Final Project

Group 22

Imanbayeva Sofya, Mazzi Lapo, Piras Mattia, Srivastava Dev

May 2023

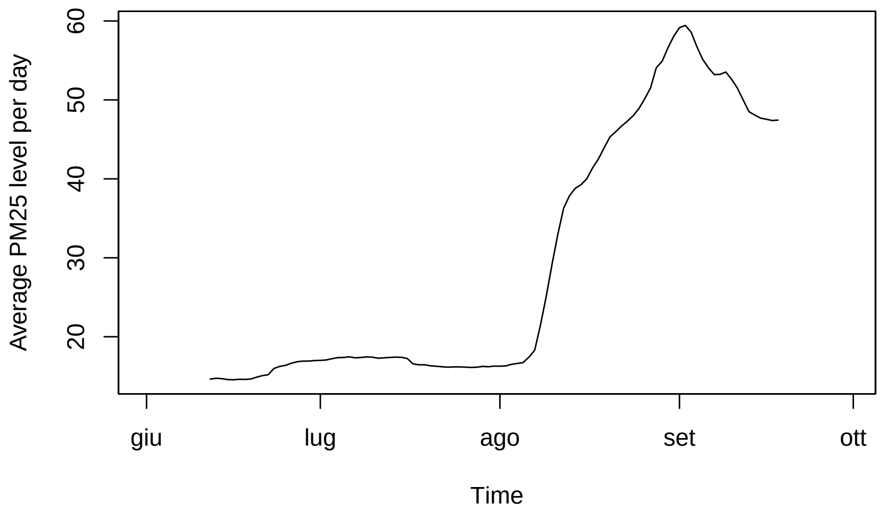
Question 1

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.

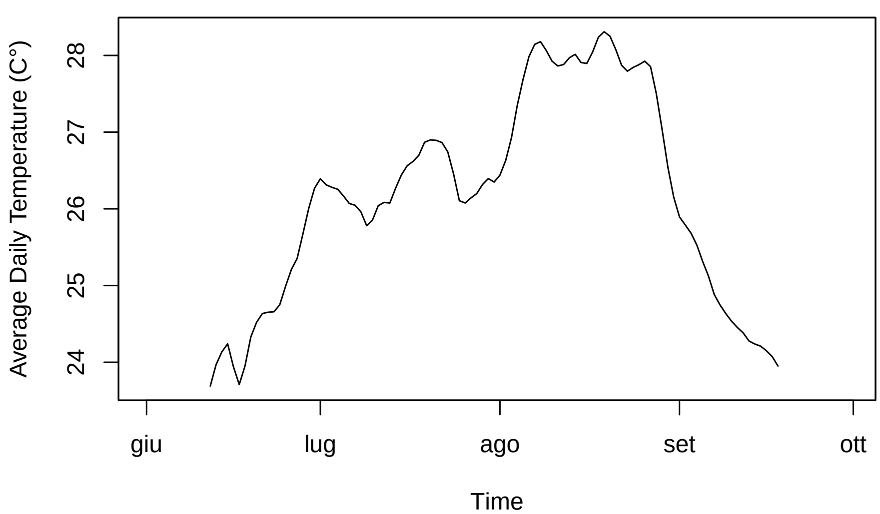
Data Visualization

Graph 1: Average PM25 levels at station 55, 2020

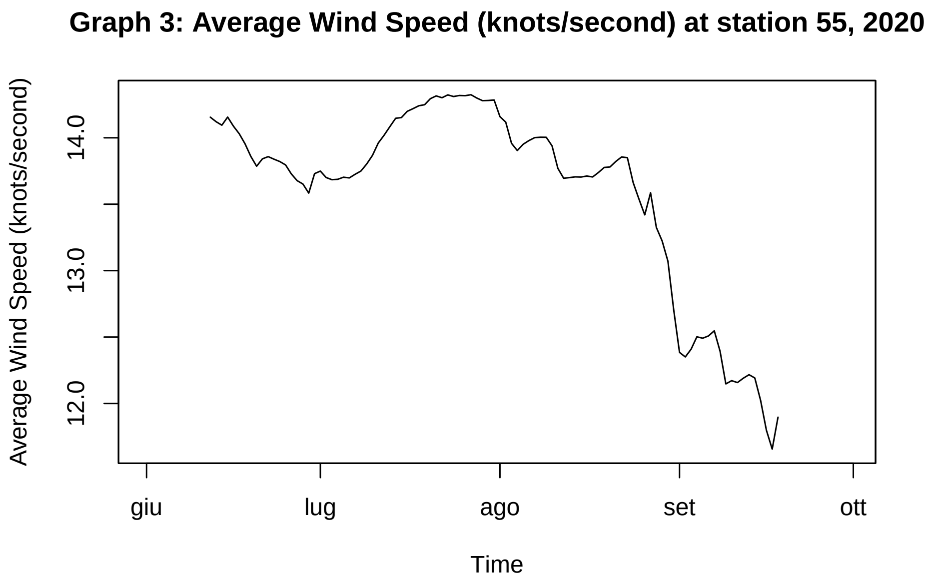


Moving average dramatically increases in the beginning of August from around 17 to its peak in the beginning of September at around 60 in its PM25 levels. The values slightly decrease in September.

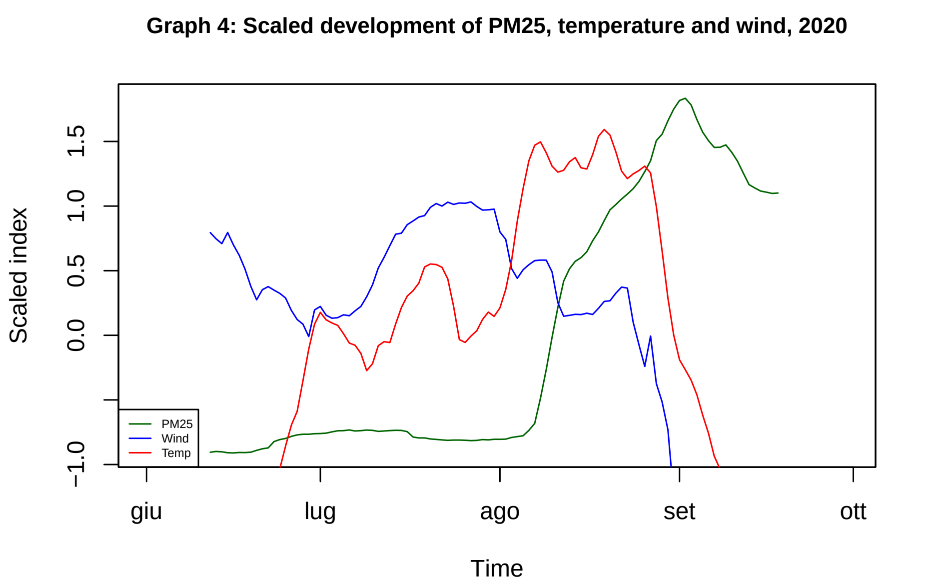
Graph 2: Average Daily Temperature (C°) at station 55, 2020



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

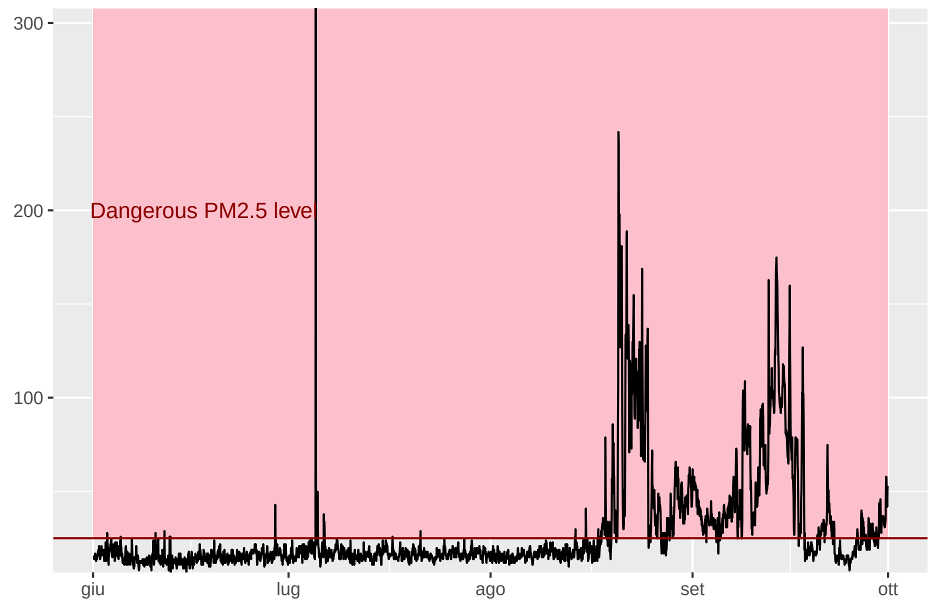


Average wind level has been stable at around 14m/s in the summer and started to decrease in the second part of August to around 12 in mid-September.



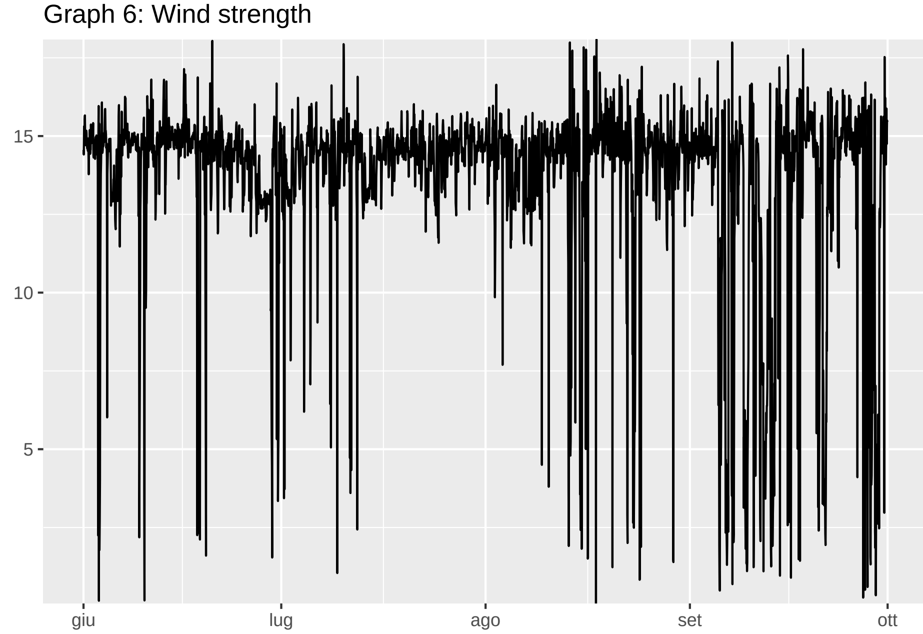
High average temperature in August and comparatively strong winds seem to have a correlation with fires, which have increased the values of PM25 particles in the air. There is around 10 day lag between the temperature increases and PM25 value increase in August.

Graph 5: PM2.5 levels at Station #55



Majority of 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 then until October, the dynamic has changed with only a few days where the values staying within the limit constraints below 25. The peaks are high, probably resulted from fires, high temperatures and strong wind.

Graph 6: Wind strength



Gaussian HMM Model

First, we set up our model. The initial probabilities and the transition matrix are just made by default values. Since, there are 3 states, they will have 1/3 probability each.

```
## Initial state probabilities model
```

```
##   pr1   pr2   pr3
## 0.333 0.333 0.333
##
## Transition matrix
##         toS1  toS2  toS3
## fromS1 0.333 0.333 0.333
## fromS2 0.333 0.333 0.333
## fromS3 0.333 0.333 0.333
##
## Response parameters
## Resp 1 : gaussian
##     Re1.(Intercept) Re1.sd
## St1               0      1
## St2               0      1
## St3               0      1
```

The results of the estimation of both the initial probabilities and the transition matrix are indicated below.

```
fmodel <- fit(model)
```

```
## converged at iteration 23 with logLik: -9089.537
```

```
summary(fmodel)
```

```
## Initial state probabilities model
## pr1 pr2 pr3
##   1   0   0
##
## Transition matrix
##         toS1  toS2  toS3
## fromS1 0.993 0.007 0.000
## fromS2 0.023 0.951 0.026
## fromS3 0.000 0.043 0.957
##
## Response parameters
## Resp 1 : gaussian
##     Re1.(Intercept) Re1.sd
## St1          15.891  3.005
## St2          33.303  8.086
## St3          91.324 36.054
```

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).

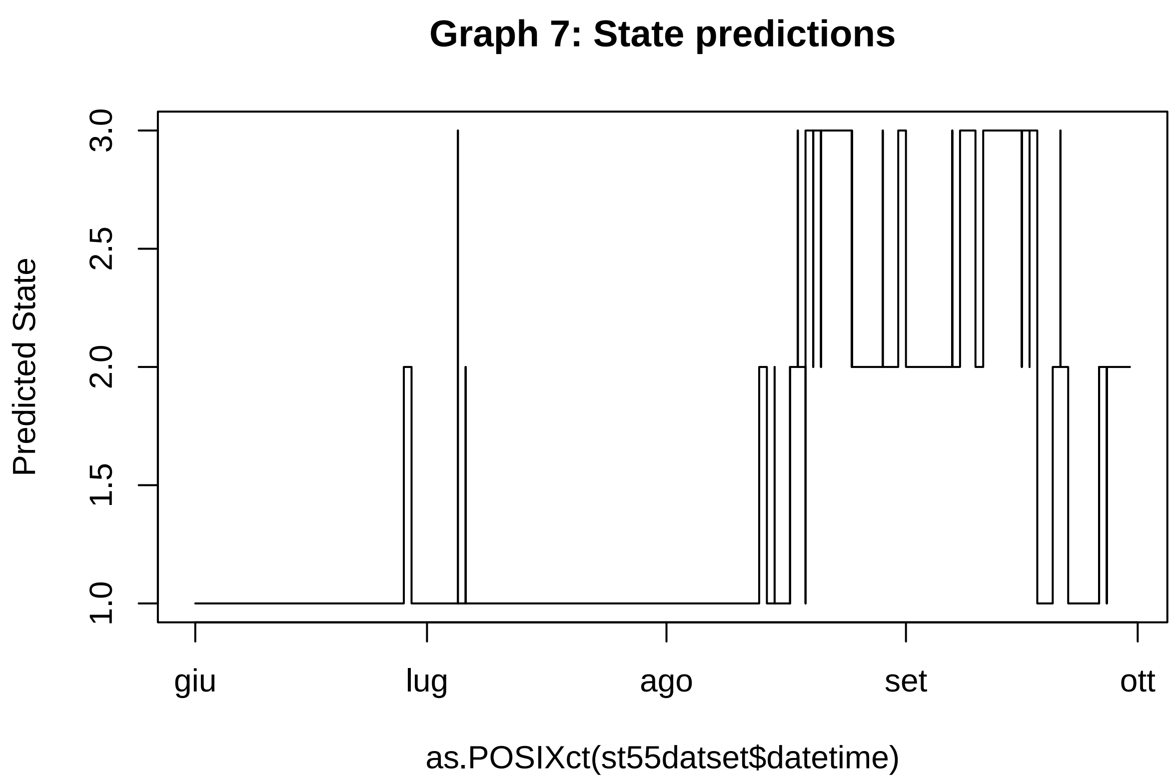
Finally, if the pollution is high prolonged strict measures should be anticipated to bring it down medium pollution and finally to low pollution (prolonged because) both high and medium pollution states being highly persistent.

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

```
##             low    medium     high
## low     0.00000 136.52717 280.8998
## medium 69.18944   0.00000 152.9230
## high   92.61855  23.42911   0.0000
```

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.

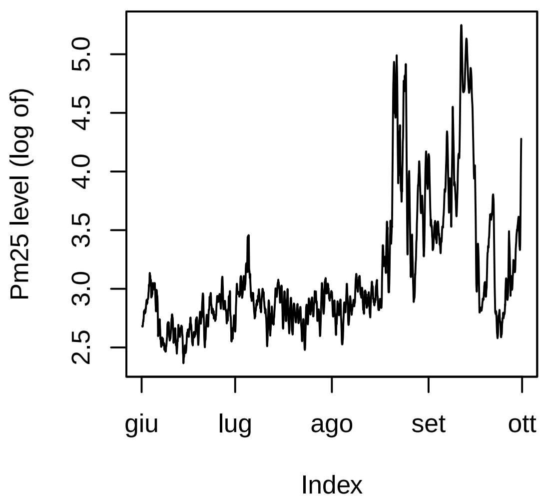
Finally, below is the prediction of the states, given the data observed.



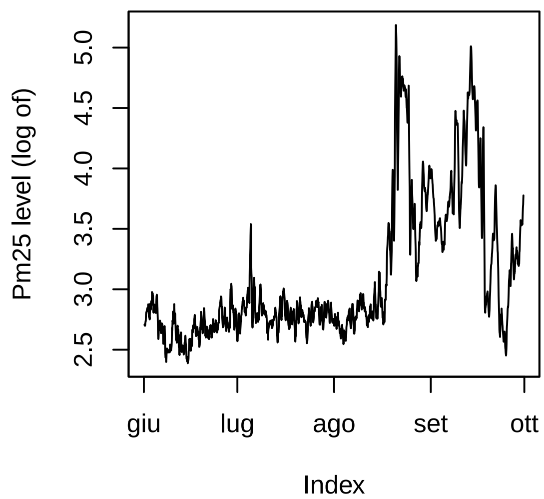
Question 2

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

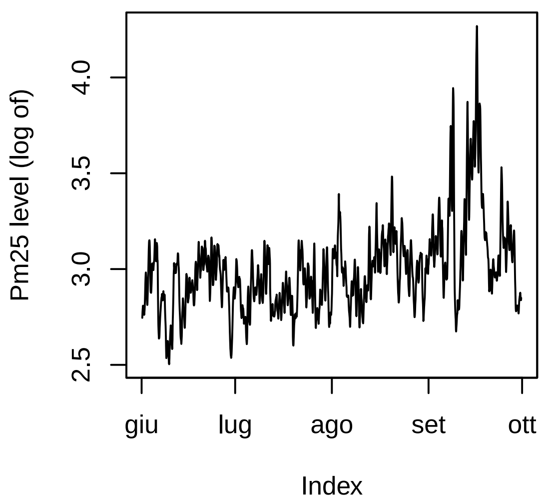
Graph 8: pm25 levels at station 41



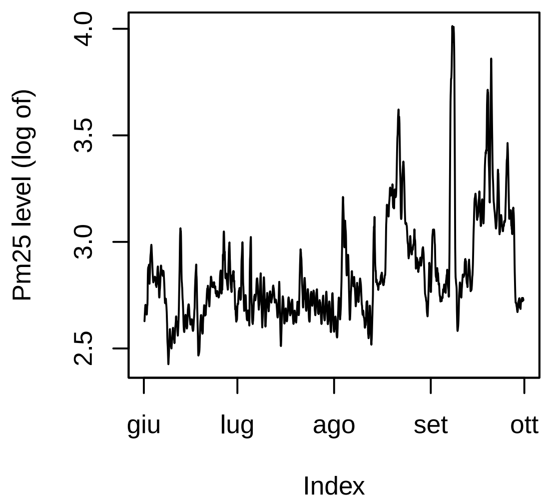
Graph 9: pm25 levels at station 55



Graph 10: pm25 levels at station 92

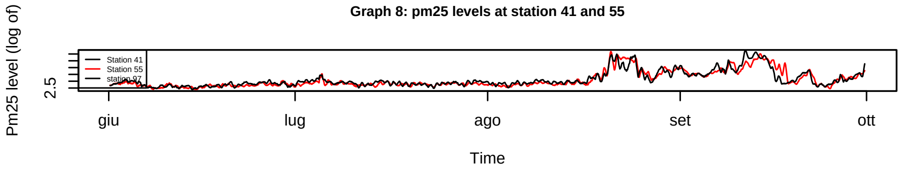


Graph 11: pm25 levels at station 97

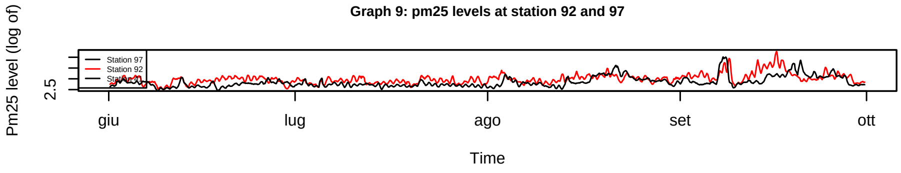


#Model specification

Graph 8: pm25 levels at station 41 and 55



Graph 9: pm25 levels at station 92 and 97



Graph 10: pm25 levels at station 92 and 41