Bocconi University, 20236 - Time Series Analysis - Final Project

Efecan Bahcivanoglu, Ozan Ocal, Francesca Manca, Clementine Lauvergne

02/06/2024

Introduction

The issue of human-induced environmental impacts has gained significant attention over recent decades, particularly concerning air pollution—its causes and the role of data in informing policy decisions. High concentrations of airborne particles like PM2.5 have been consistently linked to various adverse health effects, including severe cases of Covid-19 and other respiratory ailments. High-quality data and robust statistical models are crucial for governments and policymakers, enabling them to understand and predict air pollution trends. We have access to data from the EPA regarding hourly air quality in California during the summer of 2020. We'll use various statistical models to analyze PM2.5 concentration such as Hidden Markov Models, Univariate and Multivariate Dynamic Linear Models.

Data Summary

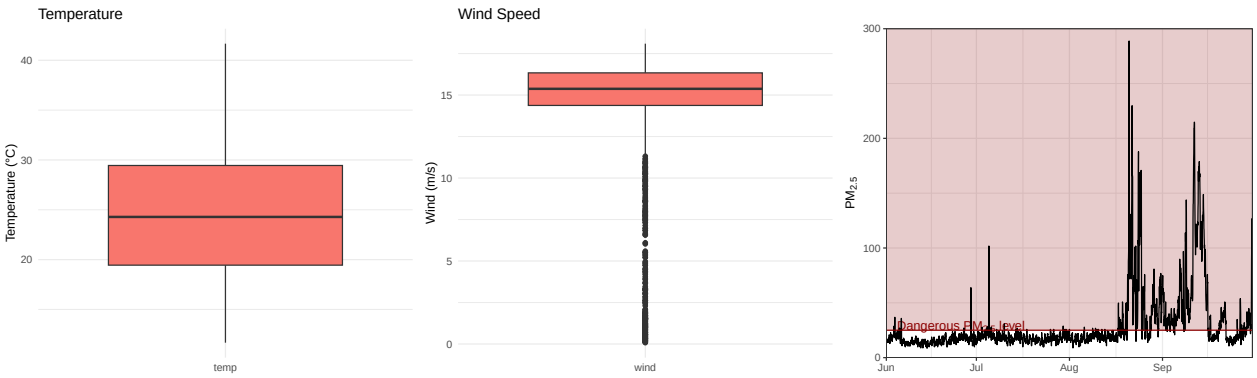

Figure 1: Station 41: Box Plot of Wind and Temp with Plot of PM2.5

First we look at the boxplots of complementary time serieses of temperature and wind speed. Existence of outliers in wind speed compared to temperature suggests while temperature follows a structured behavior throughout the summer wind speed fluctuates and thus possibly effects the PM2.5 levels. When we look at the PM2.5 levels we can observe a mildly stationary with low variance till the middle of August; however, after August as the volatility increases mean increases too. this behavior can be explained by the West Coast wildfire period, which worsens PM2.5 levels.

We furthermore analyze the cross-correlation and the covariance between wind speed, temperature and pM2.5 levels. For the wind, a negative cross-correlation value (-0.05483028) indicates an inverse relationship between wind speed and PM2.5 levels. This occurs at lag 0 and it is consistent with the understanding that increased wind can help disperse air pollutants, thereby reducing PM2.5 concentrations. For the temperature, there is a positive cross-correlation value (0.06909771) occuring at 20 hours prior. Last but not least, the magnitude of the cross-correlations suggest while there's a detectable linear relationship it's indeed quite weak.

Hidden Markov Model

To analyze our data, we opted to employ a Hidden Markov Model (HMM) to delineate the various pollution levels and their instability through the estimated states. A pivotal decision in effectively modeling the data with an HMM involved selecting the appropriate number of states to refine the model. The bulk of the data typically oscillates around a relatively low mean with modest variance. Notably, from the end of August onward, the data transitions to a phase where observations become more volatile, clustering around a higher mean value. The consideration of adding a third, intermediate state seemed justified by the observable fluctuations between June and August, helping to better accommodate the significant spikes seen in September. Faced with uncertainties about the optimal number of states, we chose to adopt an empirical approach. We fitted two models with varying numbers of states and compared their outputs to determine the most suitable configuration. The model configuration is presented as follows.

$$Y_t|S_t \overset{ind}{\sim} N(\mu_i, \sigma_i^2)$$

where latent states $\{S_t\}$ evolves according to a Markov Chain.

We analyze the data with 2 states: (high, low) and then 3 states: (high, medium, low).

Table 1: HMM 2 States: Transition Matrix

	High	Low
High	0.9911730	0.0088270
Low	0.0232636	0.9767364

Table 2: HMM 2 States: MLE of Mean and S.D.

	Value	SE	2.5%	97.5%
High(μ_1)	17.349008	0.0832255	17.185889	17.512127
High(σ_1)	3.701055	0.0611531	3.581198	3.820913
Low(μ_2)	63.594158	1.5835395	60.490478	66.697838
Low(σ_2)	43.032845	1.0940449	40.888557	45.177134

When we look at the estimated transition matrix probabilities, it's evident that markov chain is persistent in the sense that once state X is entered there's a high probability that it will stay in that state. Furthermore if we compare the values for transitioning from High to Low to transitioning from Low to High we see that it's almost 5 times more likely to transition from High to Low than vice versa. On the other hand if we analyze the MLE of the mean and standard deviation for Low and High states we can see that the mean value for High is 63.59 with 43.02 standard deviation, indicating higher volatility. Our model converged at iteration 10 with negative log-likelihood value of -10052.41. Next, we'll analyze the results of HMM with 3 states and we'll make a comparison with the 2 states.

Table 3: HMM 3 States: Transition Matrix

	High	Medium	Low
High	0.9448077	0.0000000	0.0551923
Medium	0.0043273	0.9736634	0.0220093
Low	0.0493563	0.0188738	0.9317699

Table 4: HMM 3 States: MLE of Mean and S.D.

	par	se	2.5%	97.5%
Low State (μ_3)	14.687016	0.1024573	14.486203	14.887828
Low State (σ_3)	2.198439	0.0572150	2.086300	2.310578
Medium State (μ_2)	20.049482	0.1354776	19.783951	20.315013
Medium State (σ_2)	3.006998	0.0722687	2.865355	3.148642
High State (μ_1)	64.580184	1.5958155	61.452443	67.707925
High State (σ_1)	43.127207	1.1106368	40.950399	45.304016

Again we can see that same persistent behaviour regarding staying in state X once it's reached. It's interesting to see that transition from High to Medium has 0 probability indication while increase the pm25 levels is gradual the decline is sudden. Our model converged at iteration 43 with negative log-likelihood value of -9490. While this log-likelihood value is better than the HMM with 2 states, one must know that as we increase the number of states model will have greater negative log-likelihood value but that doesn't necesarrily implies a better model.

HMM(2) States Decoding

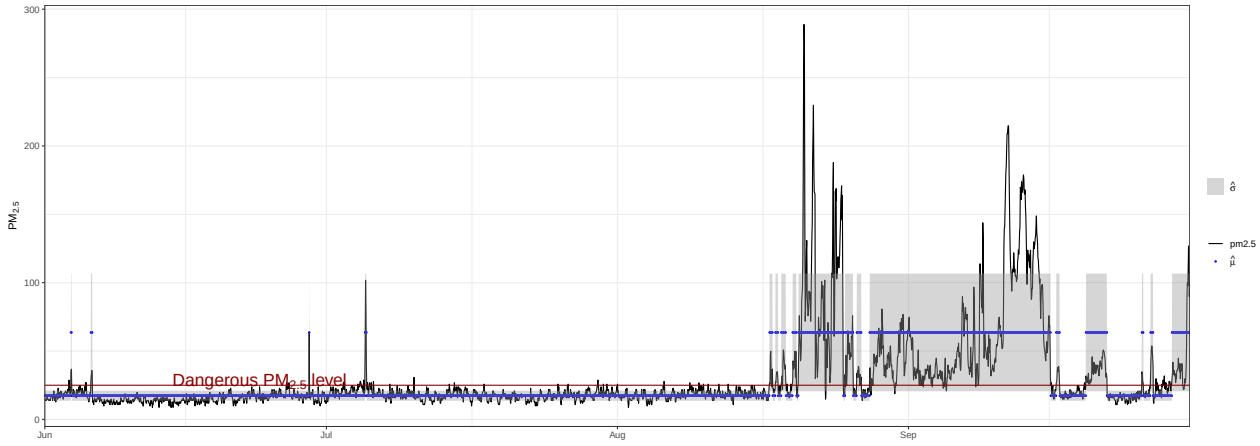

Figure 2: $PM_{2.5}$ Levels and State Values HMM 2 States

Dynamic Linear Model

Before feeding our data to a DLM we transformed it by applying log transformation. Then we tried different coarcer sclaes from 6 hours to 12 hours.It was quite intuitive that we can't use a scale that doesn't divide the 24 hours as each measuremnt then would be taken through different times over the day(10hours: 10am-8pm-4am-2pm and so on). Between 4,6,8 and 12 hours we have calculate the sample variance. As with the PCA we want to use the scale that has the greates variance as it contains information. Though, coarcer scales resulted in a higher variance so we choose the optimal scale based on the trade-off which was the 8 hours.

- Sample variance for 4-hour intervals: 0.3547907
- Sample variance for 6-hour intervals: 0.3473822
- Sample variance for 8-hour intervals: 0.3421984
- Sample variance for 12-hour intervals: 0.3335311

Our model is random walk plus noise:

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

Table 5: MLE and SE of DLM

	Parameter Estimate	SE
σ_v^2	0.0203421	0.0036792
σ_w^2	0.0301789	0.0051134

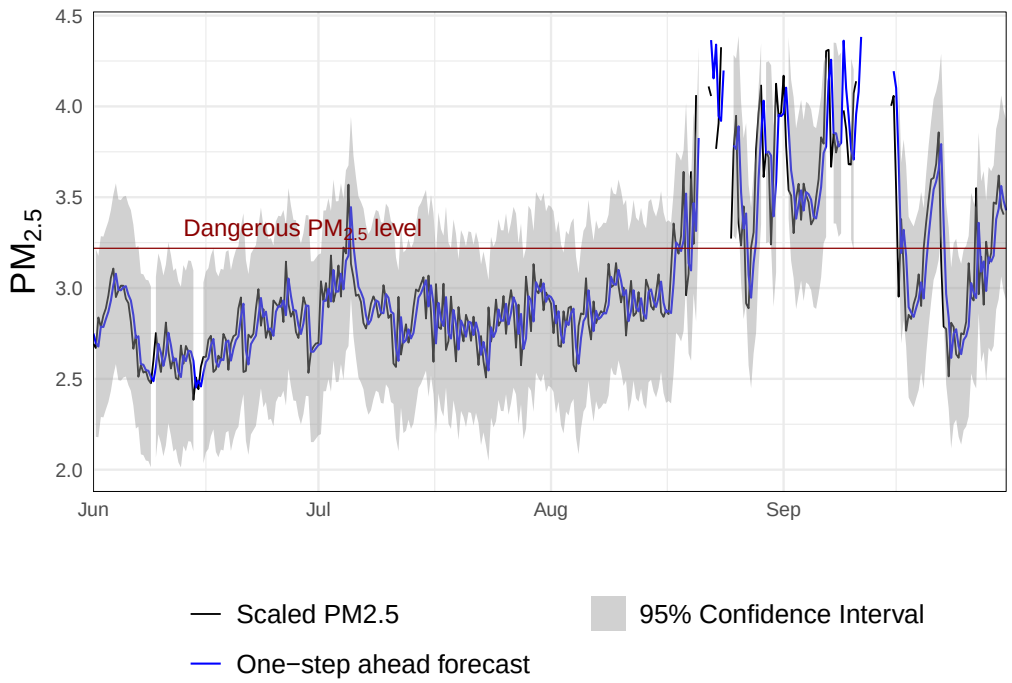

Figure 3: One-step Forecasts for Station 41

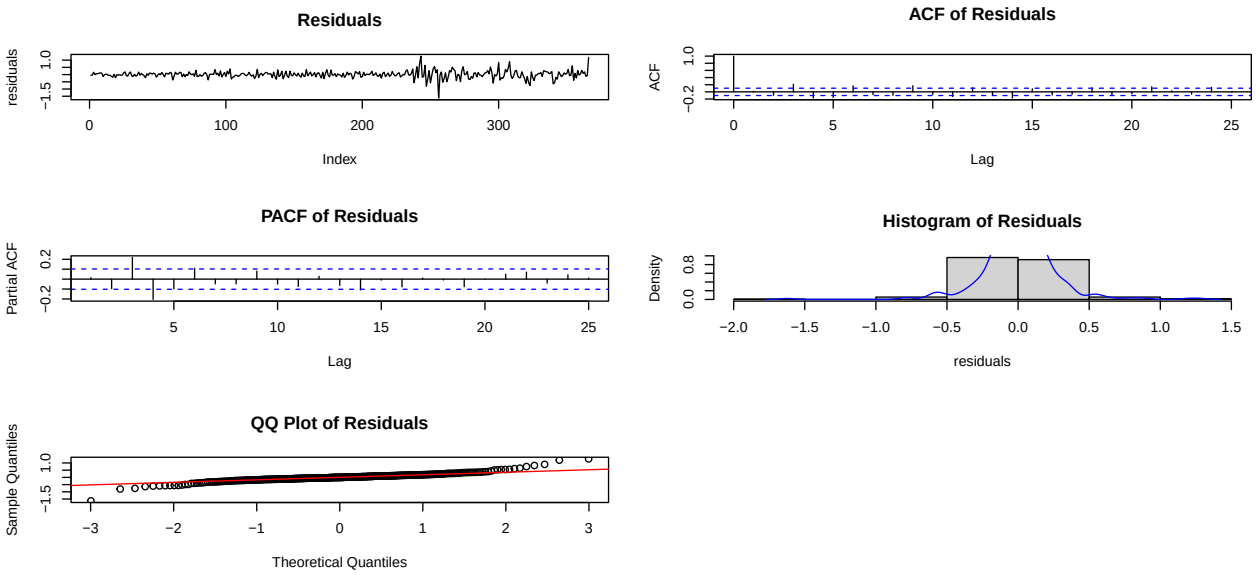

Figure 4: Diagnostic Plots for Station 41: Univariate DLM

The residuals do not display obvious patterns or trends, which generally suggests that the model does not suffer from non-random error structures.However, the variability of residuals appears slightly inconsistent, hinting at potential heteroscedasticity.This is consistent with the peaked levels of PM2.5 during the high state in the HMM model. ACF and PACF plots suggests there's not significant autocorrelation of the residuals.

Spatial Dynamic Linear Model

We created a spatial dlm with using measurements from four statins.

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

$Y_t = (Y_{t,41}, Y_{t,47}, Y_{t,96}, Y_{t,99})'$ is the vector of PM2.5 data from four stations. The F and G matrices are 4x4 identity matrices. The V matrix is also 4x4 diagonal matrix where diagonal terms are the measurement errors belonging to corresponding stations. Spatial dependence is not assumed for the measurement error but state errors are spatially dependent and defined by the $W[j,i] = Cov(w_{j,t}, w_{i,t}) = \sigma_w^2 exp(-\phi D[j,i])$. Where matrix D corresponds to the distance between the stations and ϕ is the decay parameters. In total we have 6 parameters to estimate, the diagonal terms of the V matrix, σ_w^2 and the ϕ. For the distance matrix we also include the elevation and then take the euclidean distance in 3 dimension.

To estimate the unknown parameters we use Maximum Likelihood Estimation where maxima is reached via numerical optimization algorithms. One must take it into consideration that the likelihood function for a DLM may present many local maxima. This implies that starting the optimization routine from different starting points may lead to different maxima. It is therefore a good idea to start the optimizer several times from different starting values and compare the corresponding maxima. This is exactly what we did. Furthermore we use two different methods: Simulated Annealing and L-BFGS-B to explore the MLE space.

Simulated Annealing (SANN)

1. **Initialization**: Start with an initial solution s_0 and an initial temperature T_0.
2. **Iteration**:
   - For each iteration, a new solution s' is generated by making a small random change to the current solution s.

- The change in the cost function $\Delta E = f(s') - f(s)$ is calculated.
- If $\Delta E < 0$, the new solution s' is accepted.
- If $\Delta E \geq 0$, the new solution s' may still be accepted with a probability $e^{-\Delta E/T}$, where T is the current temperature.

3. **Cooling**: Reduce the temperature according to a cooling schedule, typically $T = \alpha T$, where $0 < \alpha < 1$.
4. **Termination**: The algorithm terminates when the temperature is sufficiently low or after a fixed number of iterations.

L-BFGS-B

L-BFGS-B is an optimization algorithm in the family of quasi-Newton methods. It is specifically designed to handle bound constraints and to approximate the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, which uses an approximation to the Hessian matrix to steer its search for the optimum.

The algorithm updates an approximation H_k of the inverse Hessian matrix using the following update rule, which is derived using only first derivatives:

$$H_{k+1} = \left(I - \frac{s_k y_k^\top}{y_k^\top s_k}\right) H_k \left(I - \frac{y_k s_k^\top}{y_k^\top s_k}\right) + \frac{s_k s_k^\top}{y_k^\top s_k}$$

where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

Best negative log-likelihood value, -1956.076, is achieved by the L-BFGS-B algorithm with initial values (0.1,0.1,0.1, 0.1,0.1,0.1). Algorithm converges in 56 iterations. The estimated values of $\phi = 0.0025765$ while V and W are the following:

$$V = \begin{bmatrix} 0.016363 & 0 & 0 & 0 \\ 0 & 0.022290 & 0 & 0 \\ 0 & 0 & 0.0010679 & 0 \\ 0 & 0 & 0 & 0.0030771 \end{bmatrix} \quad W = \begin{bmatrix} 0.0232291 & 0.0204365 & 0.0055460 & 0.0054855 \\ 0.0204365 & 0.0232291 & 0.0062164 & 0.0061396 \\ 0.0055460 & 0.0062164 & 0.0232291 & 0.0224207 \\ 0.0054855 & 0.0061396 & 0.0224207 & 0.0232291 \end{bmatrix}$$

If we compare our results with the univariate DLM for station 41 we can see that V has a lower value in the multivariate case (0.020342 in univariate). This is reasonable as we have a greater information sigma algebra established by introduction of the spatial dimension. Moreover, the measurement error variance for the station 96 and 99 are significanly less than station 41 and 47. More research can be done on if there's a difference on how each individual station collects data.

In Figure 5, we have the one-step-ahead observation forecasts based on the previously estimated parameters. The colored line is the forecasted value ($\mathrm{E}[\theta_t|y_{1:t-1}]$), while the black line represents the actual observed value. The grey colored area is the 95% confidence interval. Forecast values are quite similar to the actual values implying this model can be useful for policy makers.

In Figure 6, we run diagnostics for the forecast errors of station 41.The residuals do not display obvious patterns or trends, which generally suggests that the model does not suffer from non-random error structures. However, the variability of residuals appears slightly inconsistent, hinting at potential heteroscedasticity. As one way to improve the model we can incorporate lag values of the wind and temperature in an ARMA format. Based on the Shapiro normality test even though both univariate and multivariate dlm have normally distributed forecast residuals, multivariate one has a better score implying addition of spatial dependence has a small improvement.

Conclusion

A well-constructed Hidden Markov Model (HMM) offers policymakers a valuable means to grasp the various states of air pollution. With assumptions about differing means and variances across states and since the states are discrete it offers more practical policy implications.

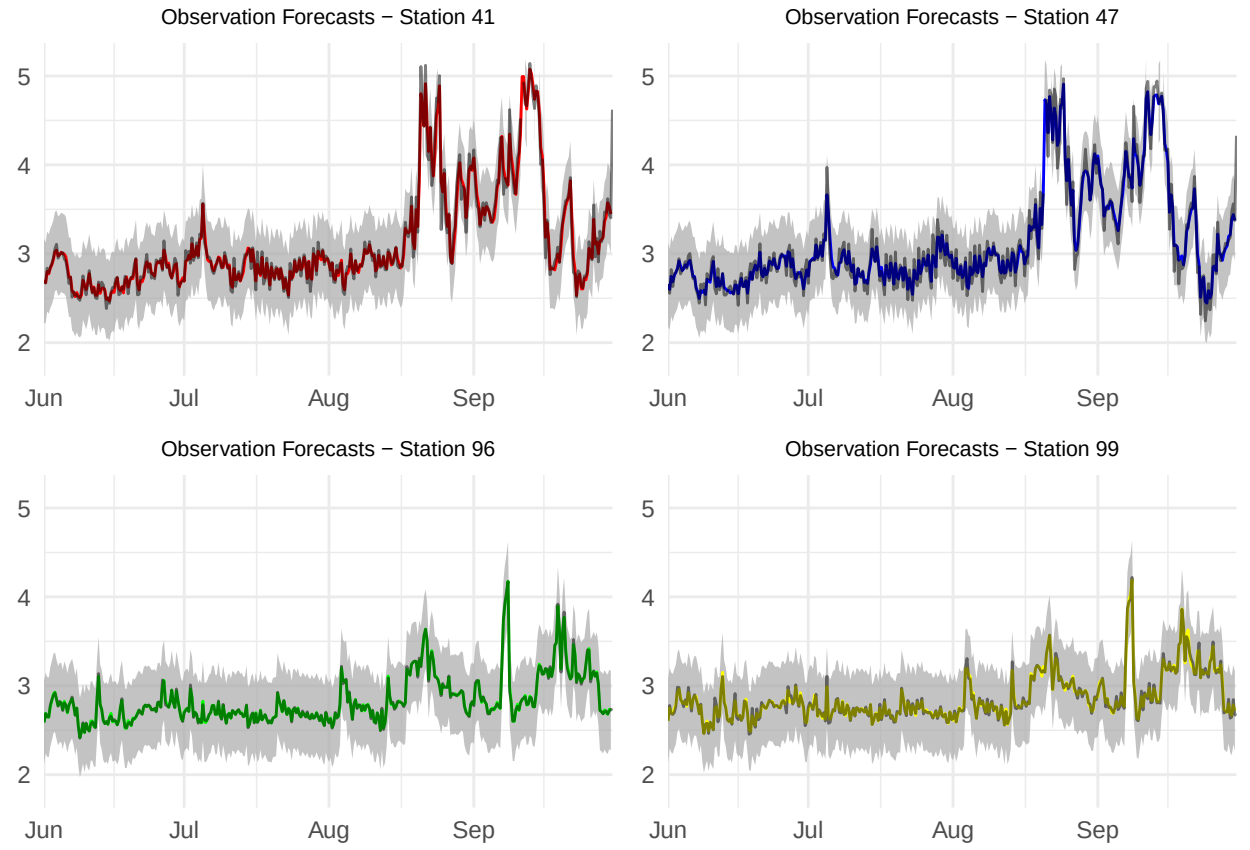

Figure 5: One-step Ahead Forecasts for 4-Stations

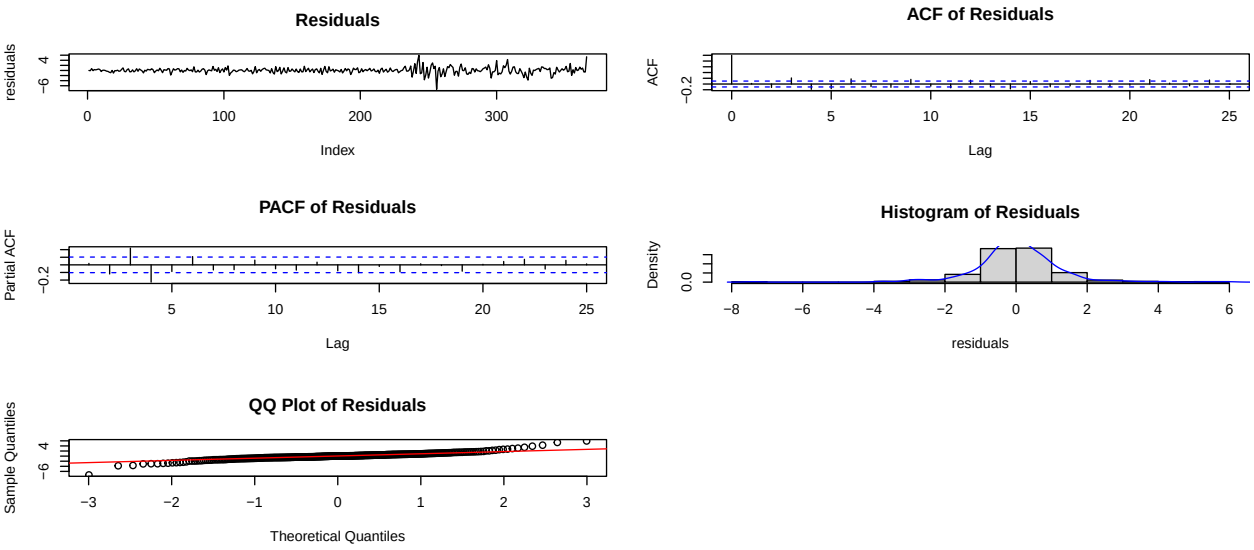

Figure 6: Diagnostic Plots for Station 41: Spatial DLM

Dynamic Linear Model (DLM) operates under the assumptions of normality and linearity in distributions, and it requires the initial distribution to be independent of the errors. We have fitted two kind of DLMs one univariate and one multivariate including the spatial dependence between stations. What makes DLMs quite useful is the fact that it can be used online, as data comes continuously.Analysis of residuals suggests improvements can be still done to this model.