Final Project - Group 11

2023-06-02

Group members:

Barni Martina (3118929), Figlino Guerino (3111009), Noce Alberto (3225732), Ungarelli Flavia (3092154)

Abstract

Fine particulate air pollution has been proven to have serious health effects, especially on cardiopulmonary health. Short-term exposure and mortality correlation has been well-documented. However, the evaluation of PM health effects at different time-scales of exposure is still ongoing, and findings are essential to design ad-hoc environmental public health policies.

In our project we aim to model the dynamics of air pollution. We use hourly air quality data coming from 10 US West Coast stations during the summer of 2020, detected by the U.S. Environmental Protection Agency (EPA). In the first part, we describe our data and try to estimate the instability of levels of air pollution, and the persistence of high levels of $PM_{2.5}$ with a three state Hidden Markov Model. In the second part, we try different model specifications (i.e., univariate and multivariate DLMs), and we compare the results in terms of estimated parameters, one-step ahead predictions and forecast errors.

The first question that we aim to answer is how to recognize the different levels of pollution, which is important for policymakers who must intervene rapidly and plan ahead. This requires an understanding of how likely it is that high $PM_{2.5}$ levels will persist in the following hours. Secondly, we provide online estimation and uncertainty quantifications, which call for a cautious monitoring of the situation. Finally, we show that interventions should be coordinated across municipalities, because $PM_{2.5}$ levels in other locations provide relevant information on the dynamics of $PM_{2.5}$ levels in any nearby location.

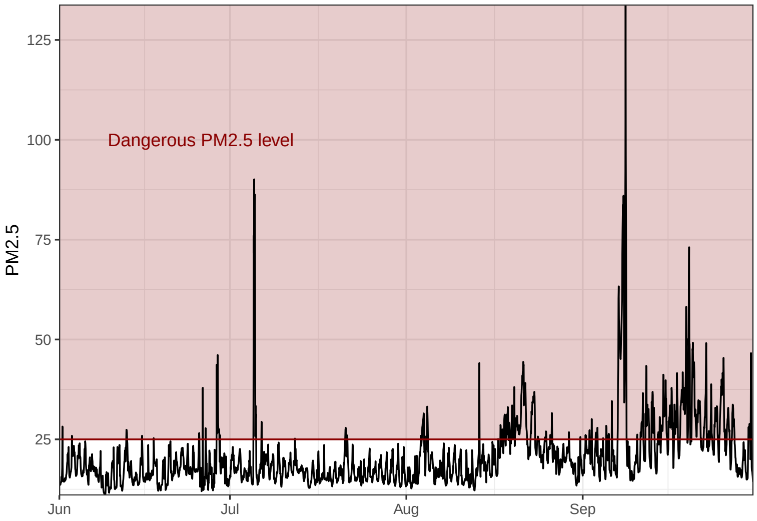
Data

The dataset used in this project reports hourly measurements of $PM_{2.5}$ for 10 stations along the Western Coast of the US. The dataset contains 29,280 observations gathered over the summer of 2020, starting from June 1st at 00:00 GMT and ending on September 30th at 23:00 GMT. $PM_{2.5}$ levels are reported in $\mu g/m^3$ throughout the report.

Station 101

Given that the time series of $PM_{2.5}$ levels recorded at each of the stations in the dataset are rather different, we focus our analysis on station 101, located in Las Vegas, Nevada.

Figure 1: PM2.5 levels at Station 101



Overall, the time series of $PM_{2.5}$ levels at station 101 during the summer months (Figure 1) can be reasonably divided in two different time periods: the first one spanning from June to the first half of August, where $PM_{2.5}$ levels are consistently below the suggested limit of 25 $\mu g/m^3$, and the second one going from the middle of August to the end of our observation period, where $PM_{2.5}$ levels are frequently exceeding the danger threshold. The time series appears also more stable in the first half of the measurement period, with instability increasing mid-August. There are several spikes in $PM_{2.5}$ levels throughout the summer months, which may be associated with local or regional air pollution events such as wildfires or industrial emissions. For instance, the outlier recorded at the beginning of July may be explained by the festivities of the 4th of July, which is a national holiday in the US. Taking into account the whole sample, the highest and lowest $PM_{2.5}$ levels recorded at station 101 are respectively, $133.81\mu g/m^3$ and $11.01\mu g/m^3$, while the mean of the observations is $20.81\mu g/m^3$ (lower than the 25 $\mu g/m^3$ threshold).

Model specification

Hidden Markov Model

We specify the model to be a HMM with three states, and Gaussian emission distributions, with state-dependent mean and variance. The HMM allows us to model a non-stationary time series that cannot easily be transformed into a stationary one. Before settling for three states, we also tried HMMs with two and four states: while the former presented higher AIC and BIC scores, the latter would have offered a better fit. Nonetheless, the three-states specification offers a good compromise because it simplifies both the computations and the interpretation for policymakers. The three states account for low levels of $PM_{2.5}$, moderate levels of $PM_{2.5}$ (both below the so-called dangerous level of air pollution), and high levels of $PM_{2.5}$, that were mainly recorded during the wildfire season (Figure 2). Data shows that two states (low and moderate $PM_{2.5}$) have kept alternating in the first period. Starting from mid-August, the minimum levels of $PM_{2.5}$ increased and air pollution occasionally reached peaks well above the dangerous level.

The HMM model that we estimate is:

$$Y_t = \mu_i + \epsilon_t \qquad \epsilon_t \overset{iid}{\sim} N(\mathbf{0}, \sigma_i^2) \quad \text{for state } S_t = i \text{ with i} \in \{1, 2, 3\}$$

The model fitting generates the following MLEs for the unknown parameters. Starting from our sample data, it estimates low levels of $PM_{2.5}$ to be $15.663\mu g/m^3$ on average (State 2), moderate levels of $PM_{2.5}$ to be $21.238\mu g/m^3$ on average (State 1) and high levels to be $34.679\mu g/m^3$ on average (State 3) (Table 2). As expected, the variance increases in higher $PM_{2.5}$ levels states (Figure 2 gives a visual representation, with the variances on the states presented in light blue).

Table 1: Transition matrix

	To Low	To Moderate	To High
From Low	0.94	0.058	0.002
	(0.007)	(0.007)	(0.001)
From Moderate	0.079	0.894	0.027
	(0.009)	(0.010)	(0.006)
From High	0.001	0.06	0.938
	(NA)	(0.012)	(0.012)

Table 2: Mean and Standard Deviation

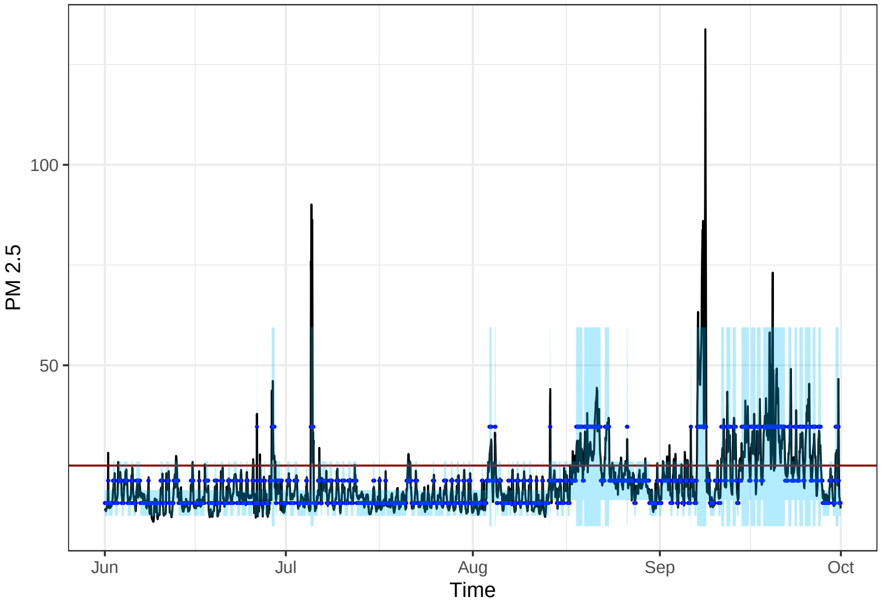
	Low	Moderate	High
Mean	15.663	21.238	34.679
	(0.0568)	(0.1178)	(0.6062)
Standard Deviation	1.606	2.419	12.401
	(0.0381)	(0.0738)	(0.4022)

Our estimated model suggests that if we detect a high level of $PM_{2.5}$, it will be followed by high levels also in the following hour in 93.8 out of 100 cases. In general, we see that the probability that there is no state change in two subsequent observations is high for all states. In particular, the probability to move from a high level of air pollution to a moderate level in the next hour is 6% and the probability to observe a significant decrease to low levels of $PM_{2.5}$ is 0.1%. This probability is almost of the same order as the one of moving from a low level of air pollution to a high one (0.2%). As Table 3 displays, after 6 hours, the probability of persistence of high $PM_{2.5}$ levels is still 70%, but after 24 hours the probability of persistence of high levels is down to 32%, which is similar to the probabilities of observing low and moderate levels (31% and 35% respectively). After two days the low state is the most likely (43%), followed by the moderate state (35%). This might be explained by the fact that in the case of an extraordinary event, like a wildfire, the danger will be contained in 24 hours and thus the $PM_{2.5}$ will be back to either low or moderate levels. Only one third of the times, the wildfire is so serious so that it will still be ravaging after 24 hours and the level of $PM_{2.5}$ will still be high. All in all, this means that intervention should be focused in the first few hours following any particular event. We find that the estimated persistence of high pollution levels is much lower in station 101 than in other stations, such as station 47, which is closer to San Francisco, where the probability of still observing high levels after 24 hours is above 70%. The persistence at the nearby stations, #97 and #103, is instead more comparable. It is important to understand what geographic and atmospheric conditions generate higher $PM_{2.5}$ persistence in order to predict how many resources will need to be employed, and for how long, in order to contain the danger.

Table 3: Persistence of High PM2.5 levels

Hours after	1	2	3	4	5	6	12	24	36	48	60
High	0.938	0.881	0.830	0.782	0.739	0.699	0.516	0.323	0.238	0.199	0.182
Moderate	0.060	0.110	0.152	0.186	0.215	0.239	0.318	0.350	0.351	0.350	0.349
Low	0.001	0.007	0.016	0.028	0.041	0.057	0.157	0.313	0.394	0.431	0.448

Figure 2: PM 2.5 data and HMM estimated means (3 states)



Univariate DLM

We observe that the time series is non-stationary and presents some evident change points, in particular in the wildfire season in August and September, where $PM_{2.5}$ levels are higher and more volatile. Therefore, we use a *local level* model to capture the main change points and other minor changes. Let $(Y_{j,t}, \theta_{j,t})$ denote, respectively, the observed measurement and the signal in station j at time t. Then, for any single station, we can consider the following simple random walk plus noise model:

$$\begin{cases} Y_{j,t} = \theta_{j,t} + v_{j,t} \\ \theta_{j,t} = \theta_{j,t-1} + w_{j,t}, \end{cases}$$

with the assumption that $\theta_{j,0} \perp (v_{j,t}) \perp (w_{j,t})$.

Considering the station 101, we estimate the parameters of the model with MLE, in order to be able to provide online estimation, using hourly data, as it streams in. We try different initial values for the optimization.

Before we do this, to smooth the sharp peaks and irregularities of the hourly data we use a log scale of $PM_{2.5}$ levels. We estimate variance of the observation, equation to be 0.000001 (0.00012) and the variance on the state equation to be 0.01044 (0.00026), where in parenthesis we report the asymptotic standard errors. However, if we use these estimates for forecasting, we obtain very noisy forecasts and a cluttered plot. To make the data analysis more informative, we therefore work with 12-hour averages. The estimates we obtain are 0.01113 (0.00280) and 0.012647 (0.00355) for the observation and state equations respectively. The estimated variance of the observations equation is higher, due to the large hourly fluctuations, but it is estimated more precisely. We plot the data, together with the one-step ahead forecasts and their 0.95 credible intervals in Figure 5. This procedure can be conducted recursively in order to monitor pollution levels and decide whether to intervene, and how intensively.

Multivariate DLM

Figure 3 displays that there exists spatial correlation of the $\text{PM}_{2.5}$ levels of stations close to each other. Indeed, $\text{PM}_{2.5}$ levels tend to be more similar over time in pairs of closer stations (i.e., 97 and 101, 101 and 103, 103 and 97), rather than farther stations (i.e., 47 and each of the other three stations considered).

Hence, it is more informative to specify a multivariate model. Given the large difference with station 47, we only consider close stations (i.e., 97, 101 and 103), and Y_t is now a 3-dimensional vector of the PM$_{2.5}$ observed at these stations.

Figure 3: 12-hour average PM2.5 levels

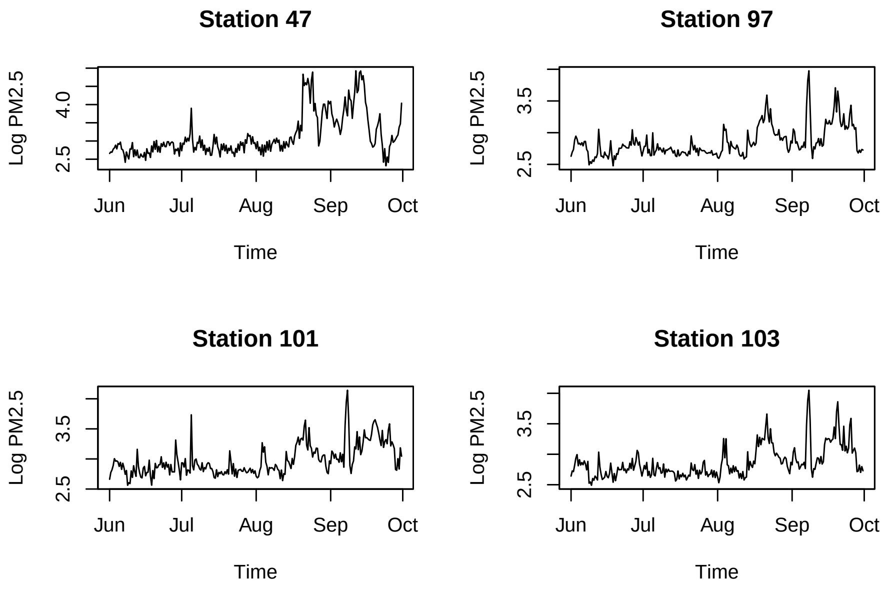
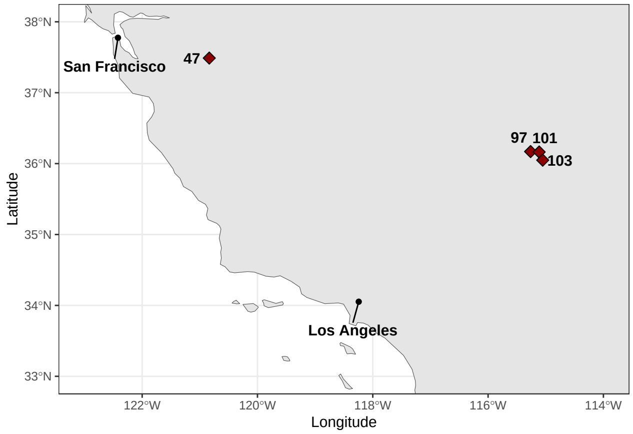


Figure 4 displays the location of the selected stations.

Figure 4: Map of the stations considered



We define a multivariate 3-dimensional DLM for $Y_t = (Y_{1,t}, Y_{2,t}, Y_{3,t})'$, so for the PM$_{25}$ observeded at stations $j = 1, 2, 3$.

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

where m=3 and each $(Y_{j,t})$ is described as a random walk plus noise as above. F and G are 3×3 identity matrices, and θ_t is 3×1 vector $\theta_t = [\theta_{j,t}]'$, for $j = 1, 2, 3$. We assume that the measurement errors $v_{j,t}$ are independent and constant across locations j, with location-specific variances. Hence, V is defined as follows:

$$V = \begin{bmatrix} \sigma^2_{v,1} & 0 & 0 \\ 0 & \sigma^2_{v,2} & 0 \\ 0 & 0 & \sigma^2_{v,3} \end{bmatrix}$$

Instead, we model the spatial dependence by assuming that the evolution errors are spatially correlated. We assume an exponential covariance function W to be:

$$W[i,k] = Cov(w_{j,t}, w_{k,t}) = \sigma^2 \exp(-\phi D[j,k]), \quad j,k = 1,\ldots,3;$$

where $\sigma^2 > 0$; $\phi > 0$ is a *decay parameter*; and $D[j,k]$ is the Euclidean distance between stations j and k. This function implies that the $Cov(w_{j,t}, w_{k,t})$ is smaller for stations that are further apart.

We report below the estimates of the unknown parameters:

Table 4: Estimated Variance-Covariance Matrix V

	Station 97	Station 101	Station 103
Station 97	0.00202	0	0
	(2.704106e-06)	-	-
Station 101	0	0.0051	0
	-	(4.161918e-06)	-
Station 103	0	0	0.00318
	-	-	(2.642621e-06)

Having estimated $\hat{\sigma}^2 = 0.8024$ (9.780938e-04) and the decay parameter $\hat{\phi} = 0.001$ (7.894010e-07), we compute the estimate for W. The covariance matrix W captures the spatial dependence and is estimated using an exponential covariance function. The estimated values of W are presented in Table 5. Similar to V, the values in W represent the variances and covariances between stations, but in this case, they reflect the spatial correlation between the evolution errors. The decay parameter ϕ controls the strength of the spatial correlation, and smaller values of ϕ indicate that the correlation decreases more rapidly with distance. Based on the estimated values of V and W, it seems that there is spatial correlation in both the measurement errors and the evolution errors. The values in the matrices indicate that stations closer to each other have higher correlation coefficients, while stations farther apart have lower correlation coefficients. This aligns with the information mentioned earlier that $PM_{2.5}$ levels tend to be more similar over time in pairs of closer stations.

Table 5: Estimated Variance-Covariance Matrix W

	Station 97	Station 101	Station 103
Station 97	0.80245	0.80233	0.80225
Station 101	0.80233	0.80245	0.80234
Station 103	0.80225	0.80234	0.80245

One-step-ahead predictions

We report the plots of one-step-ahead predictions for the univariate (Figure 5) and the multivariate (Figure 6) models. In the univariate case, a signal-to-noise ratio larger than one determines a higher reliance of the last observation. Indeed, the one-step ahead forecast in this case is close to be a one-step rightward shift of the observed data. This behavior is less evident in the multivariate case, as one-step-ahead predictions do not only rely on past observations of Station 101, but they are affected by the noise generated by the interaction with close stations.

Figure 5: One-step-ahead forecasts of 12 hour averages of log PM2.5 levels (Univariate DLM)

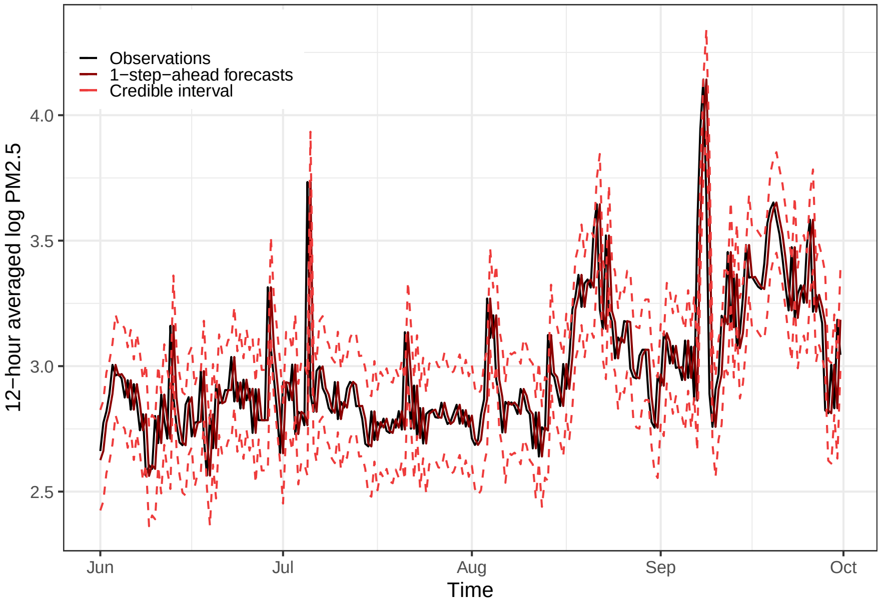
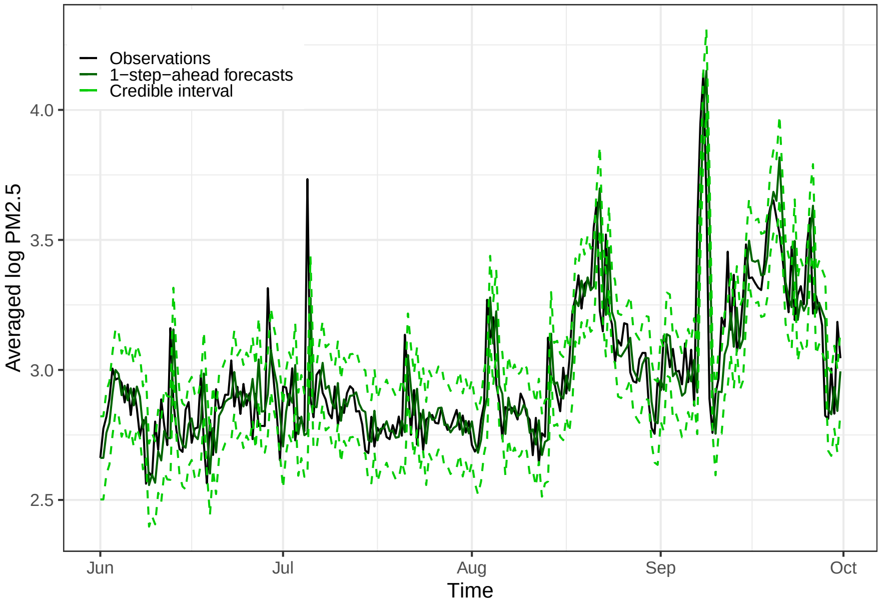


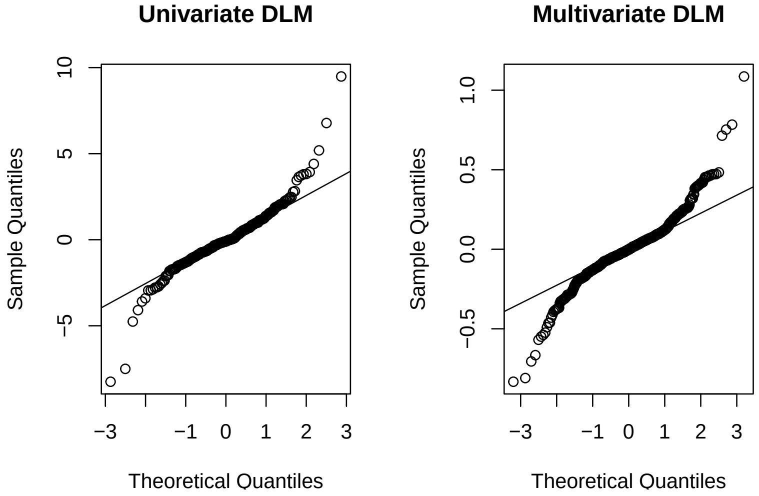
Figure 6: One-step-ahead forecasts of 12 hour averages of log PM2.5 levels (Multivariate DLM)



Model checking

We use a QQ-plot to check whether forecast errors are normally distributed. In the univariate case (left panel in Figure 7), the dots are mostly distributed along the diagonal, suggesting that there seem not to be any meaningful departure from the model assumptions. In the multivariate case (right panel in Figure 7), we observe that dots diverge from the 45 degree line in both direction, suggesting that errors are not perfectly normally distributed. However, since this deviance from normality happens in both directions, we may believe that they do not show skewness, but just thicker tails than a normal distribution. We also run a Shapiro test, and find that we can reject the null hypothesis of normality in the multivariate case.

Figure 7: Normal Q-Q Plot



The presence of non-normality suggests that the assumption of a normal distribution for the forecast errors may not accurately capture the true underlying distribution. This can lead to biased parameter estimates, inaccurate prediction intervals, and unreliable inference about the behavior of air pollution levels. One approach to solve this would be to employ alternative distributional assumptions that better capture the observed characteristics of the forecast errors, so that the model can better account for the deviations from normality and provide more reliable estimates and predictions. Additionally, we compute the MAPE for both the univariate and multivariate forecasts, and these are respectively 0.038 and 0.040, suggesting that the multivariate forecast performs slightly better on this measure, at least on the estimation sample.

Conclusions

The main assumptions underlying our two models (i.e., HMM and DLM - both univariate and multivariate), which are both state space models, are that (1) the latent state process θ_t is a Markov chain and that (2) conditionally on (θ_t), the Y_t's are independent and Y_t depends on θ_t only. These are reasonable assumptions, that allow us to deal with a non-stationary time series with change points. Moreover, in the HMM states are discrete-valued random variables. For this purpose the level of pollution, which is a continuous random variable, is discretised into three categories (i.e., low, medium, and high levels of $PM_{2.5}$). Such simplification is realistic as long as the three identified categories of pollution level are meaningful and reflect the way the intensity of pollution is classified by scientists and policy makers. Whereas, the DLM assumes the model to be linear and Gaussian. The assumption of normality has been verified and disussed in the model checking section. A random walk plus noise model is appropriate for time series that show no clear trend nor seasonal variation: we assume that there is an unobservable Markov chain (θ_t) (i.e., a level which is subject to random changes over time, described by a random walk), and that Y_t is an imprecise measurement of θ_t. A suggestion for improving our model would be to increase its explanatory power by incorporating a regression component in the DLM specification. Indeed, besides predicting the future levels of pollution, we are mainly interested in identifying the specific drivers of pollution, in order design ad-hoc policies to address the environmental issue.