HMMs incorporating additional covariates like temperature and wind for improved accuracy. The models differentiate between states representing high, moderate, and low pollution levels. HMMs also provide state transition probabilities, which can predict the likelihood of remaining in a high pollution state or transitioning to a lower state in the next hour or over a few hours. By summing these transition probabilities, we can quantify the probability of a significant decrease in pollution, offering valuable insights for real-time monitoring and decision-making in managing air quality.

Part II

Dynamic Linear Models (DLMs) are particularly beneficial for modeling such PM2.5 levels due to their flexibility in adapting to time-varying dynamics, which is essential given the fluctuating nature of air pollutants influenced by various emission sources and weather conditions. Moreover, DLMs can handle missing data effectively—an important feature for environmental datasets prone to sensor issues or data loss. These models can also integrate multiple data sources, enhancing the model's accuracy by including additional variables such as wind speed and temperature.

The model

One of the simplest Dynamic Linear Models (DLMs) is the random walk plus noise model, also called the first-order polynomial model. This model is used to handle univariate observations where the state vector is unidimensional. It is described by the following equations:

$$\begin{cases} y_t = \theta_t + v_t, & v_t \sim \mathcal{N}(0, V) \\ \theta_t = \theta_{t-1} + w_t, & w_t \sim \mathcal{N}(0, W) \end{cases}$$

The model considered in our case is constant, i.e., the various matrices defining its dynamics are time-invariant and the only parameters of the model are the observation and evolution variances V and W. In our case, we will estimate them from the available data using maximum likelihood.

Parameter Estimation and Forecasting

Before proceeding with maximum likelihood we will apply a transformation of the data. The transformation consists in re-scaling the data to the logarithmic scale and taking the 12-hours averages [Figure 4].

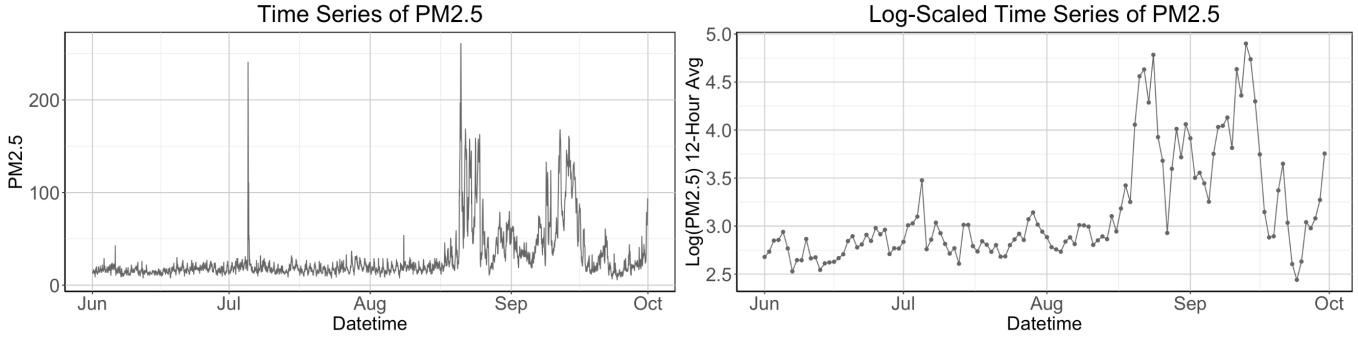

Figure 4: Original and Log-Scaled $PM_{2.5}$ Time Series

The estimated parameters for the observation variance V and the process variance W are reported in Table 3.

Parameter	V	W
Estimate	0.00287	0.07175
ASE	2.42978	0.22671

Table 3: Estimated Parameters and Asymptotic Standard Errors for the Dynamic Linear Model

The estimate for V, which represents the observation noise variance, is relatively low at 0.00287, indicating that the observations are assumed to have minimal inherent noise. However, the associated ASE of 2.42978 is considerably larger than the estimate itself, suggesting a significant degree of uncertainty in this parameter's estimation. Conversely, the estimate for W, representing the process variance in the random walk model, is 0.07175, suggesting a moderate level of variability in the process dynamics. The ASE for W is 0.22671, which, while smaller than the ASE for V, is still substantial relative to the estimate of W.

Using these estimates, we generated one-step-ahead forecasts and future prediction [Figure 5].

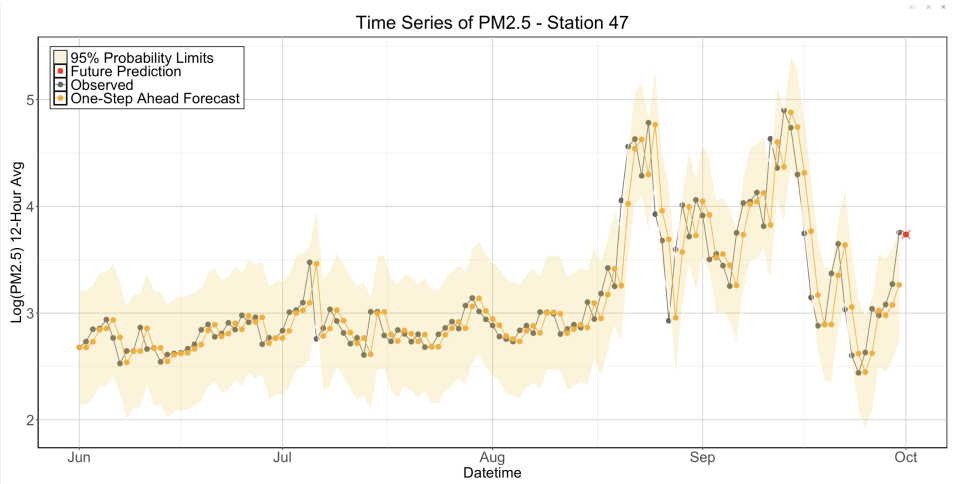

Figure 5: Time Series with One-Step Ahead Forecast and Prediction

Figure 5 illustrates the log-transformed PM2.5 12-hour average concentrations at Station 41, enhanced with one-step-ahead predictions and future prediction point, alongside 95% probability limits. The one-step-ahead predictions, represented by orange dots, align closely with the observed values, demonstrating the model's efficacy in capturing short-term fluctuations in PM2.5 levels. The future prediction (3.74), marked in red, extends beyond the available data, offering insights for preemptive air quality management. The shaded region around the predictions denotes the 95% confidence intervals, providing a measure of uncertainty in the forecasts. The narrowness of this band at certain intervals suggests a high confidence level in the model's predictions under those conditions.

Remarks

Upon evaluating the model's performance and underlying assumptions, our satisfaction is moderated by the model's simplicity and inherent limitations. The random walk plus noise model assumes linear transitions and normally distributed errors—assumptions that simplify calculations but may not capture more complex or skewed data distributions effectively. Additionally, the model's time-invariant nature could be restrictive in scenarios with evolving dynamics due to external factors such as seasonal variations or regulatory changes. Comparatively, Hidden Markov Models (HMMs), which we utilized previously, offer a more flexible framework by accommodating multiple hidden states and the potential for complex transitions, making them better suited for datasets exhibiting categorical shifts or underlying non-observable processes. This flexibility makes HMMs preferable for modeling more intricate systems where data may deviate from normality or linear patterns.

Part III

In our ongoing exploration of air quality modeling, we now aim to extend our approach to incorporate a spatial dimension into our analysis, addressing the third set of previously presented questions. This addition is crucial as it allows us to consider the interactions between different locations in our study. We propose a spatially-aware model using the familiar framework of a random walk, now adapted to handle multi-dimensional data for PM2.5 measurements collected from various monitoring stations.

Data

The log-scaled PM2.5 time series data from selected monitoring stations (IDs 41, 47, 96, and 99) from June to October reveal notable temporal patterns and suggest underlying spatial dependencies in particulate matter concentrations [Figure 6]. Each station displays synchronized fluctuations in PM2.5 levels, characterized by concurrent peaks and troughs, which may indicate the influence of regional environmental events such as changes in weather patterns or long-range pollutant transport. The alignment in the timing of these variations across stations, despite differences in magnitude, supports the presence of spatial dependence. It's important to note that the closer the stations are to each other, the more similar their observations tend to be. This observation is quite natural, as stations in proximity are likely to record similar PM2.5 values due to shared environmental conditions.

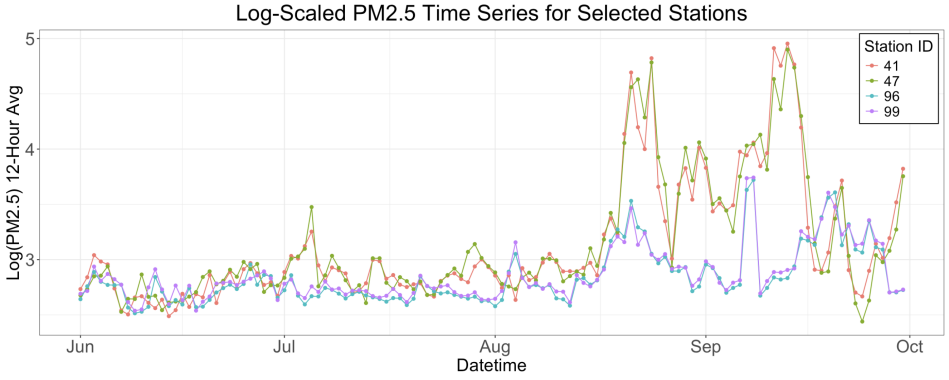

Figure 6: Time Series of $PM_{2.5}$

Model

We will consider a multivariate model for the m-dimensional $\mathbf{Y}_t = (Y_{j1,t}, \ldots, Y_{jm,t})'$, for us, the $PM_{2.5}$ observed at stations $j_1, \ldots, j_m$ where each $(Y_{j,t})$ is described as a random walk plus noise.

The model can be written as:

$$\begin{cases} Y_t = F\theta_t + v_t, & v_t \overset{\text{indep}}{\sim} \mathcal{N}_m(\mathbf{0}, V) \\ \theta_t = G\theta_{t-1} + w_t, & w_t \overset{\text{indep}}{\sim} \mathcal{N}_p(\mathbf{0}, W) \end{cases}$$

In our case, the measurement errors are assumed to be independent across location, so that V is a diagonal matrix with entries:

$$V = \begin{bmatrix} \sigma_{v,1} & 0 & 0 & 0 \\ 0 & \sigma_{v,2} & 0 & 0 \\ 0 & 0 & \sigma_{v.3} & 0 \\ 0 & 0 & 0 & \sigma_{v,4} \end{bmatrix}$$

Spatial dependence is modelled through evolution errors, so that the diagonal matrix is not diagonal and has entries:

$$W[j,k] = \text{Cov}(w_{j,t}, w_{k,t}) = \sigma^2 \exp(-\phi D[j,k]), \quad j,k = 1, \ldots, m,$$

Estimation via Maximum Likelihood

Letting $F = I_m$ and $G = I_m$, our model becomes:

$$\begin{cases} Y_t = \theta_t + v_t, & v_t \overset{\text{indep}}{\sim} \mathcal{N}_m(\mathbf{0}, V) \\ \theta_t = \theta_{t-1} + w_t, & w_t \overset{\text{indep}}{\sim} \mathcal{N}_p(\mathbf{0}, W) \end{cases}$$

Due to the structure of V and W, the parameters to estimate are: $\sigma_{v,1}$, $\sigma_{v,2}$, $\sigma_{v,3}$, $\sigma_{v,4}$, σ^2 and ϕ. The maximum likelihood estimates and their Asymptotic Standard Errors are reported in Table 4.

Parameter	$\sigma_{v,1}$	$\sigma_{v,2}$	$\sigma_{v,3}$	$\sigma_{v,4}$	σ^2	ϕ
Estimate	0.00644	0.00610	0.00038	0.00071	0.04580	0.16395
ASE	0.32050	0.32022	1.38401	0.75428	0.10769	0.16714

Table 4: Estimated Parameters and Asymptotic Standard Errors

The asymptotic standard errors (ASEs) for $\sigma_{v,1}$ and $\sigma_{v,2}$ are notably high at 0.32050 and 0.32022, respectively, compared to their parameter estimates. The ASE for $\sigma_{v,3}$, at 1.38401, significantly exceeds its parameter estimate, severely undermining confidence in this estimate and suggesting that the model may be excessively sensitive to variations in the dataset or that the data does not robustly support the state associated with this variance component. Conversely, the ASEs for σ^2 and ϕ, at 0.10769 and 0.16714 respectively, are reasonably proportional to their estimates, indicating more reliable estimations. These proportions suggest that the model effectively captures the overall dynamics and decay characteristics of the process, enhancing confidence in these aspects of the model's structure.

Given the MLE of the model parameters, the resulting Matrices V and W are:

$$V = \begin{bmatrix} 0.0064 & 0 & 0 & 0 \\ 0 & 0.0061 & 0 & 0 \\ 0 & 0 & 0.00038 & 0 \\ 0 & 0 & 0 & 0.000713 \end{bmatrix} \quad W = \begin{bmatrix} 0.0458 & 0.0421 & 0.0167 & 0.0165 \\ 0.0421 & 0.0458 & 0.0181 & 0.0179 \\ 0.0167 & 0.0181 & 0.0458 & 0.0447 \\ 0.0165 & 0.0179 & 0.0447 & 0.0458 \end{bmatrix}$$

Note the significantly larger covariance between stations in close proximity compared to those farther apart. For example, stations 41 and 47 exhibit high covariance, whereas stations 41 and 96 display much lower covariance. A similar pattern is observed between stations 91 and 99. This observation is attributed to the proximity of the stations, which typically results in more similar data patterns. This relationship is also evident in Figure 6, where nearby stations generally record comparable levels of PM2.5.

The distinct difference in the magnitudes of V and the strong correlations in W suggest that while the measurement precision might differ across stations, the dynamics of PM2.5 levels are interconnected across the geographic spread of these stations. This fitting highlights the model's capability to capture both the individual characteristics of each station's data (through V) and the broader regional dynamics affecting multiple stations (through W).

Forecasting

Figure 7 illustrate the time series of log-transformed PM2.5 12-hour average concentrations for four monitoring stations (Station 41, Station 47, Station 96, and Station 99), encompassing observed data points, one-step-ahead forecasts, and a future prediction. For each station, the observed data, depicted by grey dots, reveal significant variability with distinct spikes indicative of transient pollution events. The one-step-ahead forecasts, represented by orange dots, closely follow the observed data, underscoring the model's proficiency in capturing short-term fluctuations. The future prediction, marked by a red dot, extends beyond the observed data and provides insights into anticipated pollution levels. This consistency across all stations demonstrates the model's robustness in predicting immediate future values. The close alignment between forecasts and observations at each station reflects the model's adaptability to localized air quality dynamics. However, variability in asymptotic standard errors and parameter estimates suggests a need

for further refinement to enhance predictive accuracy and robustness, particularly for long-term forecasts. Overall, the spatial-temporal DLM effectively handles the temporal and spatial dynamics of PM2.5 levels, offering reliable short-term predictions while highlighting areas for potential improvement.

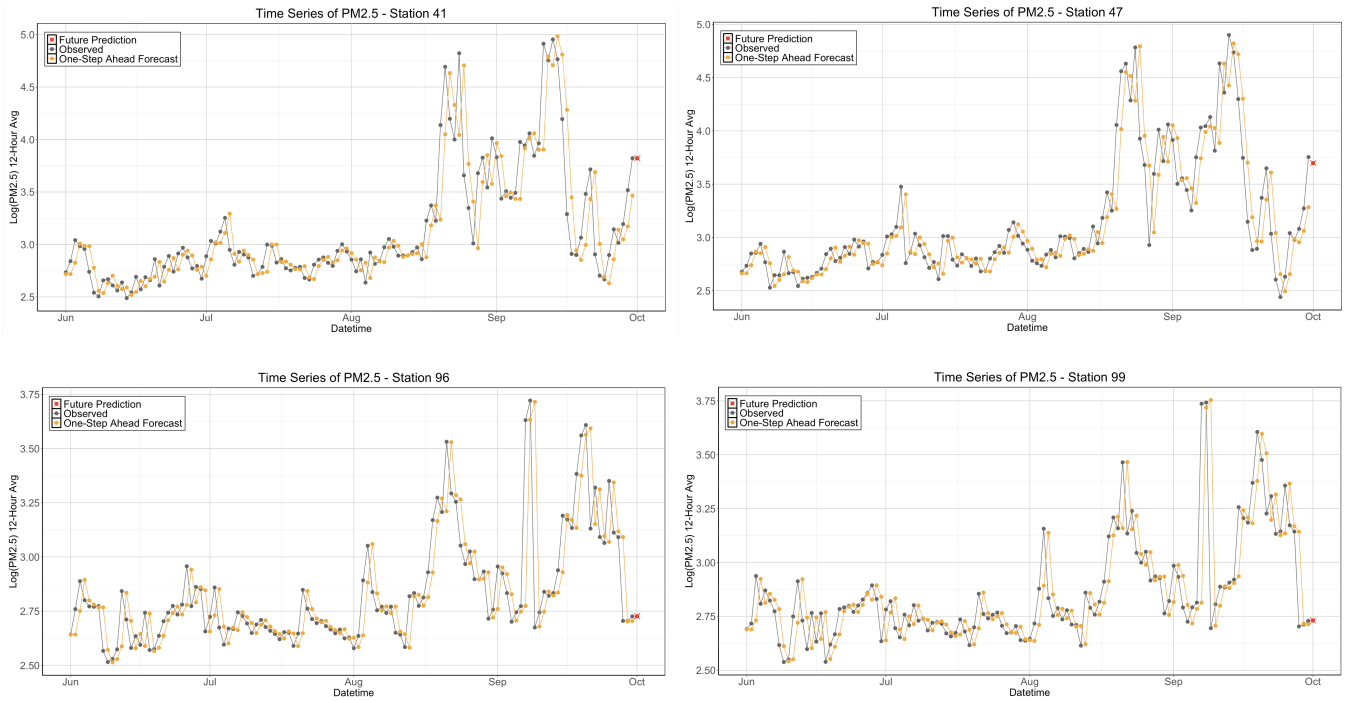

Figure 7: Time Series with One-Step Ahead Forecast and Prediction

Remarks

The spatial-temporal Dynamic Linear Model (DLM) demonstrated significant improvements over the univariate model in capturing and predicting PM2.5 levels across multiple stations. By integrating spatial dependencies, the model leverages information from neighboring stations, resulting in more accurate and reliable forecasts. The close alignment between the one-step-ahead forecasts and observed data points across all four stations (Station 41, Station 47, Station 96, and Station 99) indicates that the spatial-temporal DLM effectively accounts for localized air quality dynamics and provides a robust short-term prediction framework. The model's enhanced predictive accuracy and robustness across different monitoring stations highlight its superiority over the univariate model, which often struggles with site-specific variances. Furthermore, the spatial-temporal DLM offers a holistic view of air quality trends by incorporating data from multiple stations, thus capturing regional pollution patterns that a univariate model might miss.

However, the spatial-temporal DLM assumes that spatial dependencies between stations are accurately captured by a distance-based covariance structure, which may be restrictive in areas with complex geographical or meteorological influences. Despite these assumptions, the spatial-temporal DLM shows superior performance in forecasting PM2.5 levels, as evidenced by the tighter alignment of predictions with observed data and more reliable uncertainty estimates. Overall, the spatial-temporal DLM offers significant advancements in predictive accuracy and robustness, making it a valuable tool for environmental monitoring, though future enhancements could explore more complex spatial structures or incorporate additional environmental variables to further refine its predictive capabilities.