Modeling the dynamics of air pollution

Normanno M.A.

June 2023

1. Research Question and Description of the Dataset

A number of adverse health impacts have been associated with exposure to both PM$_{2.5}$ and PM_{10} (i.e., particulate matter of diameter 2.5 and 10 micrometer or less, respectively). Short-term exposures (up to 24-hours duration) have been associated with *premature mortality*, *increased hospital admissions* for heart or lung causes, *acute and chronic bronchitis*, *emergency room visits*, *respiratory symptoms*, and *restricted activity days*. On the other hand, long-term (months to years) exposure to PM$_{2.5}$ has been linked to *premature death*, and *reduced lung* $function$ *growth* in children. The effects of long-term exposure to PM$_{10}$ are less clear, although several studies suggest a link between long-term PM10 exposure and *respiratory mortality*. By monitoring and studying the level of pollution of the air, and by predicting its path, authorities could be able to better protect the population.

Therefore, this project aims to model the path of air pollution in the US West coast over a period that covers summer 2020 (including the 2020 wildfire season) through the data from the U.S. Environmental Protection Agency (EPA) and provide online forecasts using State Space Models.

We focus on PM$_{2.5}$ by considering it s a proxy for the level of *total air pollution*. As above mentioned, the dataset consists of various measurements collected by EPA from 10 stations located along the U.S. West Coast, mostly between San Francisco and Los Angeles, over a period that covers the 2020 summer (from June to September 2020). They are characterized by having no missing values (NA). In particular, the dataset includes:

- `Longitude` and `Latitude`: the spatial coordinates of the EPA station
- `Datetime`: the timestamp (GMT time zone)
- `pm25`: particulate matter of size 2.5 micrograms per cubic meter or less, over the minimum recorded in the data
- `Temp`: air temperature in Celsius
- `Wind`: wind speed in knots/second
- `Station_id`: station identifier within this dataset

It is important to underline that, since a major source of PM$_{2.5}$ is the blowup of fires that may be caused by high temperatures and severely exacerbated by wind, we might expect a certain degree of correlation between them.

2. Analysis of the Time Series of Station 97

As a preliminary step we show the hourly evolution of $\mathrm{PM}_{2.5}$ at *station* 97, which is located in 10556 West Pico Boulevard, Los Angeles, CA 90064.

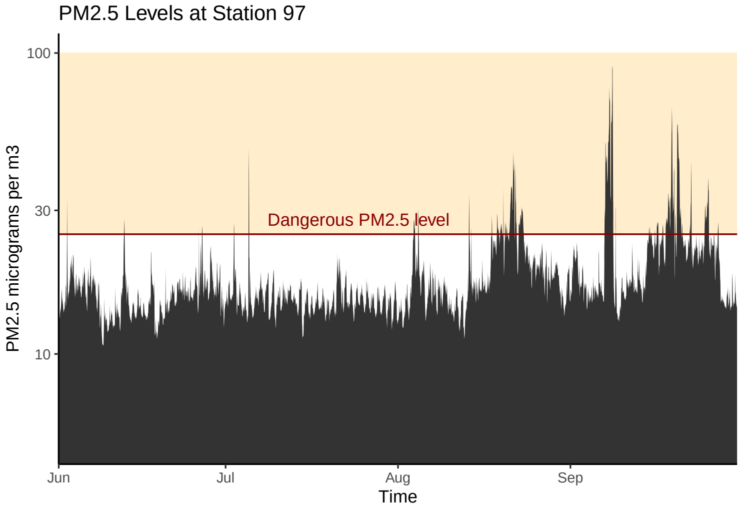

Figure 1: $PM_{2.5}$ concentration levels at station 97, log scale on y axis

Looking at the graph it is possible to easily infer that we have some very short periods of high peaks and longer periods of medium or low concentration of $\mathrm{PM}_{2.5}$, and that there are some structural breaks that lead to new temporal equilibrium in the levels of $\mathrm{PM}_{2.5}$.

Givene these many change points, we proceed by empolying a Hidden Markov Model. As mentioned previously, the rationale is the following: the estimations of pollution levels in the counties can be rather complicated due to the volatility of the observations since they are based on hourly measures prone to measurement errors.

Specifically, we may assume that the time series contains possibly three different levels; then, the process $(S_t)_{t\geq 0}$ is a homogeneous Markov Chain with 3 states representing the *Unhealthy*, the *Moderate Risk*, and a *Safe* path, and Gaussian emission distributions with state-dependent mean and variance. Moreover, conditionally on $(S_t)_{t\geq 0}$, $\mathrm{PM}_{2.5}$ level's are independent and the conditional distribution only depends on the hidden state only through the state-dependent means and standard deviations.

$$Y_t = \mu_n + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma_n^2) \quad \text{if the state } S_t = n, \text{with } n \in \mathcal{S} = (1, 2, 3)$$

Going on, proceeding with the estimation, we firstly observe that the unknown parameters of a continuous Hidden Markov Model are ϕ=(π, A, Θ) where π is the distribution of the first hidden state of the process (S_0), A is the matrix of transition probabilities from state i state j (p_{ij}) and Θ is the set of possible parameters that can determine the the distributions of the observable variable Y given the state j (θ_j).

Table 1: MLEs and associated standard errors

	Low		Medium		High	
	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3
Estimate	14.095	0.983	16.880	1.283	26.310	9.347
Standard Error	0.069	0.034	0.131	0.060	0.471	0.270

We estimate the unknown parameters of a HMM by maximum likelihood. We proceed by fitting the model and computing the MLEs of the unknown parameters $\hat{\phi}$=$(\hat{\pi}, \hat{A}, \hat{\Theta})$. We report the results obtained for $\hat{A}$,$\hat{\Theta}$.

From the estimates, we can see that there are 3 states, corresponding to different average levels of $PM_{2.5}$ in the air. In particular, we can identify a safe level of particles with an average of 14.095, a moderate risk level with an average of 16.880 in which usually sensitive individuals may experience respiratory symptoms, and a unhealthy level with an average of 26.310, above the safety threshold. It is useful to note that this latter level is characterized by a far higher volatility compared to the other two.

Table 2: Transition matrix

	To Low	To Medium	To High
From Low	0.94527	0.00404	0.05069
From Medium	0.05014	0.93144	0.01842
From High	0.00149	0.03595	0.96256

Moreover, we report the time-invariant transition matrix of the homogeneous Hidden Markov model. In particular, we are interested in understanding how long a level is likely to stay above the threshold and how likely it is to go back to safe levels It collects the probabilities of the system moving from one state to another between time (t) and (t + 1) – notice that the data are hours by hours. Generally, states tend to be very persistent. If we consider the state 3, we can easily check that, after 24 hours, the probability of staying still in this state is equal to the transition probabilities from the High state to itself over 24 steps (i.e., approximately 0.92063 = 92.063), so still probable to stay in the Unhealthy state. Moreover, if we want to consider the expected number of hours, we will need to wait before we can go back to the safe level from the unhealthy one, we can consider two possible routes: first, directly with probability 0.15%, or by first passing through to the moderate risk level (3.6%*5%). Overall, the expected average number of hours we need to stay in the unhealthy-moderate risk level before will be back to the safe state is approximately 2 hours.

As a final step, we can then decode the time series by finding the optimal state sequence associated with the the observed sequence of $\text{PM}_{2.5}$ levels.

The following chart represents the estimated path of of the state variable estimated so to maximize the following conditional probability. We add also the 95% confidence interval in order to qualitatively assess the fit of our model.

$$\max_{s_{1:T}} P\left(S_1 = s_1, \dots, S_T = s_T \mid Y_1 = y_1, \dots, Y_T = y_T, \phi\right)$$

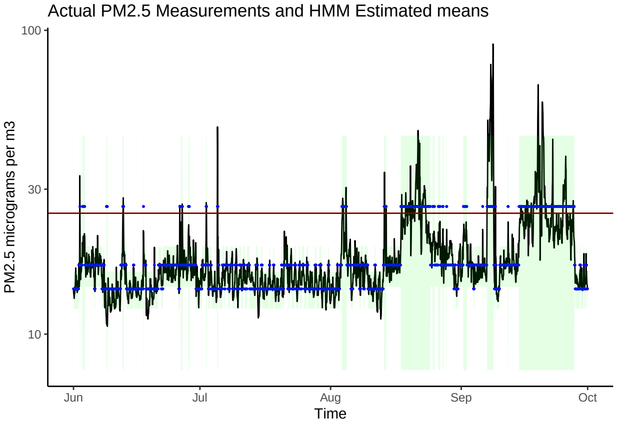

Figure 2: $PM_{2.5}$ concentration, hidden state sequence, standard deviations, log scale on y axis

We can notice that the variance of the three states is positive correlated with the mean, which means that periods in which the model infers a “high” level have a much higher variance in the possible value associated to the state. Moreover, it is also possible to see how the probability of the high state level being above 25, which is the high-risk concentration level, is very low (if computed the probability is around 23%), implying that a high state may not be attached to a high danger pollution level.

3. Forecasts and spatio-temporal analysis

In the previous section, we used an HMM to perform a type of retrospective analysis of the time series to obtain from our observational data the optimal state sequence, corresponding to different levels of air pollution. If instead, we are interested in online estimation and prediction with streaming data we are better off by using a Dynamic Linear Model that allows us to quantify the uncertainty of such prediction through the computation of the one-step-ahead forecasting distribution of the observations. In particular, it might be useful to exploit spatial dependence across observations of air pollution in nearby stations and thus use a multivariate model.

For each of our stations, we consider a random walk plus noise model for our multivariate model. In particular, the random walk plus noise model assumes the presence of a latent state that is distributed as a Markov Chain and, given (θ_t), the observations are assumed to be independent such that they have the following distribution, $p(Y_t|\theta_t)$. Our interest is in the one-step-ahead observation forecasting distribution, $p(Y_{t+1}|y_{1:t})$, which can be computed using the Kalman filter. This distribution will allow us not only to determine point forecast estimates but also fully model the uncertainty behind them.

$$\begin{cases} Y_{j,t} = \theta_{j,t} + v_{j,t} & v_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{v,j}) \\ \theta_{j,t} = \theta_{j,t-1} + w_{j,t}, & w_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{w,j}) \end{cases}$$

It is our goal to analyze more in depth the behavior of the series in 4 different stations (47, 55, 97, and 103), and describe data using a dynamic linear model (DLM) that takes into account potential spatio-temporal relationships between stations. To smooth sharp peaks and reduce the impact of possible outliers, we take the logarithm of the values of $PM_{2.5}$ and average them over 12 hours.

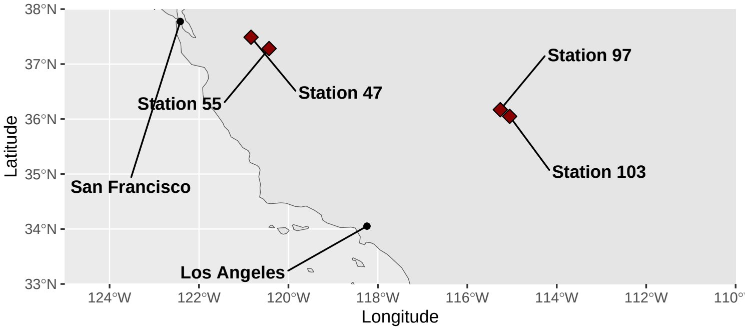

Figure 3: Location of the four stations

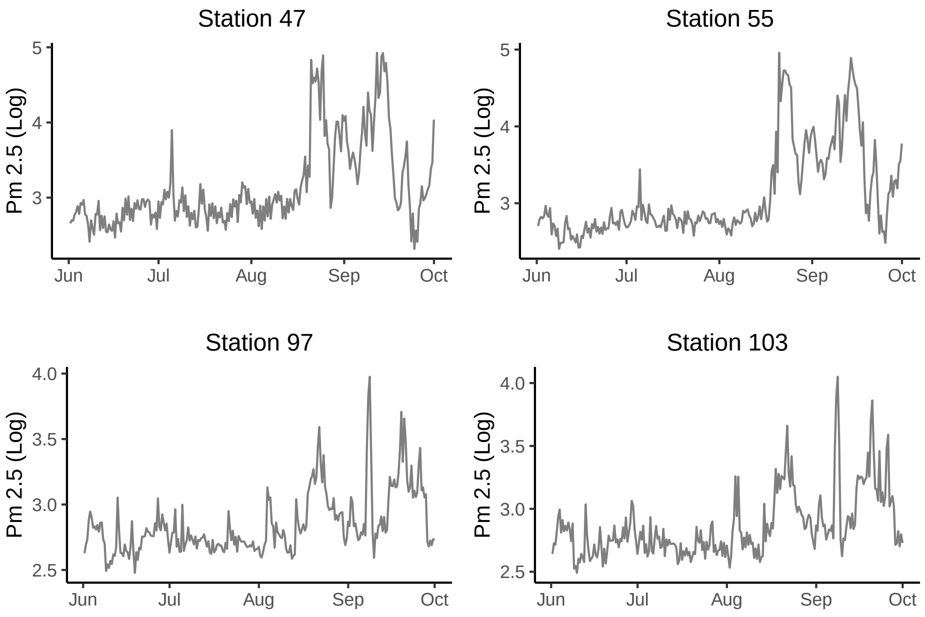

Figure 4: $PM_{2.5}$ evolution across the 4 stations

The above figure is also helpful to see how the behavior of the series associated with these four stations suggests a strong spatial dependence across observations. Particularly, we can see that the graphs of stations 97 and 103 show similar behaviors with peaks in the same periods. Similarly, the graph of station 47 and station 55 show lower volatility until August compared to the other two, even though the peaks are roughly in the same periods. Indeed, if we look at the previous map with the locations of the various stations, it can be seen that stations 47 and 55 are close to each other and far away from stations 97 and 103, and vice versa. We reasonably expect a certain degree of spatial dependence.

3.1 Parameter Estimation and One-Step-Ahead Forecasts

The stations measure the level of $PM_{2.5}$ in the air, so that in a relatively close place the values do not vary sharply. We can thus exploit what is called borrowing strength and use values from different stations to predict the level of particulate in a given location. To capture this longitudinal dependence we decided to use a multivariate local level model that builds on the random walk plus noise model introduced before and allows the error terms of the states to be correlated between different stations. More in details, our model will have six different parameters: four variances of the observation noise of each station, $\sigma^2_{v,i}$, the variance of the error in the state equation (that we assumed equal for all the stations) σ^2_W, and the decay parameter that allows us to introduce the spatial dependence ϕ.

Indeed, the correlations between the error terms of the states are represented in the covariance matrix $W[j,k] = Cov(w_{j,t}, w_{k,t}) = \sigma^2_W e^{-\sigma D[j,k]}$ with $j, k = 1, 2, 3, 4$ and $D[j,k]$ representing the distance between station j and k. Therefore, the closer the two stations are, the higher the correlation of their state errors will be.

In order to estimate the parameters of the model we use the Maximum Likelihood method. All in all, the multivariate local level model (random walk + noise) is specified as follows:

$$\begin{cases} Y_t = F\theta_t + v_t & v_t \overset{indep}{\sim} N_4(\mathbf{0}, V) \\ \theta_t = G\theta_{t-1} + w_t, & w_t \overset{indep}{\sim} N_4(\mathbf{0}, W) \end{cases}$$

In particular, $Y_t = [Y_{t,1}, Y_{t,2}, Y_{t,3}, Y_{t,4}]'$, F and G are two 4×4 identity matrices, $\theta_t = [\theta_{t,1}, \theta_{t,2}, \theta_{t,3}, \theta_{t,4}]'$ characterizes the state equation with the assumption of $\theta_0 \sim N_4(m_0, C_0) \quad (v_t) \quad (w_t)$, m_0 is the vector with the first observation of each station, C_0 is a spherical matrix with diagonal element 10^5, and V is a diagonal matrix with $V[j,j] = \sigma^2 \; with \; j = 1, 2, 3, 4$.

Below we report the estimated parameters with the associated standard errors. The resulting matrices are:

$$V = \begin{bmatrix} 0.02032 & 0 & 0 & 0 \\ 0 & 0.01073 & 0 & 0 \\ 0 & 0 & 0.00029 & 0 \\ 0 & 0 & 0 & 0.00386 \end{bmatrix}; W = \begin{bmatrix} 0.0227 & 0.0206 & 0.00693 & 0.00659 \\ 0.0206 & 0.0227 & 0.0076 & 0.00723 \\ 0.00693 & 0.0076 & 0.0227 & 0.02152 \\ 0.00659 & 0.00723 & 0.02152 & 0.0227 \end{bmatrix}$$

Moreover, we report the correlation matrix, which eases the interpretation of spatial dependence. As we anticipated, stations 47 and 55 are very highly correlated, while stations 97 and 103, which are more distant, comoves to a lesser degree, highly to each other.

Table 3: Estimated parameters

	$\sigma^2_{v,47}$	$\sigma^2_{v,55}$	$\sigma^2_{v,97}$	$\sigma^2_{v,103}$	σ^2_w	ϕ
Estimate	0.02032	0.01073	0.00029	0.00386	0.02270	0.02289
Standard Error	0.00277	0.00188	0.00046	0.00064	0.00193	0.00326

Table 4: Spatial correlations ρ between stations

	Station 47	Station 55	Station 97	Station 103
Station 47	1.000	0.907	0.305	0.290
Station 55	0.907	1.000	0.335	0.319
Station 97	0.305	0.335	1.000	0.948
Station 103	0.290	0.319	0.948	1.000

With the estimated model, we can also compute one-step ahead forecasts with the associated probability intervals. We carried out the analysis for station 97.

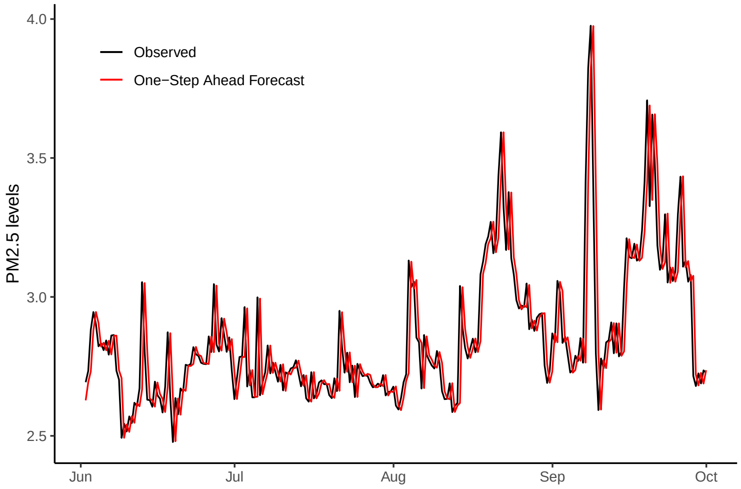

Figure 5: One-step-ahead predictions in station 97

Contrarily to the previous model, we can now make predictions not on states but on actual levels, which gives much more precise results. Furthermore, we can also compare the performance of the model for the 4 stations of interest. As we can see in *Table 5*, if we look at the Mean Squared Error of the forecasts for the 4 stations, we find that station 97 and 103 have very similar MSE, while the MSE for station 47 is significantly higher. This suggests that including neighboring stations significantly increases the accuracy of the model (given that the distance between stations 97 and 103 is lower than the one between stations 47 and 55). This is probably due to the movement of $\mathrm{PM}_{2.5}$ particles from one station to another, or from the propagation of fires. Inserting wind dynamics in the model could be of interest to further assess this relationship. Finally, this also suggests that very probably the higher the number of neighbor stations considered in the model, the better the accuracy of the forecasts. We can deduce it from *Table 6*.

Table 5: MSE for the different stations - Spatial model

Station 47	Station 55	Station 97	Station 103
0.071	0.044	0.02	0.026

Table 6: MSE for the different stations - Non-spatial model

Station 47	Station 55	Station 97	Station 103
0.07	0.351	0.059	0.07

4. Model Checking

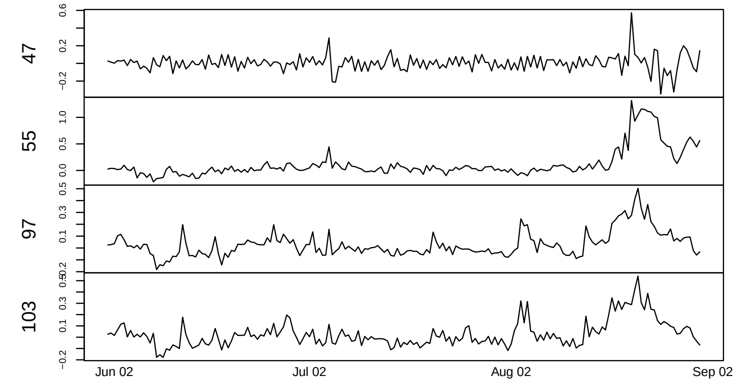

Figure 6: Residuals plot

Figure 6 reports the evolution of standardized residuals over time: while they appear to have zero mean, their variance increases in all the four stations around August, in the middle of the wildfire season. The QQplot in *Figure 7* shows how the distribution of the standardized errors is not exactly normal: their distribution is too peaked at the middle and the tails are too thin. This should not invalidate our forecast accuracy, but may play a role when we compute the credible intervals of our forecasts, which rely on the assumptions of normality in the errors.

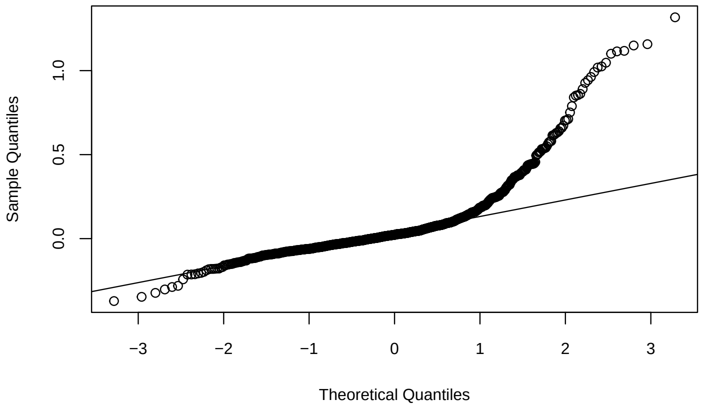

Figure 7: Normal Q-Q plot

5. Final Comments

The decision to use a Dynamic Linear Model (DLM) over a Hidden Markov Model (HHM) enables us to represent spatio-temporal dependence among stations located at different points in the region. In particular, looking at the restrictions imposed on the dynamic model, the assumption of a diagonal matrix to describe V seems quite robust since there are no evident reasons in support of a non-zero correlation in measurement errors among different locations. However, it is necessary to mention that the use of a time-invariant system-error variance (σ^2) is inadequate to account for seasonal trends: we know that between July and August there is the wildfire season in California, and our model is not able to allow for changes in variance of levels and correlations between stations during this period. This is also an issue with the homogeneous HMM that we have used in the first part, being the transition matrix constant over time. We would like to see an extension of our work that encompasses seasonal factors. In addition, in order to apply the DLM it was necessary to restrict the evolution of theta to a specific functional form - a random walk -, a restriction not imposed in the context of an HMM. Unless we have a strong rationale behind this, such as a scientific explanation of the evolution of air pollution, the HMM appears better suited to estimate the latent states. Furthermore, a DLM, as all Bayesian methods, is stochastic in updating the parameter vector, while the HMM is essentially deterministic; this peculiar feature implies that computing the one-step-ahead forecast from a DLM is rather easier and more accurate since the prediction directly gives you the point forecast associated with its variance.