Time series: Final Project

Matteo Almici (3132333), Federico Golonia (3128604)
Michele Strazza (3112839), Matteo Valetto (3136576)

1. Motivation

Air pollution is a significant issue that seriously impacts human health. In the past decades extensive studies have been carried out to detail how individuals are affected by it and more widely, the issues it generates. The most commonly monitored categories of particulate matter are $PM_{2.5}$ and PM_{10} (particulate matter of diameter 2.5 and 10 micrometer or less, respectively). When concentrations of these particles exceed established safety thresholds, they pose significant health risks and are considered hazardous. The aim of our work is to model the dynamics of air pollution, using measures of air quality, in order to provide a valid statistical analysis and useful insights on this issue to politicians, helping them in the decision making process when it comes to making policies and suggesting some positive behavioral changes to the citizens.

2. Data Description

For the aforementioned purpose, we exploit hourly air quality data from the U.S. Environmental Protection Agency (EPA). In particular, we collect information about 10 stations located along the U.S. West Coast over a period from June to September 2020, including the wildfire season. The final dataset contains several variables, including the spatial coordinates of the EPA station (Longitude and Latitude), the timestamp, air temperature in Celsius, wind speed in knots/second, station identifier and, most importantly, $PM_{2.5}$ per cubic meter. Notice that the suggested limit of $PM_{2.5}$ is 25 micrograms per cubic meter (average over 24 hours). Over this limit it is considered to be dangerous to human health.

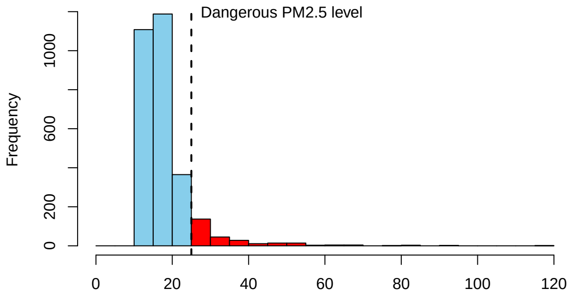

Figure 1: $PM_{2.5}$ histogram of station 96

	PM25	Temperature	Wind
Min.	10.51	13.89	0.08
1st Qu.	14.11	27.22	6.66
Median	15.71	31.11	11.46
Mean	17.92	31.04	10.10
3rd Qu.	19.01	35.00	12.91
Max.	117.11	42.78	18.76

Table 1: Summary statistics of station 96

In this first section, we are going to focus on station 96. First, we provide relevant summary statistics in Table 1 about $PM_{2.5}$, wind speed, and air temperature. The $PM_{2.5}$ data series presents some interesting characteristics. The mean value is 15.71, while the median value is slightly higher at 17.92. This is a positive aspect as both values are under the safety threshold of 25. We also observe a wide range between the minimum and maximum values, with the minimum at 10.51 and the maximum being significantly higher at 117.11.

Our distributional results are visually confirmed by Figure 1. The majority of the measurements are smaller than the prescribed limit, although there are some concentration spikes. For instance $PM_{2.5}$'s most pronounced peak was registered in September and most likely corresponds to the "Orange Skies Day", which took place in the San Francisco Bay Area. September 9, 2020 was marked by unusual orange-hued skies, a result of the smoke and ash from over 20 different wildfires scorched near the San Francisco Bay Area, causing air quality to worsen significantly.

In order to evaluate how other variables in the data set interact with air pollution, we conduct the analysis of cross-correlation between environmental factors and $PM_{2.5}$ concentration levels. Interestingly, temperature shows the highest correlation with $PM_{2.5}$ levels at a lag of -16 hours (16 hours prior) with a correlation coefficient of 0.17. This indicates a relatively low but existing relationship and, specifically, that the temperature from 16 hours before is correlated to the current levels of $PM_{2.5}$. The same considerations hold for wind. The highest wind correlation is observed with a lag of -6 hours (6 hours before) and the specific value of the coefficient is about 0.13. These cross-correlation findings suggest that both temperature and wind speed from previous hours have an influence on current $PM_{2.5}$ concentrations, although the these relationship are not particularly strong.

3. HMM Model Specification

We now turn to model specification, using data coming from station 96. Figure 2 renders explicit what we previously discussed, showing that the majority of the measurements of $PM_{2.5}$ take values smaller than the prescribed limit. To get a better understanding of the underlying process we decided to manipulate the data implementing a 24-hours moving average (given that the critical $PM_{2.5}$ is calculated over a 24 hours mean). The MA model can provide interesting insights as it smooths out random shocks which are exogenous and thus should not play an important role in the state identification.

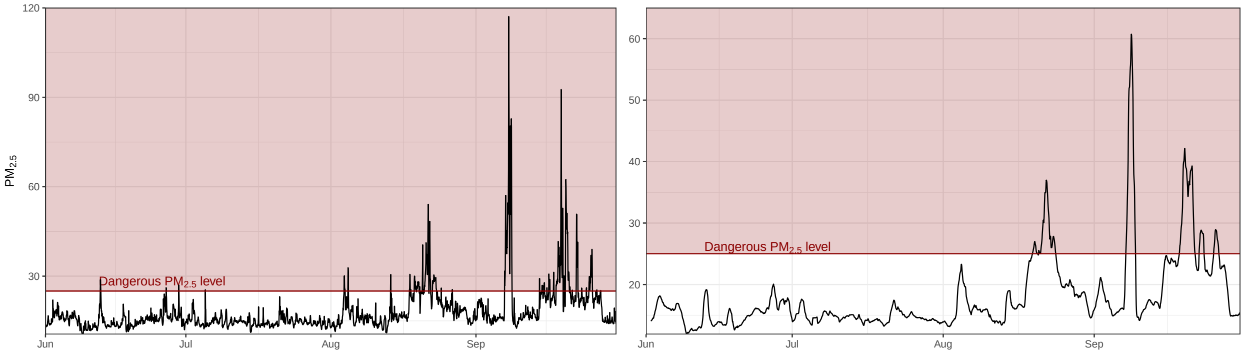

Figure 2: $PM_{2.5}$ Levels on Station 96, Hourly Data (left) and Moving Average (right)

To analyze our data we decided to implement a Hidden Markov Model (HMM) to help identify the different levels of pollution and their instability through the states that the HMM will estimate. Following the consideration above, we will model both the hourly data and the 24 hours moving average. The crucial choice for modeling well the data using the HMM is the one related to the number of states which we will use to tune the model. Observing the time series, we see that in both cases (hourly and MA data) the majority of data seems to fluctuate around a common low mean value with relatively low variance and, especially starting from the end of August, we see the observations change level and seem increasingly volatile around a new higher mean. Another possibility which seemed consistent with the data observed was that of a third intermediate state to address the shocks that we observe throughout the time series between June and August, allowing for the top state to fit better the huge spikes observed in September. Given the uncertainty we faced for the choice of the number of states, we decided to proceed with an *empirical approach* fitting two models with different number of states and then evaluating the estimates produced to choose the most promising one. Thus the model looks as below.

$$Y_{t;k}|S_{t;k} = i \overset{indep}{\sim} N(\mu_{i,k}, \sigma^2_{i,k})$$

$$\text{where } \{S_{t,k}\} \text{ is an unobservable Markov chain and}$$

$$i = \{\text{high}, \text{medium}, \text{low}\} \quad \textit{or} \quad \{\text{high}, \text{low}\} \qquad k = \{\text{Hourly Data}, \text{Moving Average}\}$$

4. HMM Model Fitting

The following are the estimates for the model with the two states and the emission parameters for the one with three states. The comparison suggests to choose the process with 2 states as it is the most accurate for the data. Indeed, the 3-states-model estimates mean values for the low and medium state are close to each other (see Tables 4 and 6) suggesting overfitting issue, whereas the 2-state model seems to "group" the two lower states and estimates the higher state with a greater value than the other model specification. Finally, we also observe that the standard errors of the 2-states specification dominate the 3-states ones as testified by the narrower confidence intervals.

Table 2: Transition matrix - Hourly Data

	to High	to Low
from High State	0.971	0.029
from Low State	0.011	0.989

Table 3: Moving Average

to High	to Low
0.990	0.010
0.004	0.996

Table 4: State parameters - Hourly Data

	par	se	2.5%	97.5%
High State (μ_1)	25.843	0.384	25.090	26.596
High State (σ_1)	10.237	0.260	9.728	10.747
Low State (μ_2)	14.980	0.040	14.901	15.059
Low State (σ_2)	1.648	0.032	1.585	1.710

Table 5: Moving Average

par	se	2.5%	97.5%
24.961	0.275	24.421	25.501
7.525	0.184	7.165	7.886
15.087	0.032	15.025	15.150
1.289	0.024	1.242	1.335

Table 6: State parameters - Hourly Data (3 states)

	par	se	2.5%	97.5%
High State (μ_1)	26.563	0.405	25.769	27.358
High State (σ_1)	10.586	0.282	10.033	11.139
Medium State (μ_2)	16.535	0.060	16.418	16.653
Medium State (σ_2)	1.286	0.037	1.214	1.357
Low State (μ_3)	13.800	0.040	13.721	13.879
Low State (σ_3)	0.904	0.024	0.857	0.950

Table 7: Moving Average (3 states)

par	se	2.5%	97.5%
25.628	0.291	25.057	26.199
7.558	0.193	7.179	7.936
16.475	0.046	16.386	16.565
0.842	0.028	0.787	0.898
14.169	0.028	14.114	14.223
0.714	0.017	0.681	0.747

Having determined the preferred specification (2 states), we now highlight the most interesting insights for our analysis. Both models (Hourly and MA) display a more stable Low state (see Tables 2 and 3), meaning that the transition probability from Low to Low is higher than the one from High to High. Moreover, in both models, the High state is significantly more volatile than the Low one (see Tables 4 and 5) which can be explained by it fitting the high peaks observed in the time series.

As for the differences between the models, the MA one is characterized by a higher state stability. For what concerns the emission estimates, the MA estimated mean for the High state is lower than the one using the Hourly data, meanwhile for the Low state it is the other way around. Notably, the MA estimates for the state variances are lower for both states. This is in line with our expectations as, taking the moving average of observations over the time period will make the peaks less pronounced, lowering the estimate for the high state as well as the variance estimate.

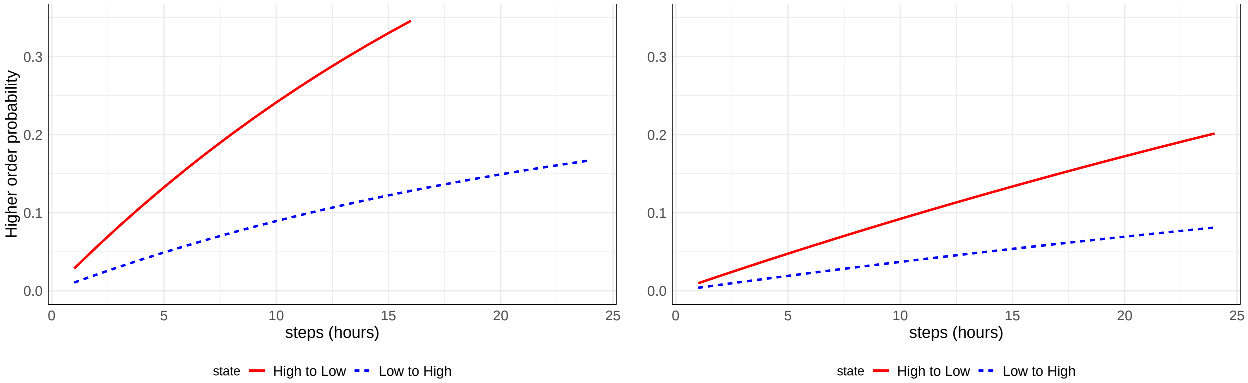

Figure 3: Higher order transition probabilities - Hourly data (left) and Moving Average (right)

To analyze the probability of seeing a state change in the following hours, we decided to compute the higher order transition probability of going from the Low to the High state and viceversa in a fixed number of steps, as shown in Figure 3. Once again, the MA model displays more stable states, as the higher order probabilities increase at a slower rate. For both models, it is evident that it takes significantly less to move from High to Low than viceversa.

5. HMM States Decoding

Having estimated the parameters for the two states of our Hidden Markov Model, in Figure 4 we plot the estimated state for each point in time together with its associated estimated standard deviation, for both models.

As expected, the models are quite similar in terms of decoding as they both assign predominantly a Low state to the period before August, and then assign a high state to match the peak periods seen from August onward. Interestingly, for both models, the peaks touched in August and September seem to represent unprecedented conditions in terms of magnitude, since they fall significantly outside the shadow cast by the High state's estimated standard deviation. Again we highlight the higher instability of the hourly model. Lastly, we underline that in the hourly model the high state $PM_{2.5}$ level is above the critical threshold, whereas in the MA one it is barely below.

6. DLM Specification and online forecasting

In order to make online predictions using the streaming data we need to implement a DLM and to do this we apply a series of improvements to our specification. Initially, we transform our data applying a logarithmic transformation and then take a 12-hours coarse average. By doing this, we get rid of some of the excess noise and have a measure that, from a policy perspective, is more insightful as being above the threshold for 12 hours has stronger consequence on health than for a single hour. Subsequently, we fitted a Dynamic

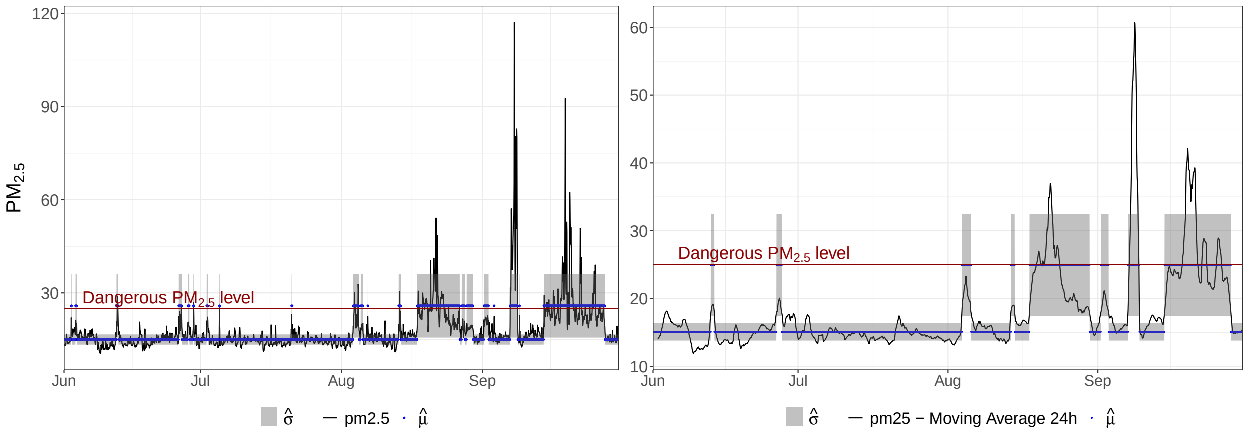

Figure 4: $PM_{2.5}$ Levels and State fitted mean values - Hourly data (left) and Moving Average (right)

Linear Model (DLM) to the transformed data from station 96. The model utilized is a random walk plus noise, which can be expressed as:

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

The relevant parameters, namely σ_v^2 (measurement error variance) and σ_w^2 (evolution error variance), were estimated through Maximum Likelihood Estimation (MLE). The results of the optimization algorithm are presented in Table 8. It's noteworthy that the process exhibits a high signal-to-noise ratio.

	Parameter Estimate	Standard Error
σ_v^2	0.0017363	0.0050258
σ_w^2	0.0196532	0.0091439

Table 8: Maximum Likelihood Estimates and Standard Errors

Its effect is clearly visible in Figure 5 (left). The one step ahead predictions very much resemble the data path given the high degree of significance assigned by the Kalman filter equations to the observations. Moreover, the CI highlights how, for most of the data points predicted to be below the critical level, the distribution of our one-step ahead forecast lies mainly below the dangerous level. This is excellent as it means that, even if our point forecast is incorrect, we rarely would be predicting a low-threat situation when in reality there is danger of high pollution.

7. Introduction to spatial dependence

To model the different stations jointly we decided to check whether the spatial dimension plays a relevant role in explaining the evolution of the pollution level. By this we mean whether, when a shock hits a specific area, it will propagate in space and influence the amount of air pollution detected from the various stations, differently according to the distance they are from the shock (i.e. closer stations see a stronger effect).

Looking at the stations'$PM_{2.5}$ levels they seem to show very similar movements. In particular, the evolution is similar for close stations: 41 and 47 evolve similarly and 96 and 99 as well (see Figure 5 on the right). In fact stations 96 and 99 are located in close proximity to Las Vegas (Nevada) and the distance between them is just of roughly 14 km. On the other hand 50 km separates stations 41 and 47 which are situated on the eastern side of San Francisco Bay Area (California). To give an idea stations 41 and 99 are located 560 km apart, in fact, the evolution of $PM_{2.5}$ for these stations is remarkably different.

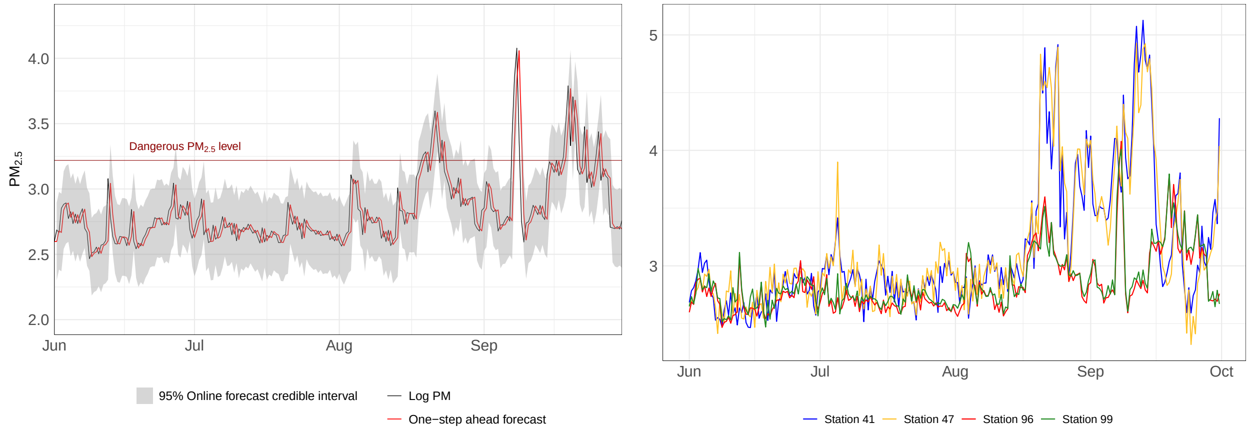

Figure 5: One step ahead forecasts for station 96 (left) - $PM_{2.5}$ Levels across stations (right)

8. Spatial DLM specification

Having analyzed the relevance of the spatial dimension for our data we proceed in implementing a joint spatial-temporal model which includes data from the four stations. The model looks as below:

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

Where $Y_t = (Y_{t,j=41}, Y_{t,j=47}, Y_{t,j=96}, Y_{t,j=99})'$ is the vector of $PM_{2.5}$ data for our stations each evolving according to a random walk plus noise with underlying state process $\theta_{t,j}$. The F and G matrices are both identity matrices of the fourth order; instead the V matrix is a diagonal matrix that includes the measurement error variances for each station ($\sigma^2_{v,i}$). On the other hand the evolution errors are spatially dependent. The structure of the W matrix is the following: $W[j,i] = Cov(w_{j,t}, w_{i,t}) = \sigma^2 exp(-\phi D[j,i])$. Where σ^2 is the common evolution variance, meanwhile the evolution covariances, which express spatial dependence, decrease with the distance across stations($D[j,i]$) according to a decaying parameter ϕ. We end up with 6 paramaters to estimate by MLE.

9. Estimation and forecasting

As a starting point for the numerical optimization algorithm, we have used the estimates of the parameters (i.e. the four measurement error variances and the evolution variance) generated by four univariate DLM's each fitting a single station's data. For the measurement error variances, we took each model's estimate, while for the common evolution variance, we computed a mean of the four estimates. Below are the estimates of V and W.

$$V = \begin{bmatrix} 0.00993 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.01205 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.00010 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.00244 \end{bmatrix}, \quad W = \begin{bmatrix} 0.03047 & 0.02740 & 0.00932 & 0.00923 \\ 0.02740 & 0.03047 & 0.01024 & 0.01014 \\ 0.00932 & 0.01024 & 0.03047 & 0.02959 \\ 0.00923 & 0.01014 & 0.02959 & 0.03047 \end{bmatrix}$$

Comparing the results with the univariate DLM (for station 96), it is evident how the latter estimates a lower value of the evolution variance with respect to the multivariate model. We deem more accurate the multivariate estimate since the underlying model includes data for the four stations and the spatial dimension which, as we have discussed, is relevant to model $PM_{2.5}$. Interestingly, observing the estimated measurement error variance, we note that the stations located in California display higher values than the ones in Nevada.

In particular, station 96 has a very low estimated measurement error variance. To further evaluate this aspect we would need more information about the measurement techniques deployed by the various stations, but it is reasonable to assume that stations located in different states (and in particular with a different proximity to major urban centers) might be equipped with different technologies to detect $PM_{2.5}$.

To visualize the outcome provided by the model, Figure 6 displays both the one-step ahead forecasts and the smoothed process for stations 41 and 47. As mentioned in the introduction, we see that these stations (which are close to San Francisco) have detected abnormally high values in the late summer in correspondence of the extreme conditions they faced.

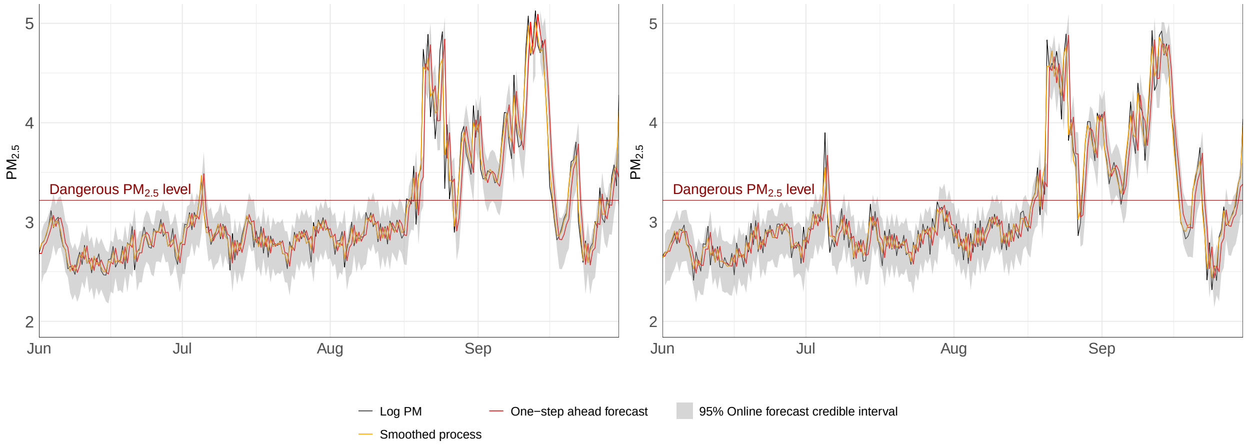

Figure 6: $PM_{2.5}$ Levels, forecast and smoothed porcesses for station 41 (left) and 47 (right)

To make a comparison with the univariate DLM, Figure 7 plots the data for station 96; as mentioned above, the filtering and smoothing procedure of the multivariate DLM for this station continues to resemble very closely the actual data, given that the spatial model assigns a very high signal-to-noise ratio. However, looking at the 7-days ahead forecast and the related credible intervals, it is clear that the model cannot be relied upon for forecasting. The uncertainty of the estimate as the forecast horizon grows is increasingly large, rendering the forecast uninformative for the future state of air pollution.

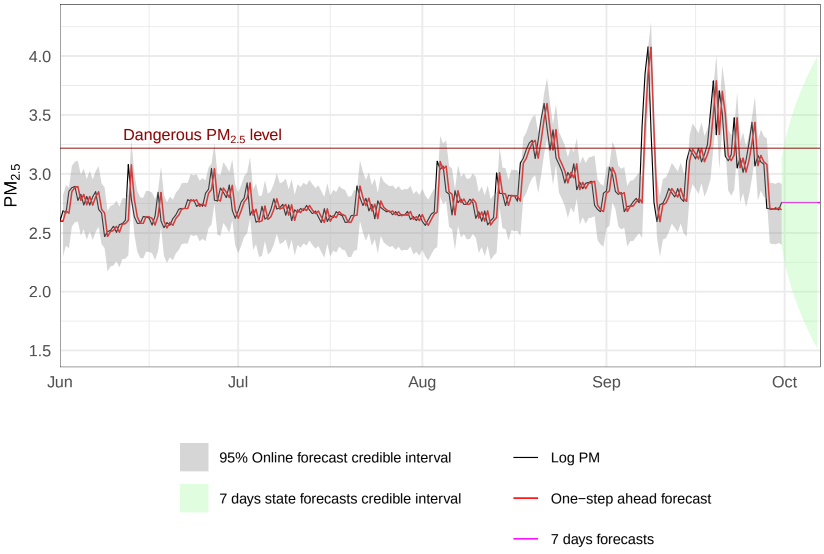

Figure 7: $PM_{2.5}$ levels, one step ahead forecasts anf seven step ahead forecast for station 96

10. Models comparison

To check the assumptions of the two models, Figure 8 plots the standardized forecast errors for station 96 against the theoretical standard normal distribution. As for the univariate model, we see a significant divergence for the extreme quantiles, which are lower in the multivariate DLM. This suggests that the univariate model is characterized by more stringent assumptions; indeed by relaxing some of them (as it happens in the multivariate DLM allowing for spatial dependence) the model's assumptions seem more credible. Furthermore, the same comparison for the other stations shows improvements for both the univariate and multivariate models. This proves again the peculiarity of station 96.

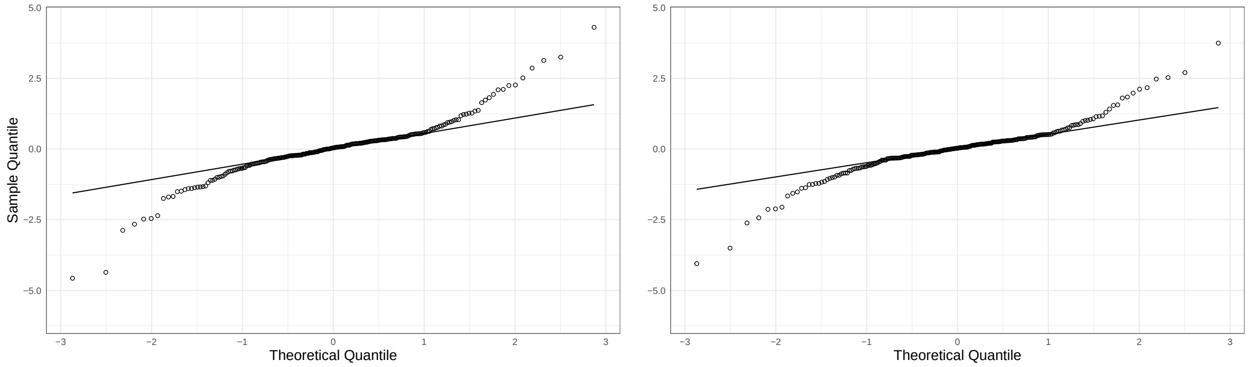

Figure 8: Model checking for the univariate (left) and multivariate (right) model for sation 96

We confirmed the better performance of the spatial model for station 96 by computing the MAPE which for the univariate models is ≈ 0.22 and for the multivariate is ≈ 0.19.

As for the criticisms, we note that the plots of the forecast errors for both DLM's models do not appear to be perfectly distributed as a standard normal. Indeed, looking at the PAC, AC and the Ljung-Box tests the one step ahead forecast error seem to be correlated also with shocks at t-1 and t-2. Thus, the across-time independence of the evolution errors assumed by both models might be restrictive. Moreover, as discussed, assuming independent measurement errors across locations might be a myopic choice given different state regulations and different distance to major cities.

11. Conclusions

Summing up our findings, we deem the HMM useful for a first intuitive understanding of the different levels of pollution and their respective instability. On the other hand, the DLM's models could be used for practical applications such as nowcasting and noise reduction, given their precision in estimating the underlying process and the relatively low MAPE.

Yet the DLM's lose reliability when it comes to h-step ahead forecasts. This in part has to do with the fact that we are trying to predict future values of pollution based on its lagged levels. We might improve this aspect by allowing for additional regressors chosen consulting experts and scientific literature.

We could think of different kind of relevant predictors such as demography, urbanization and climate variables, which are driven by long-term trends that are generally orthogonal to policy actions. Other factors, that can be more directly affected by public policies, might be energy and transportation efficiency.