Time Series Analysis of Economic-financial Data
Final Project

Group 6 - Del Frari Elisa, Durmush Derya, Vastola Giorgia

May 2024

1 Introduction

This paper delves into the analysis of air quality data, focusing on PM2.5 levels collected from station 96 of the U.S. Environmental Protection Agency (EPA), located along the U.S. West Coast during the summer of 2020. Given the inherent temporal dependencies in PM2.5 levels, our approach employs time series models to effectively capture these dynamics. Leveraging these models, we aim to discern varying pollution levels, generate real-time forecasts for streaming data and explore spatial dependencies among different monitoring stations.

2 Exploratory Analysis

The dataset consists in hourly data from 10 Californian stations spanning from June 1st to September 30th, 2020, encompassing a total of 29,280 observations, with no missing values. Among these, 2,928 observations originate from station 96. The variables under scrutiny include station ID, longitude, latitude, datetime (timestamp in GMT timezone), pm25 (particulate matter of size 2.5 micrograms per cubic meter or less), temperature (in Celsius), and wind speed (measured in knots per second).

The time plot of PM2.5 levels, wind speed and temperature reveals that these three environmental variables exhibit parallel patterns and simultaneous shocks. For instance, the spike in PM2.5 levels observed at the beginning of September 2020 coincides with a drastic decrease in temperature, suggesting a potential causal relationship. These associations are quantified in the correlation matrix, which indicates a negative association between temperature and wind (-0.28), while also revealing a very weak correlation between PM2.5 levels and both wind and temperatures (0.06 and 0.03, respectively). The time series of interest, depicted in Figure 1, exhibits non-stationary behavior, as mean and variance are not constant.

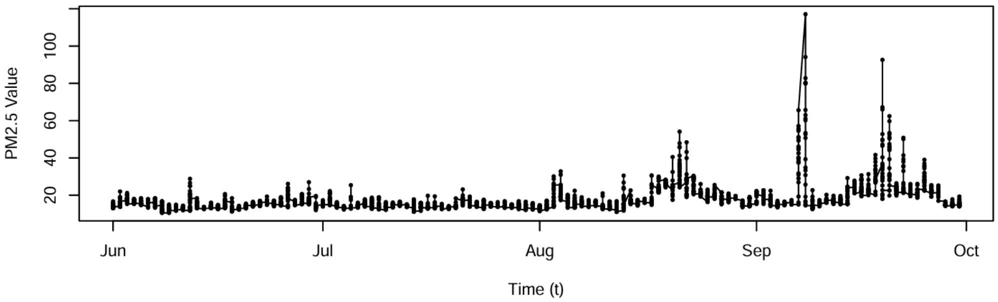

Figure 1: Time plot of PM2.5 levels at station 96

To analyse the correlation between the series and its own lagged values at different time lags, the Autocorrelation Function (ACF) is produced. The plot reveals a steadily decrease in autocorrelation, indicative of an autoregressive (AR) process. On the other hand, the PACF, which also controls for the effects of shorter lags, shows a high value at lag 1, a significant negative value at lag 2 and a weaker influence of lags 3, 4 and 5 on the current value. These observations suggest a AR(5) model, with some potential negative autoregressive parameters. The temporal dependencies within the variable of interest are further examined using a lag plot of PM2.5 levels for lags 1, 2, and 3. The plot reveals a clustering of data points along the diagonal line, indicating again a degree of autocorrelation within the data. However, as the lag increases, the density of points along this diagonal diminishes, suggesting a weakening autocorrelation strength with increasing time intervals between observations. These patterns imply that past values possess some predictive capability for future values, a foundational premise in time series analysis. It suggests that the data is not purely stochastic but may exhibit underlying patterns that can be modelled. Additionally, deviations from the diagonal line, particularly evident at higher lags, hint at potential non-linearities in the data.

Furthermore, day, week, and month decompositions of time series of PM2.5 levels are produced. The daily decomposition yields less autocorrelation in its residuals, as observed in the ACF and PACF plots of the residuals, suggesting a more effective modeling of the data's underlying patterns at the daily frequency.

3 Modelling

3.1 Hidden Markov Model

To identify different levels of pollution and predict persistence or decrease of its levels, a Hidden Markov Model is employed, as it allows to model change points and transitions between different states.

3.1.1 Model specification

To formalize, a hidden Markov model is a discrete-time stochastic process $((Y_t, S_t))_t$, where the observable time series $(Y_t)_{t\geq 1}$ depends on the state of a non-observable hidden process, $(S_t)_{t\geq 0}$, called state process. In this application, the time series $(Y_t)_{t\geq 1}$ is given by the observed PM2.5 concentrations, which depend on the state process $(S_t)_{t\geq 0}$, modelled as a homogeneous Markov chain, with state space S={1,2,..,k}. The emission distribution, namely the distribution of the observable variable Y_t given the state $S_t = j$ for j={1,2,..,k} is assumed to be Gaussian with state-specific parameters.
Initially, a two-state HMM is employed. However, this approach is inadequate to capture the full spectrum of variability present in the dataset. Notably, the highest state, observed at around 25 micrograms per cubic meter, corresponds to the actual prescribed limit. Subsequently, a three-state HMM is fitted to categorize emissions into low, moderate, and high levels. Despite this refinement, the high state remains centered near the limit, at 26.563 micrograms per cubic meter, again failing to capture exceptionally high levels. Subsequently, a four-state HMM is evaluated. In addition to yielding a higher log-likelihood, this model effectively captures pollution levels exceeding the prescribed limit, specifically around 36 micrograms per cubic meter. Thus, the model that satisfactorily meets our requirements and aligns with the goals of our analysis, is composed as follows:

$$Y_t = \begin{cases} \mu_1 + \epsilon_t & \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_1^2) & \text{if state } S_t = 1 \\ \mu_2 + \epsilon_t & \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_2^2) & \text{if state } S_t = 2 \\ \mu_3 + \epsilon_t & \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_3^2) & \text{if state } S_t = 3 \\ \mu_4 + \epsilon_t & \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_4^2) & \text{if state } S_t = 4 \end{cases}$$

3.1.2 Model estimation

The unknown parameters for which Maximum Likelihood Estimates are sought are: the initial probability law π of S_0; the state transition matrix $A = [p_{ij}]$, where $p_{ij} = P(S_t = j|S_{t-1} = i)$ for $i, j = 1, .., 4$; the parameters of the emission distributions $(\mu_1, \sigma_1^2,\ \mu_2, \sigma_2^2,\ \mu_3, \sigma_3^2,\ \mu_4, \sigma_4^2)$.

Table 1: Estimated parameters of emission distributions of 4-state HMM

	Estimated mean	Estimated sd
State1	15.987	0.933
State2	13.639	0.883
State3	21.142	2.450
State4	36.525	13.812

Figure 2 plots the estimated means alongside with the original time series and Table 1 presents the estimated mean and standard deviation of the Gaussian emission distributions for each state. The table shows that the estimated mean for the high-emission state (State 4) is 36.535, with a standard deviation of 13.812. In contrast, the low-emission state (State 2) has an estimated mean of 13.689, with a standard deviation of 0.883. This indicates that higher emission levels are associated with significantly greater variability. Additionally, the estimated initial probability law π of S_0 indicates that the system starts in State 2 (the low scenario) with a probability of 100%, which reflects the observed temporal pattern of the data.

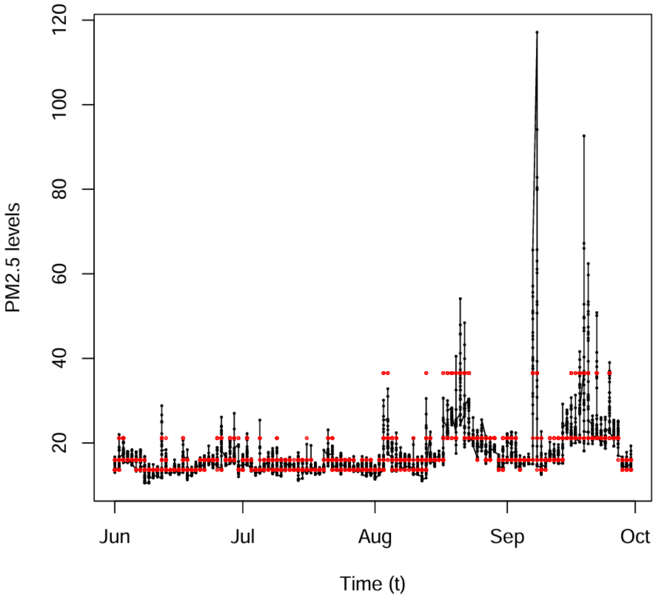

Figure 2: Time plot of PM2.5 levels and 4-state HMM estimated means

The estimates of the emission probabilities (the likelihood of observing different pollution levels given each state) and the transition probabilities (the likelihood of transitioning from one state to another, as shown in Table 2) are useful tools to aid researchers and decision makers to better understand the phenomenon and make better-informed decisions regarding the timing and type of policies to implement. For instance, forecasting the persistence of high pollution levels in the subsequent hour becomes feasible by analyzing transition probabilities. In fact, when the system is experiencing a high level of pollution, one can examine the transition probabilities from the current high pollution state to all other states (last row of Table 2). If the probability of remaining in the high pollution state surpasses that of transitioning to any other state, it suggests a prospensity for the current environmental conditions to persist without significant change. Specifically, the estimated transition matrix indicates a 93% probability of remaining in a state of very high pollution.

Table 2: Estimated state transition matrix of 4-state HMM

	to S1	to S2	to S3	to S4
from S1	0.913	0.049	0.038	0.000
from S2	0.051	0.947	0.002	0.000
from S3	0.000	0.005	0.912	0.025
from S4	0.000	0.000	0.073	0.927

On the other hand, when the aim is to quantify the probability to see a significant decrease in the next few hours, a multi-step prediction approach is needed. In fact, a significant decline over a short period can happen in different ways, for instance through a one-step transition from very high to low pollution or through a more gradual decline. For example, in a scenario where pollution levels are currently at their peak, one can assess the likelihood of a gradual decline by determining the probability of transitioning from state 4 to state 3, then to state 1, and finally to state 2 over the next three hours. This probability calculation can be formulated as follows:

$$P^{(3)}_{4\to 2} = P_{4,3} \cdot P_{3,1} \cdot P_{1,2} = 0.073 \cdot 0.000 \cdot 0.049 = 0$$

The situation becomes more complex as one examines the transition from S4 to S2 in 3 or more hours, considering all the combinations of paths in the middle. In this case, what should be done is to consider the probabilities of all the possible paths to S4. Following that, for every potential path, the transition probabilities step by step would be multiplied. The final step would be to aggregate the probabilities of all such paths that end in S4.

3.2 Dynamic Linear Model

Accurately estimating and forecasting streaming data in the scenario under study is crucial for timely implementing effective mitigation strategies. One highly effective approach to address this challenge involves the application of state space models, particularly dynamic linear models. Indeed, while Hidden Markov Models (HMMs), assuming a finite number of hidden states, excel in identifying distinct pollution levels and modeling transitions between them, DLMs offer greater flexibility by allowing for continuous state space and are particularly useful for scenarios where real-time forecasting is essential, due to the sequential updating of parameters and states based on incoming data.

Before advancing to the modeling steps, a series of elaborations is conducted on the data. Due to the presence of sharp peaks in the data, a logarithmic transformation is applied. As a result, the range of values is compressed, shrinking larger values while expanding smaller ones. Despite this transformation, the underlying data appears to maintain its inherent structure. In addition, to mitigate the effects of noise and irregularities present in hourly data, experimentation with coarser scales is conducted, namely 12, 24, 36 and 48-hour averages. In the proceeding analysis, the latter option is utilized, as it yields uncorrelated model residuals.

3.2.1 Model specification

The modelling approach consists in a random walk plus noise, represented by the following system of equations:

$$\begin{cases} y_t = \theta_t + v_t, & \text{where } v_t \sim \mathrm{N}(0, \sigma_v^2) \\ \theta_t = \theta_{t-1} + w_t, & \text{where } w_t \sim \mathrm{N}(0, \sigma_w^2) \end{cases}$$

with $\theta_0 \sim N(m_0, C_0)$ and $\theta_0 \perp (w_t) \perp (v_t)$ both within them and between them.

3.2.2 Model estimation

As a first step, maximum likelihood estimation is performed to estimate the unknown parameters of the model of interest, σ_v^2 and σ_w^2. Parameter inference and predictions is conducted using the same data, a practice that can lead to overfitting in predictions. However, in the context of state models, this approach is often employed. Two different optimization algorithms are evaluated: the default L-BFGS-B algorithm and Simulated Annealing. While the first is better suited for global optimization problems where the objective function may have multiple minima, the second is more appropriate for local optimization of smooth, convex functions. Despite this difference, both algorithms return approximately the same optimum value in our application. Thus, we opt for the default algorithm. Additionally, lower bounds for the variances are set at $\sigma_v^2 \geq 0.0001$ and $\sigma_w^2 \geq 0$. Several sets of initial parameter values are tested, including $(0.001, 0.001)$, $(0.01, 0.01)$, $(0.1, 0.1)$, $(1, 1)$. The likelihood is optimized with the parameter values $(0.001, 0.001)$, though the estimates and likelihood values for all tested cases are nearly identical, with differences as minimal as 0.0001.

To quantify the uncertainty in the estimated parameters, the asymptotic covariance matrix of the estimates is computed as the inverse of the Hessian matrix obtained during the optimization process and that contains the second-order partial derivatives of the negative log-likelihood function with respect to the parameters. The standard errors of the MLEs are computed by extracting the square root of the diagonal elements of the asymptotic covariance matrix. These estimates are shown in Table 3. The MLE of σ_w^2 is 0.0064 with a standard error of 0.0070, while the MLE of σ_v^2 is 0.0247 with a standard error of 0.0097. This indicates that the underlying state is relatively stable while the observed data are subject to higher random fluctuations. Since the observation noise is greater than the variability in the true signal over time, the signal-to-noise ratio is estimated to be smaller than 1, specifically 0.26. This ratio influences the adaptive coefficient. Consequently, the model places less trust in the observations and adapts the state estimates more slowly to changes in the observed data, providing smooth state estimates as it assumes that the true state is relatively stable and that changes are more likely due to observation noise.

	σ_v^2	σ_w^2
MLE	0.0247	0.0064
Standard error	0.0097	0.0070

Table 3: MLEs and standard errors of the unknown variances σ_v^2, σ_w^2 (univariate DLM)

3.2.3 Forecasting

Next, the MLEs are plugged in the dynamic linear model and one-step-ahead forecasts are computed. By aggregating data into 48-hour averages, short-term fluctuations are smoothed out, allowing the model to focus on capturing and predicting the underlying trends. In Figure 3, the forecasts exhibit a close correspondence with the observed data trends, suggesting that the model accurately captures the general patterns and fluctuations in PM2.5 levels. Despite the overall accuracy of the forecasts, significant deviations are observed where the model appears to lag in response to the peaks observed towards the end of summer 2020. This lag can be attributed to the sudden and disruptive nature of the fire-induced shock, which perturbs the typical patterns and assumptions of the state space model, including constant variance. Consequently, the model encounters challenges in accurately predicting the affected portion of the time series, as the random walk plus noise framework struggles to account for substantial deviations from the anticipated trajectory. Furthermore, it is important to recall that the influence of the adaptive coefficient results in forecasts adapting more gradually to changes.

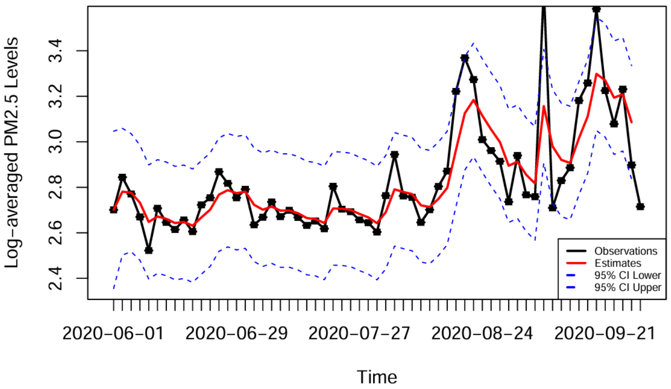

Figure 3: One-step-ahead forecasts and credible intervals of log-averaged PM2.5 levels at station 96 (univariate DLM)

3.2.4 Model checking

To check model validity, forecast errors are analyzed. In particular, a plot of the autocorrelation function (ACF) is generated to verify the assumption of independence of the residuals. Autocorrelated residuals violate the assumption of independence, potentially resulting in inaccurate statistical inference. The ACF plot reveals no significant correlation. As previously mentioned, uncorrelated residuals are achieved only after averaging the results over 36 and 48 hours, indicating that short-term fluctuations are the primary cause of autocorrelation. Moreover, the p-values of the Ljung-Box test are above 0.05. This indicates that the assumption of independence among residuals is met. Thus, we are satisfied with the validity of the model. Furthermore, to check the assumption that model's residuals follow a Gaussian distribution, a QQ plot is produced. The plot reveals that the standardized innovations follow the reference line, especially in the center of the distribution. However, there are some deviations, particularly in the tails of the distribution, indicating that the distribution assigns different probabilities to the tails compared to the Gaussian distribution. Such deviations may imply that the Normality assumption could be overly restrictive or that the local level model might be too simplistic for the time series analyzed.

3.3 Spatio-temporal Analysis

This section expands on the previous model by incorporating multiple stations, allowing for the analysis of spatial dependencies among them. Specifically, time series from stations 96, 99, 41, 47 are jointly modelled.

The time plots and the correlation matrix reveal significant positive correlation between stations 96 and 99 (around 0.89) and between stations 47 and 41 (approximately 0.87). Conversely, the correlations between station 96 or station 99 and station 47 or station 41 are notably weaker (around 0.25). This spatial relationship is further confirmed by the Euclidean distances, computed from the latitudes and longitudes. Indeed, stations 96 and 99 are in close proximity, with a Euclidean distance of 0.13, while stations 47 and 41 are also closer, with a distance of 0.51. In contrast, the distance between stations 99 or 96 and stations 41 or 47 extends to approximately 5 units, indicating a notably higher spatial separation between these pairs.

Thus, the approach now consists in a multivariate model for the m-dimensional vector $\mathbf{Y}_t = (Y_{j1,t}, \ldots, Y_{jm,t})'$ of $\text{PM}_{2.5}$ levels observed at stations $j_1, \ldots, j_m$ at time t, where $m = 4$. Again, a random walk plus noise model is considered:

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

where F is a $m \times p$ identity matrix, G is a $p \times p$ identity matrix, $\theta_0 \sim N(m_0, C_0)$, $\theta_0 \perp (w_t) \perp (v_t)$ between them.
The measurement errors $v_{j,t}$ are independent across different locations, so the matrix V is diagonal:

$$V = \begin{bmatrix} \sigma^2_{v,1} & 0 & 0 & 0 \\ 0 & \sigma^2_{v,2} & 0 & 0 \\ 0 & 0 & \sigma^2_{v,3} & 0 \\ 0 & 0 & 0 & \sigma^2_{v,4} \end{bmatrix}$$

Conversely, the evolution errors $w_t = (w_{j1,t}, \ldots, w_{jm,t})$ are spatially correlated: the covariance matrix W is not diagonal, but it is modelled through the exponential covariance function:

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

where $\sigma^2 > 0$ and constant across locations; $\phi > 0$ is a decay parameter; and $D[j,k]$ is the distance between stations j and k, so that the $\mathrm{Cov}(w_{j,t}, w_{k,t})$ between two different stations j and k decreases as the distance between them increases.

3.3.1 Model estimation

As before, maximum likelihood estimation is performed to estimate the unknown parameters of the model of interest, namely $\sigma^2_{v,1}$, $\sigma^2_{v,2}$, $\sigma^2_{v,3}$, $\sigma^2_{v,4}$ (the observations variances at station 96, 99, 47, 41), σ^2_w (the state variance) and ϕ (the decay parameter). Additionally, lower bounds for the variances are set at 0.0001 for all the $\sigma^2_{v,i}$, $i = \{1,2,3,4\}$ and for ϕ and at 0 for σ^2_w. Multiple sets of initial parameter values for the variances are tested, including 0.001, 0.01, 0.1, and 1, while keeping the initial value for ϕ constant at 0.1. The likelihood optimization consistently converges to similar estimates, with nearly identical likelihood values across all tested cases. The standard errors of the MLEs are again computed by extracting the square root of the diagonal elements of the asymptotic covariance matrix. These estimates are shown in Table 4. The observation noise variances are notably reduced compared to the one in the previous model, suggesting that the observed data points are now considered more precise and close to the underlying state. Conversely, the estimated σ^2_w increases from 0.0064 in the previous model to 0.07497, indicating a higher level of uncertainty in the evolution of the underlying state process. Overall, the standard errors of the estimates of the observation variances experience a notable reduction with respect to the one of the univariate model. In contrast, there is a marginal increase in the standard error associated with the state variance, which augments from 0.007 to 0.01. Thus, the signal-to-noise ratio exceeds 1 for all stations.

	σ^2_{v1}	σ^2_{v2}	σ^2_{v3}	σ^2_{v4}	σ^2_w	ϕ
MLE	0.00010	0.00010	0.0001	0.00548	0.07497	0.13324
Standard error	0.00046	0.00046	0.00183	0.00190	0.01063	0.026296

Table 4: MLEs and standard errors of the unknown parameters: observations variances at stations 96, 99, 47, 41, state variance and ϕ (multivariate DLM)

3.3.2 Forecasting

Afterwards, the MLEs are plugged into the model to generate one-step-ahead forecasts. Figure 4 presents these forecasts for series 96, allowing for a comparison with the previous model. Compared to the previous model, the forecasts still closely track the observed observations, but this time they are significantly more precise, particularly in accurately capturing the sudden shocks observed at the end of summer 2020. This increased precision highlights the multivariate model's ability to leverage information and borrow strength from multiple time series by incorporating spatial dependence among the time series through the covariance structure of the evolution errors. In addition, given that the signal-to-noise ratio exceeds 1, forecasts now adapt more quickly to changes, as each new observation provides reliable information about the state of the system, due to the lower σ_v^2s values.

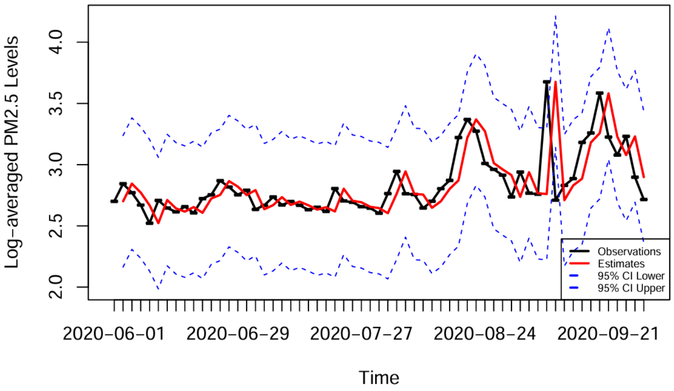

Figure 4: One-step-ahead forecasts and credible intervals of log-averaged PM2.5 levels at station 96 (multivariate DLM)

3.3.3 Model checking

The plot of the autocorrelation function (ACF) and the p-values of the Ljung-Box reveal no significant correlation, indicating that the assumption of independence among residuals is once again met. Thus, we are satisfied with the validity of the model. However, the QQ plot of the standardized residuals still shows deviations from the reference line at the tails of the distribution. Again, this might signal that the Normality assumption could be overly restrictive or that the current model specification is not correct.

4 Conclusions

In summary, this work employes a 4-state Hidden Markov Model to identify different pollution levels and quantify the likelihood of future pollution values, allowing for timely responses to anticipated pollution changes. Afterwards, real-time forecasts are generated using both a univariate DLM and a multivariate one that also accounts for spatial dependence among four stations. By explicitly modelling spatial dependence and leveraging information from the time series, the multivariate DLM leads to improved parameter estimation, evident in the enhanced precision of the one-step-ahead forecast errors.

To conclude, it is essential to acknowledge the limitations present in this work. Firstly, the distribution of the standardized residuals deviates from the Gaussian distribution in the tails. Additionally, the analysis focuses exclusively on PM2.5 particle levels, overlooking the potential influence of other variables such as wind speed and temperature, which could provide valuable insights into the dynamics of air pollution.