20236 Time Series Analysis: Final Project

Group 7

Stefano Pacifico (3185497)
Matteo Ticli (3077833)

Spring 2022

Analysis of Pollution Level Dynamics in the US West Coast

In the course of this analysis, we are going to try to model the dynamics of air pollution, using hourly air quality data from the U.S. Environmental Protection Agency. The data refers to 10 stations located along the U.S. West Coast and was collected during the summer of 2020. The data provided is collected hourly over the time frame of four months. In particular, we are going to focus on the levels of PM2.5, that is, particulate matter of diameter 2.5 micrometers or less. The study of this kind of particulate matter is of great importance for human health, as high levels of pollution have been linked to respiratory diseases and also to more severe Covid-19 outcomes. We carry out the analysis using different models and approaches, that may be useful in answering several policy-relevant questions.

Descriptive Analysis

The first stage is to carry out a descriptive analysis of the data. To this end, it is useful to plot the data for PM2.5 and to comment on the characteristics of this time series. The plot is reported in Figure 1. For this purpose, we take under consideration station 95; notice that similar considerations would hold for other stations too, since all of them show similar behavior. We will use data from station 95 also in defining the models discussed later on.

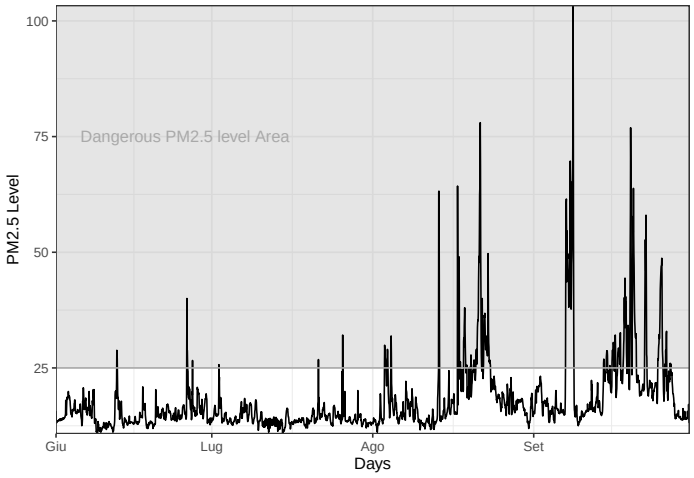

Figure 1: PM2.5 data for station 95

We can see from the plot that there is great variability in the observations. In particular, for the first two months the level of pollution moves around a lower level compared to that observed in the last two months. Moreover, the variance of the level of pollution greatly increases in the second half of the observation period: this is evident from the high peaks shown in the graph. This is likely related to the start of the wildfire season, which has the effect (as it is fair to assume) of increasing air pollution. We will need to keep these considerations in mind in carrying out our analysis.

Here we present the summary statistics of the data:

Table 1: Summary Statistics Station 95

pm25	temp	wind
Min. : 10.81	Min. :12.78	Min. : 0.165
1st Qu.: 13.71	1st Qu.:26.11	1st Qu.: 9.365
Median : 15.31	Median :30.56	Median :12.098
Mean : 17.91	Mean :30.25	Mean :11.217
3rd Qu.: 18.71	3rd Qu.:35.00	3rd Qu.:14.311
Max. :103.31	Max. :42.78	Max. :18.190

Looking in a more rigorous way at the data, we report here the most relevant summary statistics. In particular it is possible to see that the average (hourly) pollution level is 17.91 micrograms per cubic meter, which is below the suggested limit. Moreover, more than 75% of the observations display a PM2.5 level below the 25 micrograms limit. In any case, we also see that there are very large peaks, up to 103.31. It is relevant to notice that the maximum observation among all stations for which we have data is 307.81.

First Model

After this first description of the data, we can move to analyze its dynamics using various models and techniques. In the first section of the report, we analyze station 95, reducing the number of records at our disposal, with respect to the degree of air pollution. We aim at identifying levels of pollution that can opportunely cluster observations relative to the PM2.5 present in the air. We can see in Figure 1 that there is great variability in the observations. In particular, for the first two months the level of pollution moves around a lower level compared to that observed in the last two months. Moreover, the variance of the level of pollution greatly increases in the second half of the observation period: this is evident from the high peaks shown in the graph. This is likely related to the start of the wildfire season, that has the effect (as it is fair to assume) of increasing air pollution.

The first question that we try to answer is: can we define different levels of pollution (and associated instability), in order to guide the needed interventions by the decision-makers? Moreover, can we quantify the probability that a high level of pollution will remain such in the next hour, and the probability that pollution will decrease in the next few hours? For this purpose, we have defined a model that divides the observations into three main states: low pollution, normal pollution and high pollution. In particular, this is an Hidden Markov Model (from now on, HMM). Below is presented the structure of the model:

$$\begin{cases} Y_t = \mu_1 + \epsilon_t, & \epsilon_t \sim \mathcal{N}(0, \sigma_1^2), \quad \text{if} \quad S_t = 1 \\ Y_t = \mu_2 + \epsilon_t, & \epsilon_t \sim \mathcal{N}(0, \sigma_2^2), \quad \text{if} \quad S_t = 2 \\ Y_t = \mu_3 + \epsilon_t, & \epsilon_t \sim \mathcal{N}(0, \sigma_3^2), \quad \text{if} \quad S_t = 3 \end{cases}$$

So, we estimated the relevant parameters of this model using maximum likelihood estimators(μ_t and σ_t^2 for $t = 1, 2, 3$), that is, using estimates that best fit the observed data. In table 2 we report the estimates for the 3 states' parameters. In particular, for each of them we report the expected pollution level together with the associated instability. Notably, we see that only 1 state (state 3) has an expected pollution level above

the suggested pollution limit. We decided to define three states because it allows us to have a parsimonious specification and, at the same time, to describe with satisfactory precision the data observed.

Table 2: Response param Mat of HMM with three state

	mean	sd
S1	16.5057	1.4990
S2	13.4283	0.8364
S3	28.8199	11.6352

In Figure 2, we plot the data together with the estimated states at each observation. As it is fair to expect, we can see that state 3 is associated with periods of high pollution. However, we observe that some datapoints were misclassified: they were incorporated with the observations above the maximum threshold (above 25 micrograms per cubic meter) of PM in the air being nonetheless under the critical level.

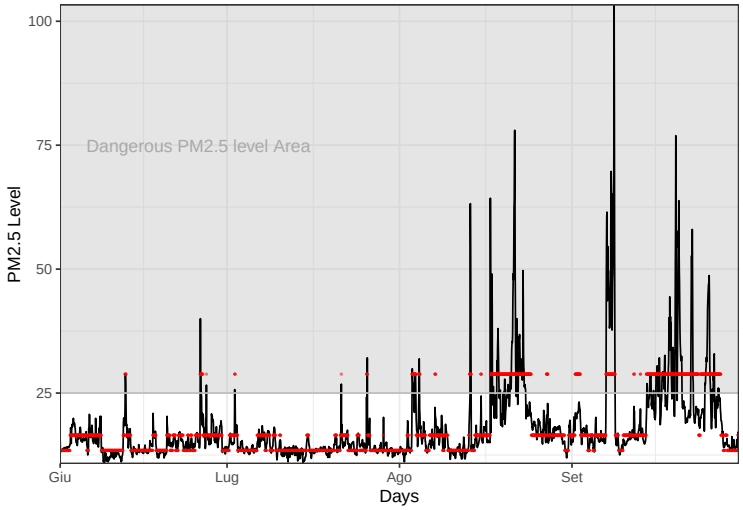

Figure 2: PM2.5 levels at Station 95 estimated with a three states HMM

One major advantage of this model is that it allows us to quantify the probability of switching between different states. We report the estimates of these probabilities in Table 3. In other words, we can quantify the probability of moving from different pollution levels and this indication is important, because it can suggest the expected level of pollution and so can guide the actions of policy-makers. For instance, in the station under analysis, we can conclude that the hourly pollution level is most likely to remain in the same state after one period; this is consistent with the high frequency of the observations, as it is unlikely to have abrupt changes in the level of pollution in the course of one hour. However, we can also conclude that in the time frame of a few hours, the probability of remaining in a high state of pollution drops significantly.

Table 3: Trans Prob Mat of HMM with three states

	toS1	toS2	toS3
fromS1	0.9460	0.0370	0.0170
fromS2	0.0314	0.9669	0.0017
fromS3	0.0302	0.0014	0.9684

The model just presented is useful for the reasons explained above; however, one major drawback is that it

does not allow to make precise one-step-ahead forecasts. In other words, it does not give a precise indication of what pollution level we should expect in the next hour.

Second Model

In order to answer more precisely to this other question, we defined a different model, that links each observation only to a state θ and observations are independent among them given the state parameter. As in the previous model, the state depends only on the one in the previous period and adds to it a random error. More precisely, we defined a Dynamic Linear Model (DLM from now on) using a random walk plus noise model:

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

With the standard assumption of: $\theta_0 \perp\!\!\!\perp (v_t) \perp\!\!\!\perp (w_t)$. We set as initial value for the parameter θ_0 the first observation of the PM2.5 series. We assume that the initial value is independent from the errors terms. Moreover, the error terms for the observation equation and the state equation are independent among them. Most notably, we are assuming that the variances will remain constant throughout the series; this is a strong assumption given the data, nonetheless necessary to define the model.

The main advantage of this model is that it allows to make online estimation and prediction, exploiting the streaming nature of the data available; in fact, given that data are coming in quite frequently (hourly), it makes sense to produce one-step-ahead forecasts based on the previous observation. Below, we report the estimates of the parameters of the model (the variance of the observation error and that of the state error), together with the associated uncertainty (i.e., their variance).

Table 4: MLEs and standard deviation for DLM with hourly data

	MLE	sd
param 1	0.0000	0.0049
param 2	7.5291	0.1968

Where param 1 is σ_v^2 and param 2 is σ_w^2.

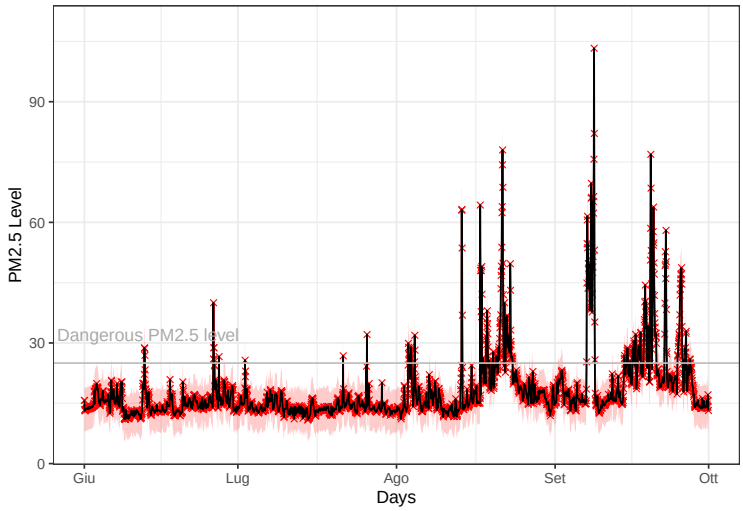

Figure 3: Station 95 hourly data of PM2.5 levels with respect to its one-step-ahead forecast

In Figure 3 we represent the PM2.5 hourly observations (in black) and the one-step-ahead forecast (in red); the shaded area in red is the 95% confidence interval of the forecast. We can see that the performance of our model is somewhat disappointing: in fact, the one-step ahead forecasts obtained using the hourly observation with the DLM stated above mimic the just occurred observation. This problem arises from the impossibility of correctly estimating the variance for the error term (σ_v^2) of the observation equation: since hourly data are very noisy and irregular, the maximization algorithm fails in estimating correctly this parameter; indeed we obtain a value very close to zero, which doesn't seem reasonable looking at the data. Deep diving into a more statistical framework, the signal-to-noise ratio (σ_w^2/σ_v^2) is really high and this leads to the problem aforementioned.

To overcome this problem, we used the same model but with the daily averages of the observations for PM2.5 instead of the hourly data. In fact, we expect daily averages to be more stable and less noisy over time; this should help in obtaining more reliable estimates for the model's parameters. From a policy perspective it is fair to assume that the policymakers would undertake decisions on a daily or weekly basis and not on an hourly one, also because the suggested limit for pollution is expressed as a daily mean. Below, we report the estimates of the parameters of the model (the variance of the observation error and that of the state error), together with the associated uncertainty (i.e., their variance).

Table 5: MLEs and standard deviation for DLM with daily data

	MLE	sd
param 1	15.9418	4.7683
param 2	8.1420	4.4949

Where, once again, param 1 is σ_v^2 and param 2 is σ_w^2.

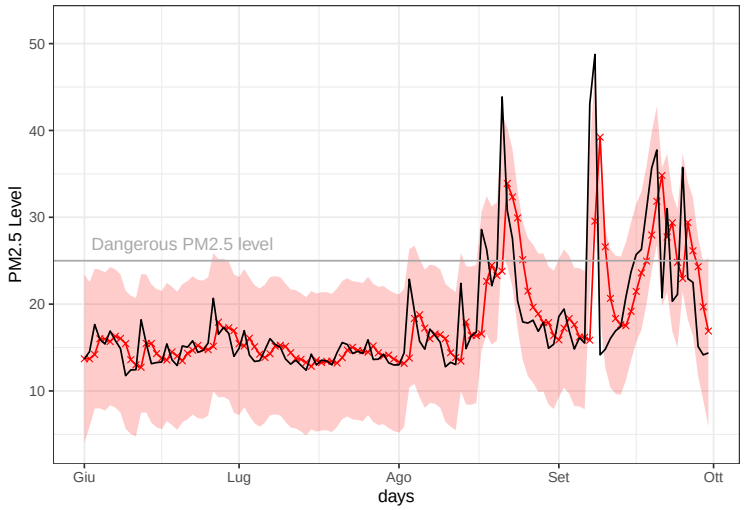

Figure 4: Station 95 daily means data of PM2.5 levels with respect to its one-step-ahead forecast

In Figure 4 we represent the PM2.5 daily means observations (in black) and the one-step-ahead forecast (in red); the shaded area in red is the 95% confidence interval of the forecast. We can evince, first of all that the series is smoother than with hourly observations. As a consequence, the one-step-ahead forecast follow quite well the trend of the series except for periods in which there are high peaks. Still, we have to bear in mind that the assumptions under this model are quite strong, and the results will be valid if the assumptions are respected.

Third Model

After having considered station 95 alone, we decided to consider three stations jointly, in order to take into account the possible correlation among stations located in different places. In fact, pollution is not restricted by any boundary and is likely to move across locations thanks, for instance, to the wind. In other words, the intuition is that, if we observe a high level of PM2.5 in a given station, sooner or later we will observe an increase also in stations nearby.

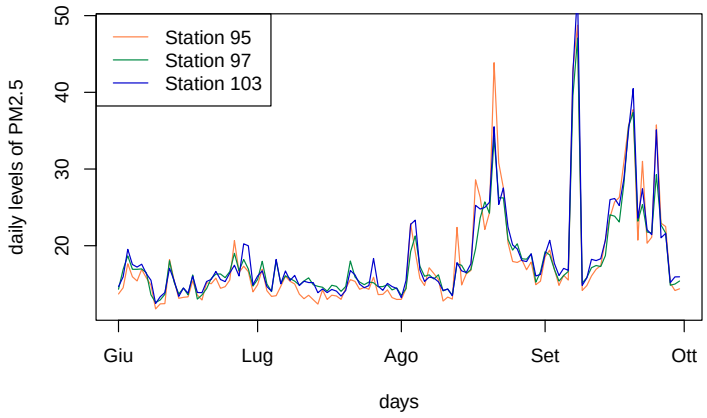

Figure 5: Daily means of PM2.5 levels for Station 95, 97, 103

In Figure 5 we can observe the data relative to the daily average of PM2.5 in station 95, 97, 103. The intuition aforementioned is reinforced by the evidence suggested by the data: the trend of pollutions seems common to all three stations.

For this reason, we adapted the DLM model used before to accommodate not only the temporal, but also the spatial dimension:

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

With the assumptions that:

$$\theta_0 \perp\!\!\!\perp (v_t) \perp\!\!\!\perp (w_t)$$

Now we define the components of the aforementioned model:

$$Y_t = \begin{bmatrix} Y_{t,95} \\ Y_{t,97} \\ Y_{t,103} \end{bmatrix}$$

which is the vector that holds the observations of PM2.5 for the stations considered,

$$\theta_t = \begin{bmatrix} \theta_{t,95} \\ \theta_{t,97} \\ \theta_{t,103} \end{bmatrix}$$

the vector holding the hidden states of the model.

The variance covariance matrix of v_t and w_t are:

$$V = \begin{bmatrix} \sigma^2_{v,95} & 0 & 0 \\ 0 & \sigma^2_{v,97} & 0 \\ 0 & 0 & \sigma^2_{v,103} \end{bmatrix}$$

$$W = \begin{bmatrix} \sigma^2 & Cov(w_{95}, w_{97}) & Cov(w_{95}, w_{103}) \\ Cov(w_{97}, w_{95}) & \sigma^2 & Cov(w_{97}, w_{103}) \\ Cov(w_{103}, w_{95}) & Cov(w_{103}, w_{97}) & \sigma^2 \end{bmatrix}$$

The generic of the W matrix is defined as: $W[i,k] = Cov(w_{j.t}, w_{k,t}) = \sigma^2 \exp(-\phi D[j,k]), \quad j,k = 95, 97, 103$

Moreover, we define the two weighting matrices F and G which are simply defined as:

$$F, G = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The spatial dimension has been inserted into the variance-covariance matrix of the error term of the space equation, where we add a covariance term that depends on the distance between stations. We used this model to analyze the dynamics of stations 95, 97 and 103 jointly. Looking at the assumptions that we made, we assumed that the observation errors are independent among them (i.e., their variance-covariance matrix is diagonal) and that each station has a different observation error. We have not been able to estimate correctly the decay parameter that is associated with the distance, hence we fixed this parameter equal to one, in order not to hinder the precision of the other estimates.

Below, it is possible to see the estimates of the parameters of the model as done before:

Table 6: MLEs and standard deviation for DLM with daily data

	MLE	sd
param 1	15.8815	3.6568
param 2	7.3858	2.5158
param 3	13.3703	3.2634
param 4	8.2154	2.5633

Where param 1 refers to $\sigma^2_{v,95}$, param 2 is $\sigma^2_{v,95}$, param 3 is $\sigma^2_{v,95}$, param 4 refers to σ^2.

From the table 6 we can compare the results obtained using the spatio-temporal model with respect to the one presented as single location temporal model. The result for $\sigma^2_{v,95}$ are almost the same, with a slight decrease in both the expected value and standard deviation of the estimate using the spatio-temporal model.

Finally, we report in Figure 6 the daily means observations of PM2.5 (in black) and the one-step-ahead forecasts (in red) using the spatio-temporal model just defined. We do not report the confidence intervals in this case since we were not able to retrieve the standard errors of the forecasts from the model.

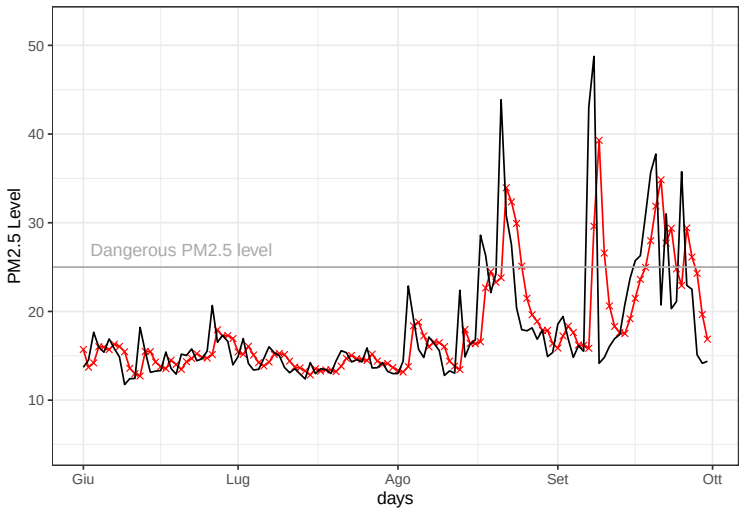

Figure 6: Station 95 daily means of PM2.5 levels with respect to its one-step-ahead forecast (spatio-temporal model)

We can see that results collected in Table 7 are not so different from the one obtained from the simple DLM model used before (no spatio-temporal). To see whether this intuition is sustained by statistical evidence, we computed the MAPE and the MSE for both models and indeed, the spatio-temporal model has lower forecast performance. Nonetheless, the difference is quite small as we can see from the table below:

Table 7: Evaluation metrics of the models

	MAPE %	MSE
DLM 1	7.6583	7.8482
DLM 2	9.4003	8.8658

Where DLM 1 is the simple DLM model while DLM 2 is the spatio-temporal one.

All in all, we can conclude that it is not wise to use the spatio-temporal model since we have a lower precision at the expense of estimating more parameters, incurring in a higher risk of misspecification. It makes sense to stick to a more parsimonious specification such as the one of the simple DLM. In general, the data under analysis represents a complex case study because it displays high variability induced by unpredictable events such as wildfires.