Final Project

Group 22

Imanbayeva Sofya, Mazzi Lapo, Piras Mattia, Srivastava Dev

May 2023

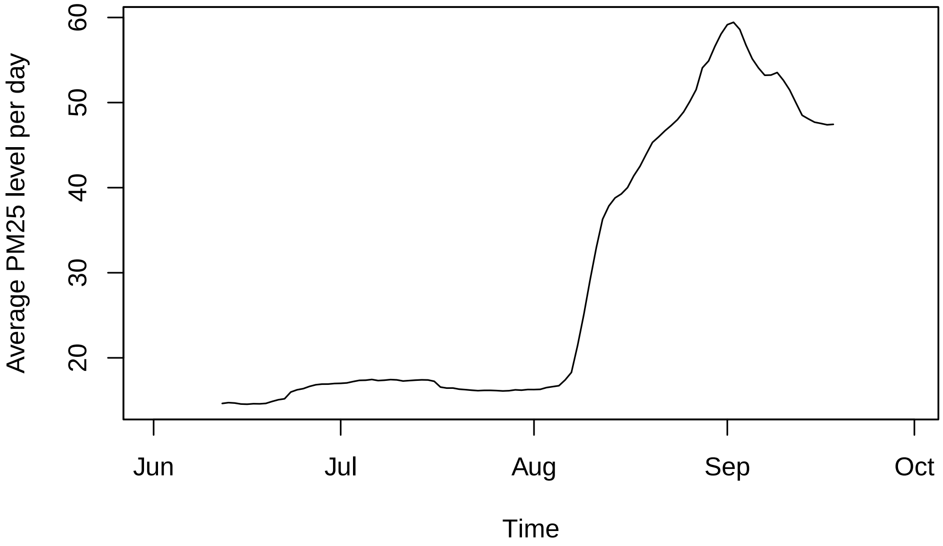
Question 1

We have chosen Station 55 for our analysis.

We will first present some time series plots to understand the data observed. After that we will fit a Gaussian HMM model and use it to interpret the first question of interest.

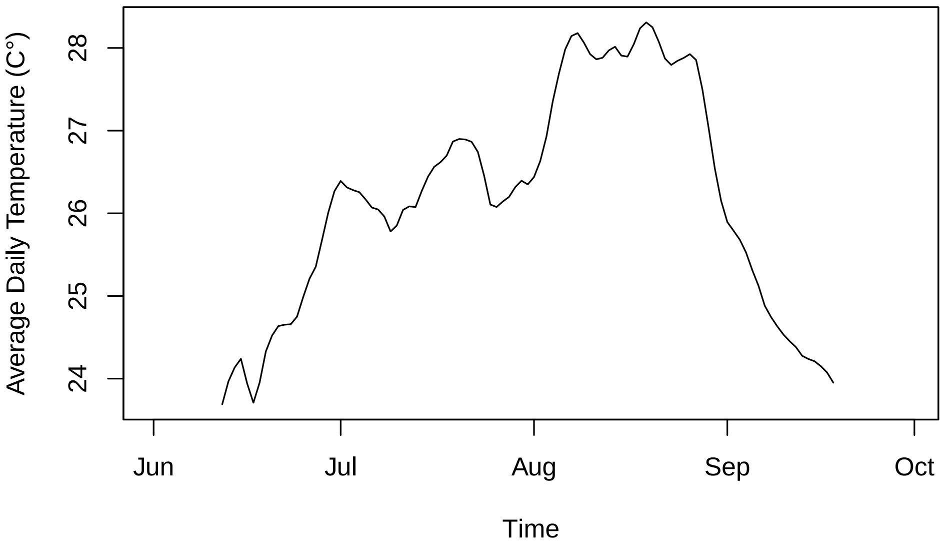
Data Visualization

Graph 1: Average PM25 levels at station 55, 2020



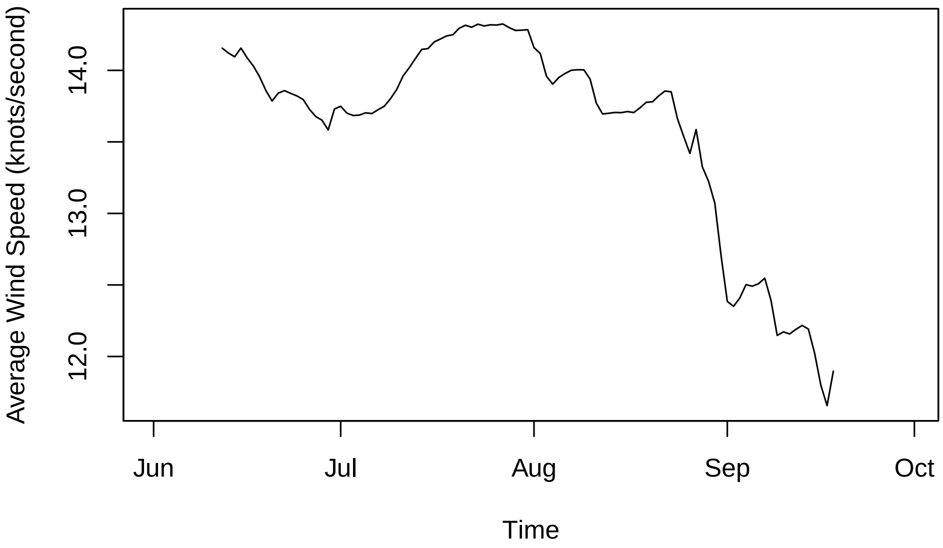
Moving average dramatically increases in the beginning of August from around 17 to its peak in the beginning of September at around 60 in its PM25 levels. The values slightly decrease in September.

Graph 2: Average Daily Temperature (C°) at station 55, 2020



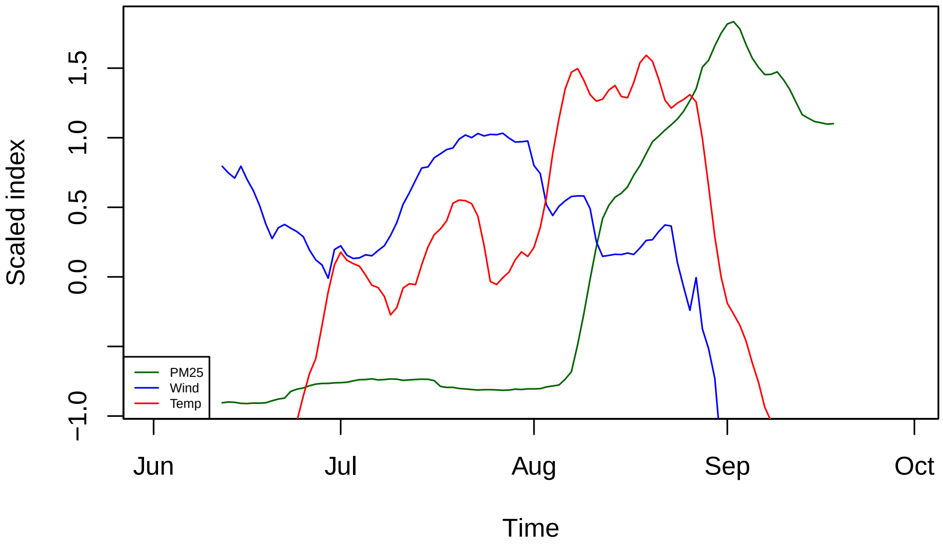
The temperature has been increasing from June's values of 24 degrees Celsius to its peak in August at around 28.

Graph 3: Average Wind Speed (knots/second) at station 55, 2020



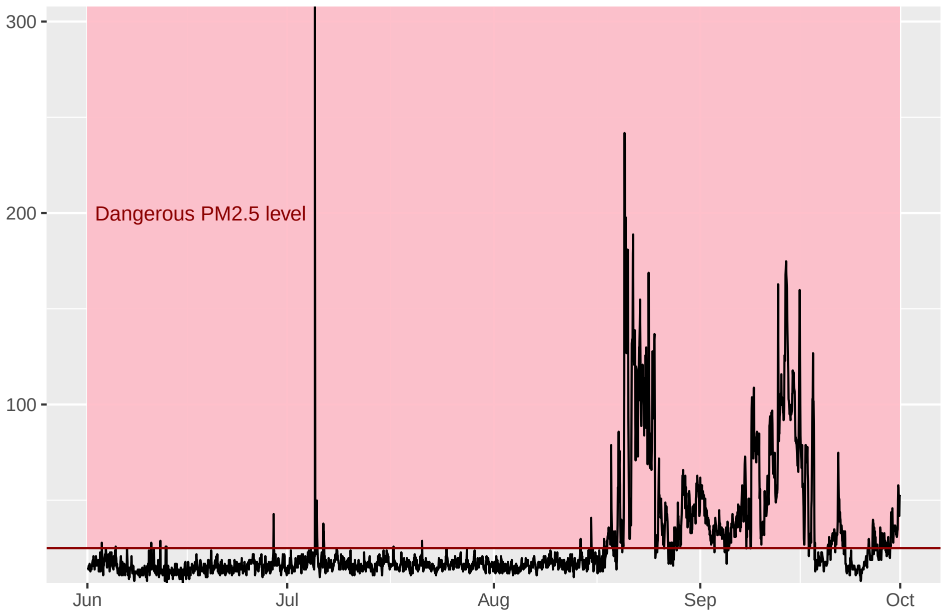
Average wind level has been stable at around 14m/s in the summer and started to decrease in the second part of August to around 12 in mid-September.

Graph 4: Scaled development of PM25, temperature and wind, 2020



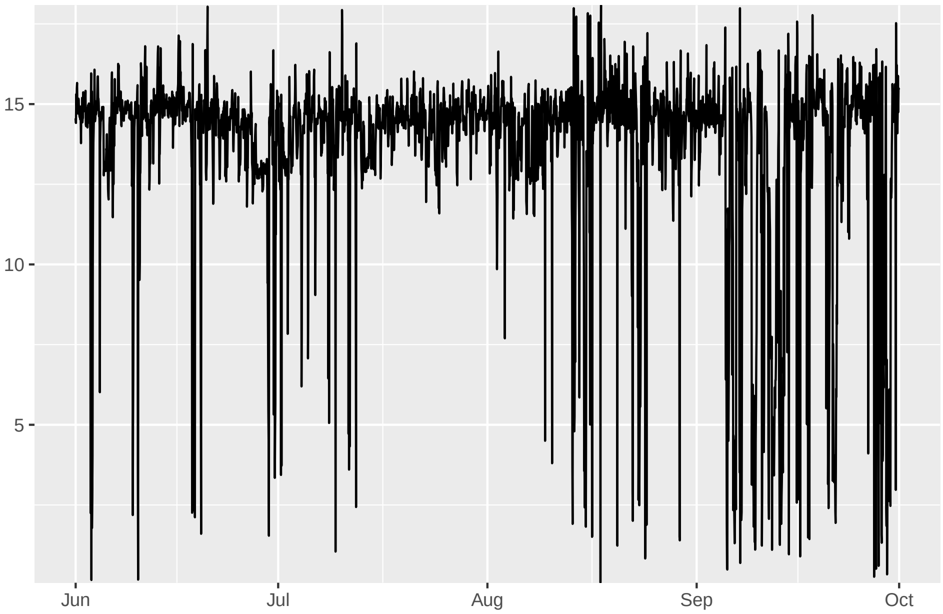
High average temperature in August and comparatively strong winds seem to have a correlation with fires, which have increased the values of PM25 particles in the air. There is around 10 day lag between the temperature increases and PM25 value increase in August.

Graph 5: PM2.5 levels at Station #55



Measurements from June to mid-August are smaller than the prescribed limit with the exception of a peak of an outlying 307.81 in July. However, since August until October, the dynamic has changed with only a few days where the values stayed within the limit constraints below 25 and most data being above the safe limit. The peaks are high, probably resulted from fires, high temperatures and strong wind.

Graph 6: Wind strength



Gaussian HMM Model

We will model the PM2.5 levels in the air using a 3-state Gaussian Hidden Markov model with a simple random walk. The following equations describe it -

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

The initial states and the transitions from state i to j are considered to have equal. Since, there are 3 states, they will have 1/3 probability each. Similarly, the transition probabilities from any state i to j are also initially set to 1/3.

Running the HMM model outlined above on the PM2.5 data gives us the following transition probabilities from state i to state j -

Table 1: Estimated transition probabilities of the states[1]

Transition	To state 1	To state 2	To state 3
From state 1	0.993	0.007	0.000
From state 2	0.023	0.951	0.026
From state 3	0.000	0.043	0.957

Table 2: MLE estimates of the mean and standard deviation of the states[3]

	Coefficient	Standard Error	Upper Bound	Lower Bound
St1 Intercept	15.891	0.070	16.029880	15.752120
State 1 Standard Deviation	3.005	0.052	3.108168	2.901832
St2 Intercept	33.303	0.476	34.247384	32.358616
State 2 Standard Deviation	8.086	0.372	8.824048	7.347952
St3 Intercept	91.324	2.261	95.809824	86.838176
State 3 Standard Deviation	36.054	1.355	38.742320	33.365680

Firstly, we can identify the three states that we wanted to study. State 1 is the one relative to low pollution, state 3 is relative to high pollution levels and state 2 is the one that we can associate to a medium pollution levels. Looking at the transition matrix we note that there are steps that are never possible, state 3 to 1 and vice versa.

Further, the states are very persistent, and probability of state transition by a single step is very low, and by two steps virtually zero. Thus, the current state is a very good predictor of the state in the next hour.

If the current state is of low pollution, there is not much need to enforce any strict measures, with an extremely high probability (0.993) of staying the same low pollution state.

If the current state is of medium pollution, there is a need to enforce strict measures for some time, given the large probability of staying in the medium state (0.95), and also a possibility of transitioning to high pollution state in the next hour(0.026).

Finally, if the pollution is high prolonged strict measures should be anticipated to bring it down medium pollution and finally to low pollution (prolonged because) both high and medium pollution states being highly persistent.

The following table gives the expected number of hours for pollution state to move from i to j (where i and j are different).

[1] The probabilities were estimated through MLE

[2] Upper and lower bound computed at the 95% confidence level

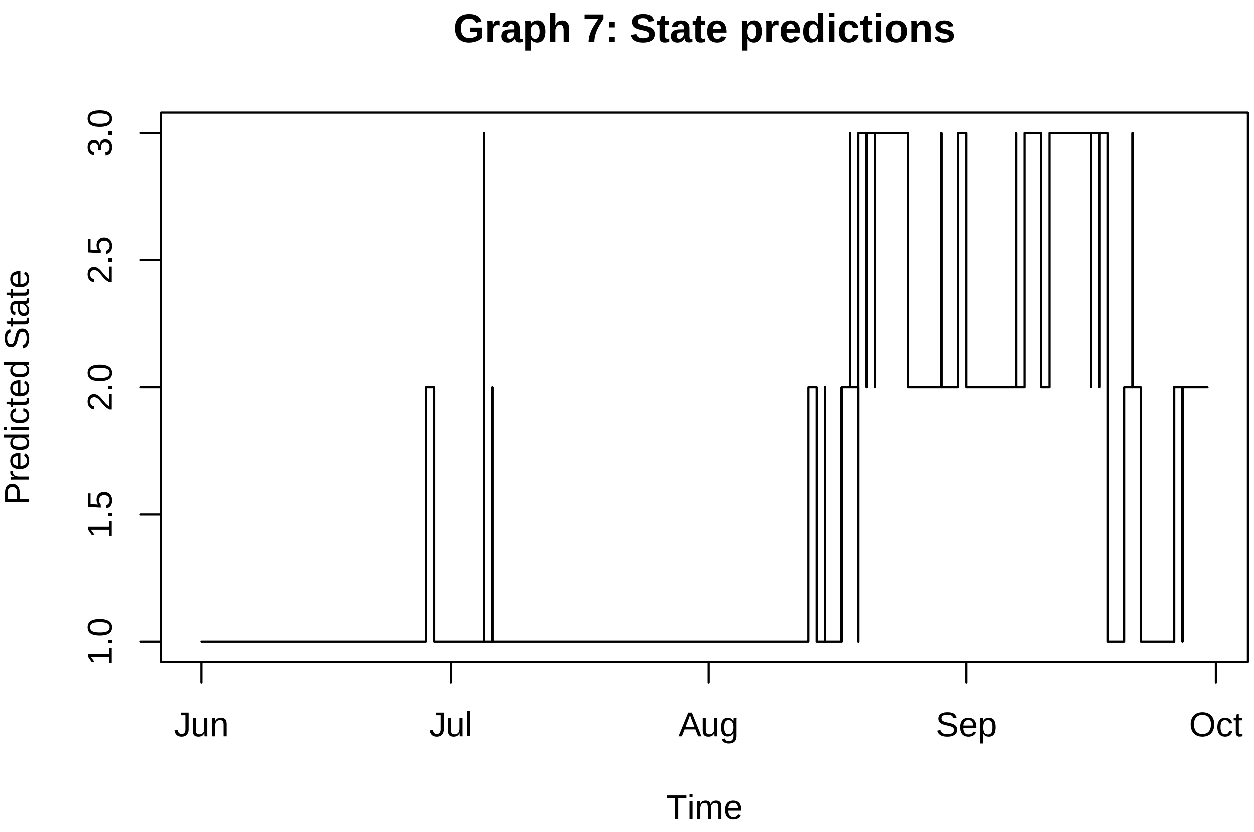
[3] Upper and lower bound computed at the 95% confidence level

Table 3: Expected number of hours for transition from state i to j

Transition	To state 1	To state 2	To state 3
From state 1	0.00	69.19	92.62
From state 2	136.53	0.00	23.43
From state 3	280.90	152.92	0.00

Thus we can see that if the current state is of high pollution, it'll take an expected time of 92 hours for the pollution to reach the low pollution state and 23 hours to reach the medium pollution state. Again given the persistence of each state, the first passage time is a good metric for predicting outcomes over the next hours.

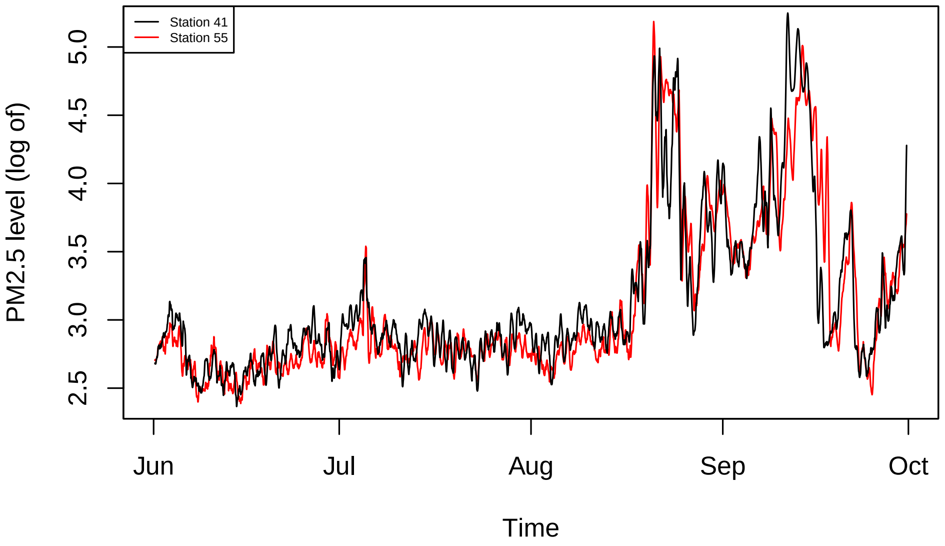
Finally, below is the prediction of the states, given the data observed.



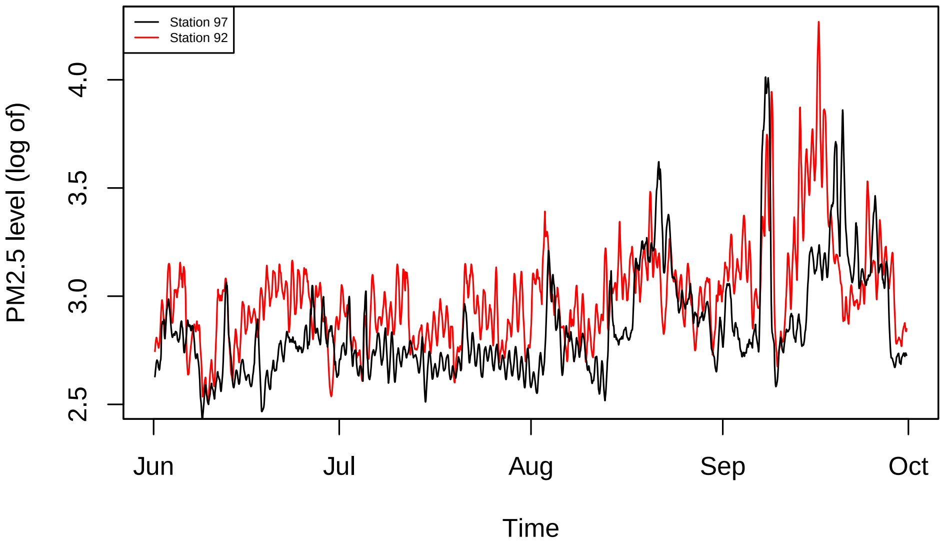
Question 2

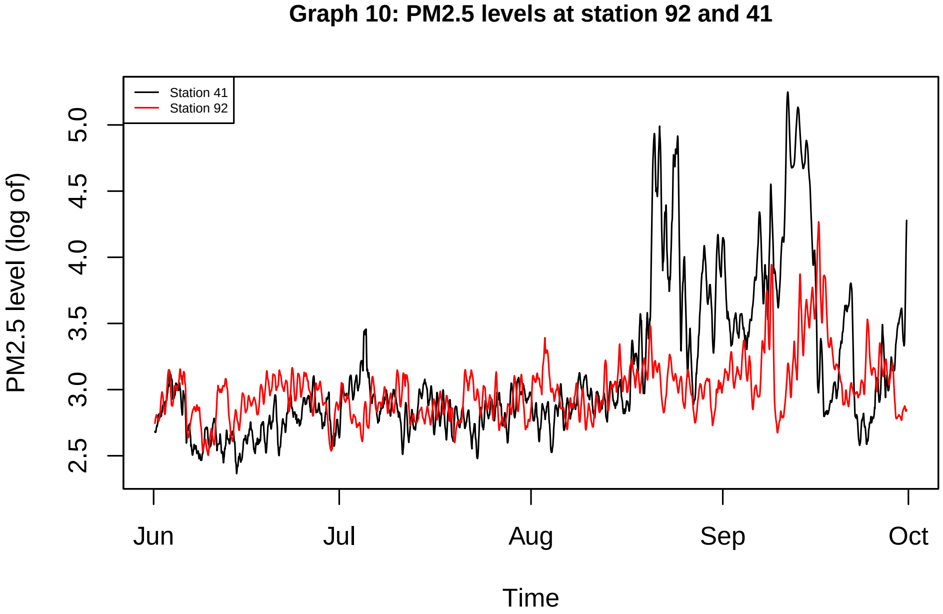
For our analysis we decided to use stations 55, 92, 97 and 41.

Graph 8: PM2.5 levels at station 41 and 55



Graph 9: PM2.5 levels at station 92 and 97





In case of spatial dependence we expect the graphs of the closer stations to be more aligned compared to the ones which are more distant. Stations 55 and 41 are the closest one (only 100km apart) and we can see from Graph 8 that their measurement of pm25 almost coincide, supporting the spatial dependency hypothesis. Stations 92 and 97 are almost 300 km apart, and we can see from Graph 9 that their overlap is lower compared to the previous figure. This is even more clear if we look at Graph 10. Stations 41 and 92 are the furthest apart (650 km) and we can see that their observations are the less synchronized, especially in the Autumn months. Thus, we can affirm from this rough first analysis that there is the case for a phenomenon of spatial dependence.

Model specification

To set up a model that accounts for the spatial dependency we first need a formula to calculate the distance between stations. We decided to use the geometrical based on the coordinates of the stations.

$$distance_{i,j} = \sqrt{(long_i - long_j)^2 + (lat_i - lat_j)^2}$$

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

Clearly, the values for State 2 Expected Value and one-step-ahead prediction are the same, as it seems that for all stations, the value of PM2.5 is expected to be high.