Final_Project_TSA

Georgia Koutsoura - Agostino Ruta - Nicolas Sourisseau

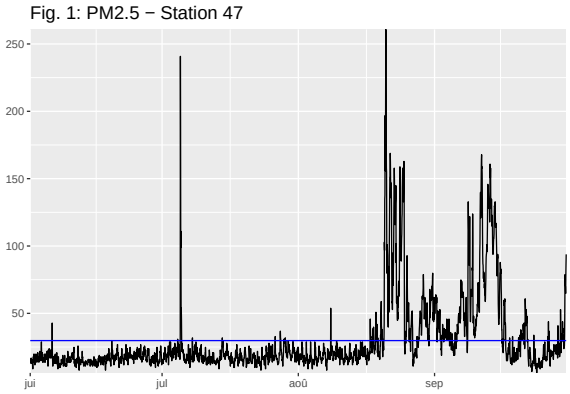
Group 9

Research Question: How to anticipate an increase in the level of PM_{25}? To answer this question, we first give employ a simple model, namely a Hidden Markov Model to have a first idea on the data. Then, we apply more complex structures, including the spatial dependence or other factors such as the wind and the temperatures for example to have a more precise estimate.

For the first part of the analysis of the data, we decided to focus on the station 47. Our data set contains information on the hourly level of PM_{25}, wind, and temperature. In most of our analysis, we are going to take into account only the former to try to predict it since it is our variable of interest. However, it is reasonable, at this stage, to believe that the level of PM_{25} is affected by the wind and the temperature since it is impacted by wildfires.

Description of the Data:

Let us first plot the time series corresponding to the level of PM_{25} measured at the station 47 to get a first idea about the time series and the way we can try to modelize it.



Min	Med	Max	Mean	SD
5.81	18.81	260.81	29.79	28.77

Table 1: Summary Statistics

As we can see from Figure 1, the time series shows no clear trend nor seasonal component. However, it is not stationary (*the average between June and August is clearly below the average between mid-August and mid-September*) and since there is no trend or seasonal component, we can not use a simple transformation to make it stationary. Nevertheless, if we look closely to Figure 1, we can see that the time series admits change points, hence we may want to use a Hidden Markov Model to modelize it.

Hidden Markov Model:

A Hidden Markov Model is a discrete time stochastic process $\big((Y_t, S_t)\big)_t$, where one assumes that the observable time series $(Y_t)_{t\geq 1}$ depends on a latent process $(S_t)_{t\geq 0}$ (*state process*). Therefore, in our case, the stochastic time process is $(PM2.5_t)_{t\geq 1}$.
Our assumptions concerning this model are that:

- $(S_t)_{t\geq 0}$ is an homogenous Markov chain, with state space $\{1, 2, 3\}$, initial distribution π, and transition matrix $P = \big(p_{ij}\big)_{i,j=1}^{3}$.
- Conditionally on $(S_t)_{t\geq 0}$, the $PM2.5_t$'s are independent and $\forall E$, it is that:

$$\mathbb{P}\big\{PM2.5_t \in E | PM2.5_{1:t-1}; s_{0:t}\big\} = \mathbb{P}\big\{PM2.5_t \in E | s_t\big\}.$$

Note that, here, we define a model with three hidden states. However, concerning the number of states for our HMM estimation, from Figure 1, we can see either two or three states. On one hand, we prefer to use a model that is more accurate, this argument would be in favor of the three-state hypothesis. But on the other hand, if we add to many states, then our model is too tailored and will not be useful anymore, this argument would be in favor of the two-state hypothesis.

To have a better idea, let us first try a model with three states. Note that the estimations were made by maximum likelihood. In that case, the obtained transition matrix and response parameters are respectively given by Table 2 and 3:

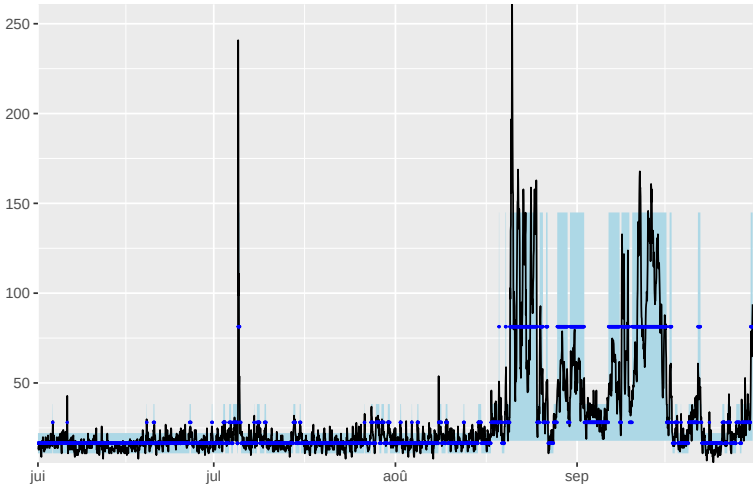
	state 1	state 2	state 3
state 1	0.891	0.026	0.083
state 2	0.030	0.970	0.000
state 3	0.028	0.000	0.972

Table 2: Transition Matrix - 3 States

	Mean	SE	SD	SE
state 1	28.047	0.579	6.213	0.247
state 2	81.363	1.857	38.638	1.231
state 3	16.464	0.143	3.432	0.099

Table 3: Response Parameters - 3 States

Fig. 2: Gaussian HMM for PM2.5 − 3 States



As we can see, two of the states have a very close estimated average, therefore, according to the parsimony principle, we should favor a model with only two states. Nevertheless, note that the 95% intervals of the state means do not intersect, hence, the model can be used. Figure 2 clearly shows that the distribution (*in light blue, the* 90% *prediction interval of the distribution*) in the high-level state is way more dispersed than in the middle- and low-level states. Moreover, the distributions of the middle- and low-level states clearly overlap, therefore we have reasons to believe that introducing a third state may be redundant. Using now the same model as before but with only two states, the obtained transitioned matrix and response parameters are respectively given by Table 4 and 5:

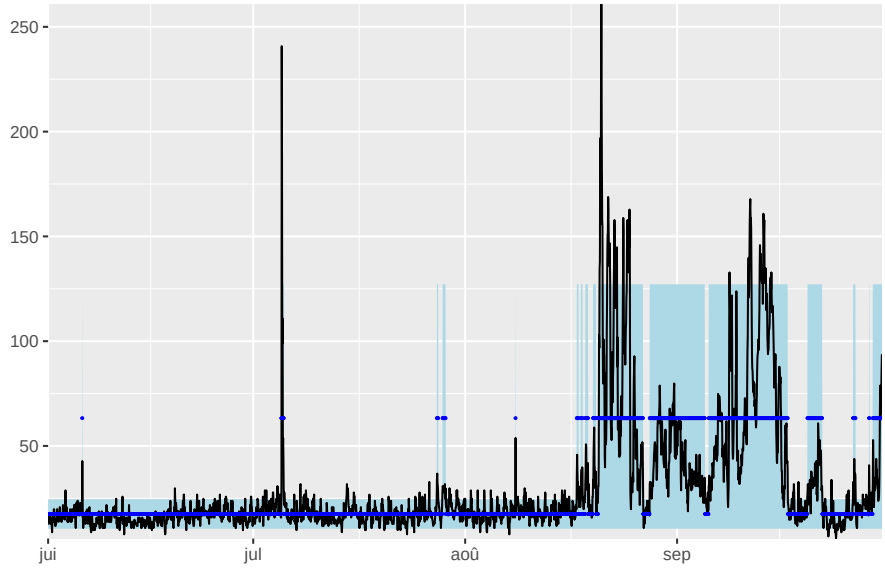
	state 1	state 2
state 1	0.979	0.021
state 2	0.008	0.992

Table 4: Transition Matrix - 2 States

	Mean	SE	SD	SE
state 1	63.279	1.421	38.832	0.981
state 2	17.533	0.096	4.286	0.070

Table 5: Response Parameters - 2 States

Fig. 3: Gaussian HMM for PM2.5 − 2 States



In this Figure 3, we observe that the average amount of PM_{25} in state 1 is way above the threshold while in state 2, it is below. Therefore, we can consider that the two-state modelization is useful because it allows us to make predictions about the fact that the level will be dangerous or not, and, this dichotomy is a very appropriate simplification for our analysis. Note that in both cases (three or two states), the estimated averages of the states are all statistically different from each other (*their* 95% *confidence interval do not cross*). Indeed, for the high state, the 95% confidence interval of the mean is $\left[60.493, 66.063\right]$. Concerning the low state, we have $\left[17.345, 17.721\right]$. Therefore, when we are in the high state, the average level of PM_{25} is really worrying (*clearly above the threshold*) while in the low state the average is clearly below the average. However, in Figure 3, we can also see that the 90% prediction interval for the high-level state distribution is very large, including values under the threshold.
Nevertheless, considering the points mentioned before and the parsimony principle, we will keep the two-state model for our analysis, since it may be the most useful one (*for this specification*).

Analysis:

To identify and estimate the different levels, the first idea, is to use a Hidden Markov Model. We have have estimated 2 of them, one with 2 states and one with 3.
Using this method, we obtained the transition matrix (*Table 4*) which gives us the probability of going from one state to the other. This means that we have the probability of remaining above the danger threshold ($p_{11} = 97.9\%$), in which case some immediate action should be taken and the probability of remaining below the threshold ($p_{22} = 99.2\%$), in which case no action is required. This is consistent with the fact that in Figure 3, several observations accumulate around the same level before switching to another level, hence we generally tend to stay in the same state in the next few hours. We also have the probability of going from the dangerous level to the not dangerous one ($p_{12} = 2.1\%$) and conversely ($p_{21} = 0.8\%$).
In the three-state model, we have the opportunity to set an intermediate state in which some less constraining action should be taken.

Using the two-state model, we can see that the probability that, given that we have a high level of pollution at a certain time, the state will remain the same in the next hour is 97.9%. Therefore, the probability of a significant decrease in the next hour is 2.1%. Concerning a significant decrease in the next few hours, by taking the n^{th}-order of the transition matrix, the probability p_{12} will provide the probability of being in the low level n hours after having been in the high level. However, note that this probability also takes into account the possibility of several state transitions in those n hours.

Spatial Dependence:

The first modelization that we have done seems to be appropriate for the data, however, as explained before, we could include additional factors to have a more accurate forecast. We could consider other variables such as the wind and the temperature, or account for the spatial dependence of the different stations. For now, we are going to focus on the second approach by fitting a random walk plus noise (*dynamic linear model*).

To proceed, we are first going to downsize the data, grouping it by non-overlapping day and night average (*12 hour averages*) for each date. Note that we also need to include other stations to measure the spatial dependence. Therefore, we will incorporate, from now on, the data from stations 55 and 92 in our analysis. Note that we are also taking the logarithm of the variables to reduce the variance of the time series.

Fig. 4: Downsized log−data

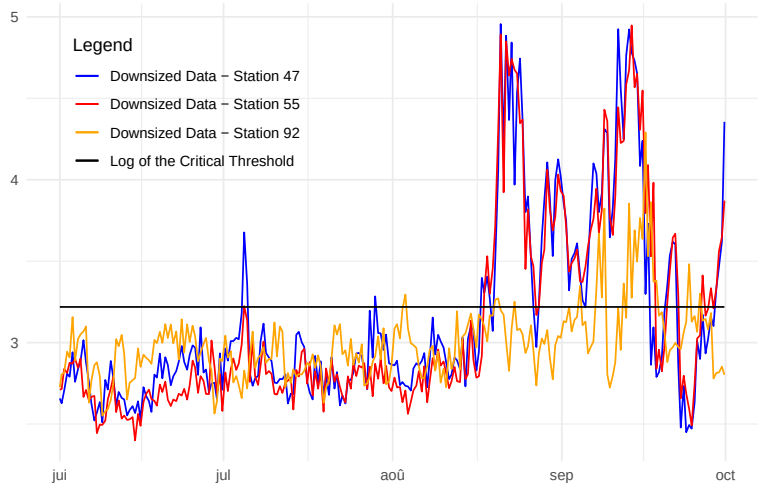


Figure 4 depicts the three downsized time series and the logarithm of the critical threshold. We notice a high correlation between stations 47 and 55 (0.952), which are close to each other. On the other hand, the correlation between each of these and station 92 is lower (*0.392 and 0.472 respectively*) and this station is in a further location. Indeed, it makes sense that the levels of PM_{25} of close locations are correlated. Therefore,

this would tend to confirm the fact that including the spatial dimension in our model would be relevant.

To modelize this dependence, we are using a multivariate dynamic linear model, in which the three time series are dependent only through the covariance of state-errors (*and not through the fact that each state can directly affect the other stations*). Therefore, both F and G are diagonal matrices. Since we want to modelize random walks plus noise, we have:

$$F = G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Concerning the rest of the parameters of our model, after trying the optimization the first time, we observed that the variances of $v_{j,t}$ were very close (*or equal*), therefore, we can add to our model, the assumption that:

$$\sigma^2_{v_i} \equiv \sigma^2_v, \forall i \in \{47, 55, 92\}.$$

Therefore, we have:

$$V = \begin{pmatrix} \sigma^2_v & 0 & 0 \\ 0 & \sigma^2_v & 0 \\ 0 & 0 & \sigma^2_v \end{pmatrix}$$

Contrary to basic models, our W matrix is not diagonal to allow for the spatial dependence (D *represents the distance between the stations*). It is given by:

$$W = \begin{pmatrix} \sigma^2 & \sigma^2 e^{-\phi D(47,55)} & \sigma^2 e^{-\phi D(47,92)} \\ \sigma^2 e^{-\phi D(47,55)} & \sigma^2 & \sigma^2 e^{-\phi D(55,92)} \\ \sigma^2 e^{-\phi D(47,92)} & \sigma^2 e^{-\phi D(55,92)} & \sigma^2 \end{pmatrix}.$$

Therefore, the model that we want to estimate is given by:

$$\begin{cases} Y_t = F\theta_t + v_t \\ \theta_t = G\theta_{t-1} + w_t \end{cases} \Rightarrow \begin{cases} \begin{pmatrix} Y_{47,t} \\ Y_{55,t} \\ Y_{92,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_{47,t} \\ \theta_{55,t} \\ \theta_{92,t} \end{pmatrix} + \begin{pmatrix} v_{47,t} \\ v_{55,t} \\ v_{92,t} \end{pmatrix} \\ \begin{pmatrix} \theta_{47,t} \\ \theta_{55,t} \\ \theta_{92,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_{47,t-1} \\ \theta_{55,t-1} \\ \theta_{92,t-1} \end{pmatrix} + \begin{pmatrix} w_{47,t} \\ w_{55,t} \\ w_{92,t} \end{pmatrix} \end{cases} \Rightarrow \begin{cases} \begin{pmatrix} Y_{47,t} \\ Y_{55,t} \\ Y_{92,t} \end{pmatrix} = \begin{pmatrix} \theta_{47,t} \\ \theta_{55,t} \\ \theta_{92,t} \end{pmatrix} + \begin{pmatrix} v_{47,t} \\ v_{55,t} \\ v_{92,t} \end{pmatrix} \\ \begin{pmatrix} \theta_{47,t} \\ \theta_{55,t} \\ \theta_{92,t} \end{pmatrix} = \begin{pmatrix} \theta_{47,t-1} \\ \theta_{55,t-1} \\ \theta_{92,t-1} \end{pmatrix} + \begin{pmatrix} w_{47,t} \\ w_{55,t} \\ w_{92,t} \end{pmatrix} \end{cases}$$

Where:

$$v_t \overset{indep}{\sim} \mathcal{N}_3\left(\mathbf{0}, \sigma^2_v I\right)$$

$$w_t \overset{indep}{\sim} \mathcal{N}_3\left(\mathbf{0}, \begin{pmatrix} \sigma^2 & \sigma^2 e^{-\phi D(47,55)} & \sigma^2 e^{-\phi D(47,92)} \\ \sigma^2 e^{-\phi D(47,55)} & \sigma^2 & \sigma^2 e^{-\phi D(55,92)} \\ \sigma^2 e^{-\phi D(47,92)} & \sigma^2 e^{-\phi D(55,92)} & \sigma^2 \end{pmatrix}\right)$$

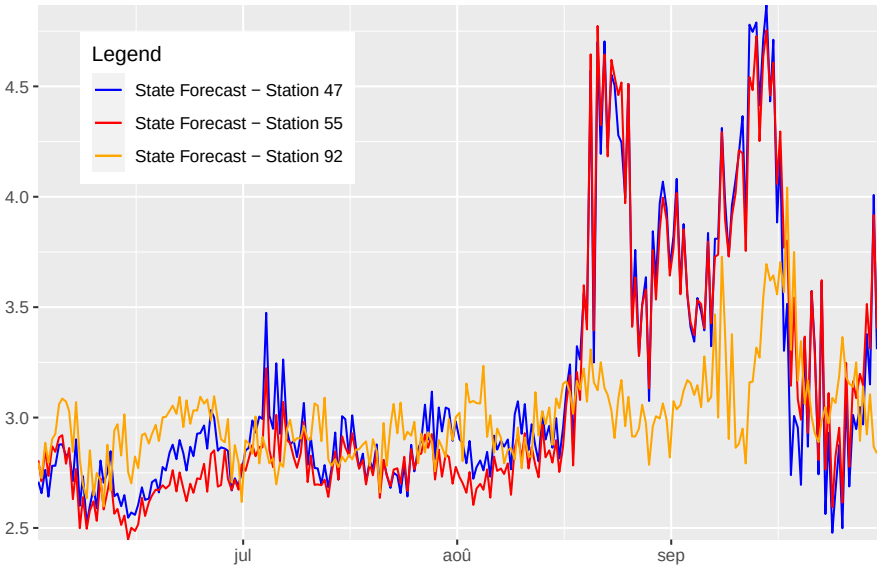
When estimating the parameters of the model, we get the following maximum likelihood estimates (*note that the three parameters are imposed to be positive*):

	Sigma_v2	Sigma2	Phi
MLE	0.015	0.027	0.181
SE	0.002	0.004	0.039

Table 6: MLE

Hence, taking 95% confidence interval, we have $\sigma^2_v \in \left[0.011, 0.019\right]$, $\sigma^2 \in \left[0.019, 0.0.031\right]$, and $\phi \in \left[0.105, 0.095\right]$. Now, using these estimates, we can use Kalman Filter to compute the one-step ahead forecast for the states and for the observations, as shown by Table 6.

Fig. 5: 1−step ahead state forecasts



As we can see in figure 5, the one-step ahead state forecast of stations 47 and 55 are very similar, which confirm our expectations. On the other hand, in station 92, the level of $PM_{2.5}$ tends to vary less, starting at a higher level but never reaching a level as high as the level reached by the other stations.
From the previous point, we can see that the level of the $PM_{2.5}$ in the station 92 has a lower variance than the one of the other stations. However, the formula that we are using considers the variances to be equal, hence, it might be too restrictive.

Fig. 6: 1−step ahead obs. forecasts − Station 47

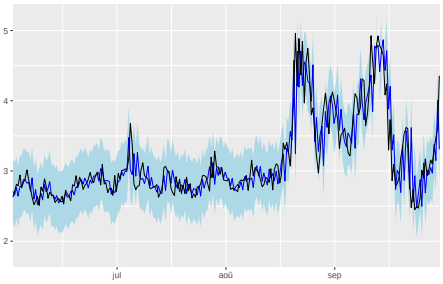
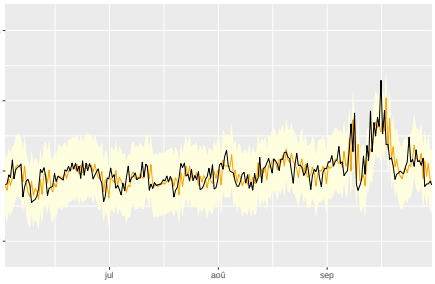


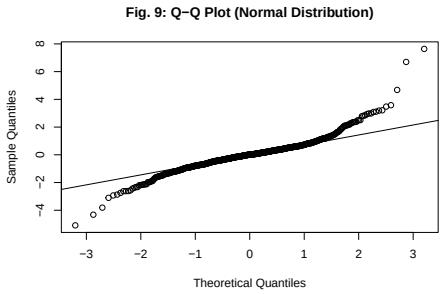
Fig. 7: 1−step ahead obs. forecasts − Station 55

Fig. 8: 1−step ahead obs. forecasts − Station 92



Figures 6-8 show the measured data for each station (*in black*) and the one-step ahead forecast of the observations (*in color*). Note that the light-colored regions represent the 95% confidence intervals. In all three graphs the forecasts seem to be accurate, therefore, the model provides a good estimate of the data, in which case it would be a good solution for our research question. To confirm it and quantify the precision of our forecast, let us look at the errors. Note that we are also doing a comparison with a model with no spatial dependence but allowing the state-errors variances to differ as it seems to be the case in the graphs. In this case, the model is the same as before, except for:

$$W = \begin{pmatrix} \sigma^2_{w_1} & 0 & 0 \\ 0 & \sigma^2_{w_2} & 0 \\ 0 & 0 & \sigma^2_{w_3} \end{pmatrix}$$



	Errors_Mean	Errors_SD	Errors_t-stat	MAPE	MSE
Spatial Model	0.015	1.070	0.014	5.696	8.168
Not Spatial Model	0.013	0.935	0.014	7.378	15.106

Table 7: Errors

As we can see from figure 12, we can conclude that the distribution of the forecast errors in the spatial model is approximately normal, and since $t - stat < 1.96$, we can reject the fact that the errors are significantly different from zero for both models. Note that the mean absolute percentage error is of 5.696% (*and mean squared error of* 8.168%) for the model with spatial dependence and 7.378% (*and mean squared error of* 15.106%) for the model without, we can thus conclude that the spatial-dependence model is more accurate.

Analysis:

Using dynamic linear models, which are based on Bayesian statistics, it is possible to provide an online prediction for streaming data, as we have just did in the previous section. Indeed, these models allow us to have computational efficient online predictions through recursive formulae, therefore, it also allows us to quantify the uncertainty of the estimates. Since our data is very noisy, a good idea to reduce the noise is to downsize it. Moreover, the fact that it is linked to wildfires tends to favor a day/night grouping of the data since fires are more likely to occur during the day. This imply that the variance of our data set is significantly reduced and thus it is easier to estimate the model.
In the previous part, we have introduced a spatial dimension to our model. As we have seen from Table 7, this increases the accuracy of our estimations and predictions. Therefore, we have a promising hint for the estimation of an accurate model to predict the evolution of the level of PM_{25}. Indeed, this spatial dependence could be included in a model in which the precedent level of PM_{25} impacts the future level of each station (*the impact would decrease with the distance*). In that case, the model would be slightly different from the first one that we estimated, indeed, G would not be the identity matrix anymore but would be of the form:

$$G = \begin{pmatrix} 1 & \psi D(47,55) & \psi D(47,92) \\ \psi D(47,55) & 1 & \psi D(55,92) \\ \psi D(47,92) & \psi D(55,92) & 1 \end{pmatrix}$$

Other Factors:

Until now we have only used the level of PM_{25} to predict the level of PM_{25}. However, it is important to notice that this particle is linked to the presence of wildfires. Therefore, it should be impacted by the wind and the temperature that are aggravating factors.
Thus, an idea would be to add some lags of these variables to predict in advance a potential increase in the level of PM_{25} or to use a dynamic linear model with temperatures and wind as states. We can also combine them. This would mean to fit the trend (*global warming*), the seasonality (*seasonality*), and to use an ARIMA model for the detrended time series (*temperature and wind*). Obviously, if needed, we would need to use some other tools to make the temperature and the wind stationary. After these estimates are done, we can combine them with a dynamic linear model, such as the one computed before that includes the spatial dependence.

However, to compute such a model, we would need data from several years in order to fit the trend and the seasonality. Also, this model may require to estimate to many parameters which can bring more complications than solutions for our research question. A first-step would be to compare the previous models with a simple model including the wind and temperatures to check whether it is comparable in terms of prediction power.

Conclusion:

In this project the two main categories of model that we are using are Hidden Markov Models and Dynamic Linear Models. In general, no statistical model is perfect, therefore,it is necessary to compare the assumptions needed by the different models as well as the fit of each model to the data.
The HMM relies on the conditions required in the definition stated in page 2 and additional assumptions of normality and independence between the stations.
On the other hand, the DLM assumes that the initial distribution of the hidden state is independent from the distribution of the errors (v_t) and (w_t). We also assume that the relations are linear and that the distributions are Gaussians. Furthermore, we assume some additional conditions such as the fact that we have a random walk plus noise or that the states have the same variance, but we allow for spatial dependence that are specified in the way we explained in page 5.
In our case, the DLM with spatial dependence provides a better fit. Moreover, the additional assumptions made in this model are reasonable therefore, out of the models that we have tried, it is the one that we would keep to forecast and make decisions.
Another way would be to investigate the relationship between the level of PM_{25} and some other variables (*wind, temperature*).