POLITECNICO
MILANO 1863

BAYESIAN SPATIOTEMPORAL MODELS FOR $PM_{2.5}$ IN THE PO VALLEY

Bayesian Statistics course

Lorenzo Angiolini, Alessandro Benelli, Giulia Bergonzoli,
Greta Camplese, India Ermacora, Lucia Gregorini

Tutors: Alessandra Guglielmi, Michela Frigeri
Academic year 2022-2023

Contents

1 Introduction

The term $PM_{2.5}$ identifies all the pollutant particles, both solid and liquid, partly emitted directly from sources into the atmosphere and partly formed through chemical reactions between other polluting species, with an aerodynamic diameter of less than or equal to 2.5 μm. These particles are considered dangerous to health because they are characterized by a long permanence time in atmosphere and are able to penetrate deeper into the human respiratory tract.
We have focused our analysis on the Po valley, where about 40% of the Italian population lives: over 23 million people who together produce more than 50% of the national GDP. The Po Valley is indeed one of the most populated and industrialized places in Europe, leading to high levels of pollutant emissions. However, air quality is particularly critical in this area also due to the orographic and weatherclimatic conditions of the Po Valley. In fact the plain is surrounded on all sides by mountains (the Alps and the Appennines), except to the east where it overlooks the Adriatic; the area is also characterized by atmospheric stability which favors the increase in concentrations of pollutants and makes their dispersion difficult and slow.
The goal of our project is to obtain spatio-temporal models for $PM_{2.5}$ emissions collected in different detection stations located in the Po Valley, taking into account the context characteristics of each station.

The Dataset Our dataset includes $PM_{2.5}$ daily concentration measurements collected in different stations over Lombardia, Emilia-Romagna, Piemonte and Veneto from 2014 to 2020. It also includes some characteristics of the stations themselves.
In particular for every observation we considered:

- **Date**: date in which the observation was registered (day-month-year)
- **Station's name**: we have 32 stations in Lombardia, 30 in Emilia Romagna, 6 in Piemonte and 20 in Veneto
- **Area**: division of the stations into *Urban*, *Suburban* or *Rural* area
- **Type**: division of the stations into *Traffic*, *Industrial* or *Background* type
- **Altitude**: altitude of the station where the $PM_{2.5}$ is measured
- **Latitude and Longitude**: coordinates which describe the precise location of the station

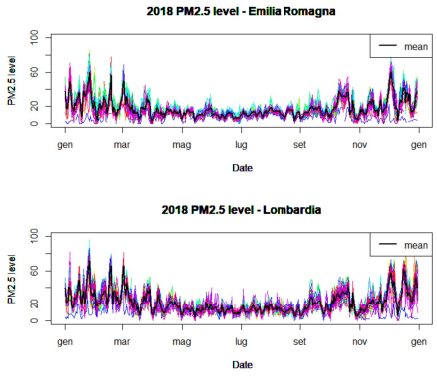
In our analysis we focus on data of 2018 related to the 62 stations located in Emilia Romagna and Lombardia, ignoring those of Piemonte and Veneto, since during the collection period the majority of the stations in these two regions were not active, having many days in which data were not gathered. We did not replace the remaining NA values in order to avoid modifying the original data; they will be dealt with the Bayesian approach later on. Finally, since we have noticed a certain periodicity between different years, with our final model we will be able to extend our results to other years.

2 Exploratory data analysis

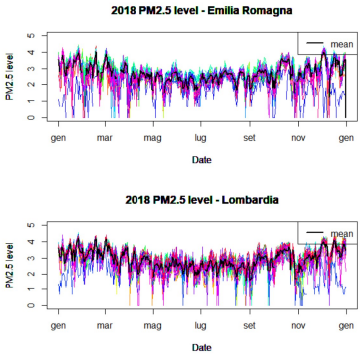
We begin our inspection with an exploratory data analysis trough a frequentist approach, taking care of both the *temporal* (distribution over the year) and *spatial* (different locations of all the stations) information.

2.1 Temporal dimension

From the plotted time series of the levels of PM2.5 of all the stations we can notice a clear U-shape due to seasonality. In fact, in the coldest months the levels of pollution are higher than in the warmer ones, being extremely influenced by domestic heating and vehicular traffic in towns, and there is also a higher variability as we can see by the steeper peaks. Moreover we can notice a strong correlation between the different stations, since they follow the same shape.
Additionally, as is common in literature, we log-tranformed our data in order to reach a more symmetric and gaussian-like behaviour of the observations distributions, also taking into account the fact that pollution levels are strictly positive.



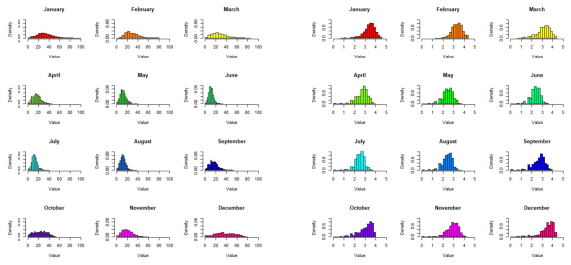
(a) PM2.5 data in Emilia Romagna and Lombardia



(b) log-PM2.5 data in Emilia Romagna and Lombardia

Figure 1

Studying the evolution of pollutant levels over months for the two considered regions, we plotted the histograms of the joint monthly distributions of PM2.5 levels in figure 2. There is a higher variability during the colder months, while in the warmer ones measurements are distributed around low values. We assume that this particular trend is due to the shutting down of domestic heating and the less vehicular traffic in towns. Obviously these differences are attenuated in the log-transformed data.
We then conclude that seasonality plays a key role in the definition of the pollution level and it will be an important aspect to be taken into account in defining our models.



(a) Lombardia (b) Lombardia

Figure 2: Histograms of levels of PM2.5 for each month

2.2 Spatial dimension

From the territorial point of view, data is partitioned in three different areas: *Rural*, *Suburban* and *Urban*.

- **Rural**: station located in a mostly not urbanized area
- **Suburban**: station located in a zone in which both built-up and not urbanized areas are present
- **Urban**: station located in a predominantly built area

We report in figure 4 the graphical overview of the PM2.5 time series of both Lombardia and Emilia Romagna and their means, dividing the stations by area. It seems evident that there are no big differences between the three, we will investigate if our hypothesis are true in our future models.

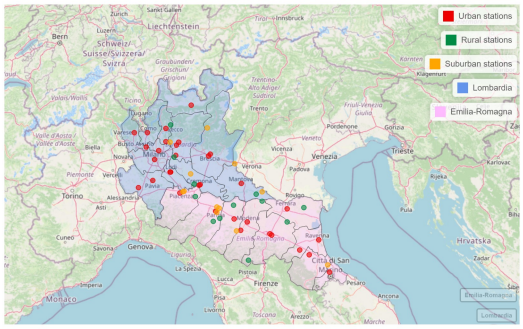


Figure 3: Localization of stations divided by area

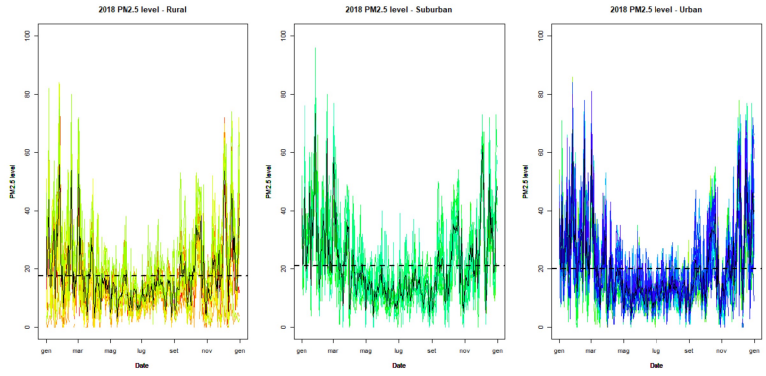


Figure 4: Concentration of PM2.5 in Rural, Suburban and Urban areas

Another partition is made over the data by zones. We have three types of zones: *Background*, *Industrial* and *Traffic*.

- **Background**: stations located in areas for which it cannot be assumed that the air pollution level is mainly due to specific sources
- **Industrial**: stations located close to industrial areas
- **Traffic**: stations located in areas in which the air pollution is mainly due to vehicle traffic emissions

As before we reported in figure 6 the graphical overview of the PM2.5 time series divided by type: the results are quite similar. Finally, we will consider the altitude of the different stations since we expect that higher

values of pollutant will be found at lower height and vice versa. Indeed the main sources of PM2.5 pollution are connected to the human presence, such as transport emissions and domestic heating, which is lower at high altitudes.

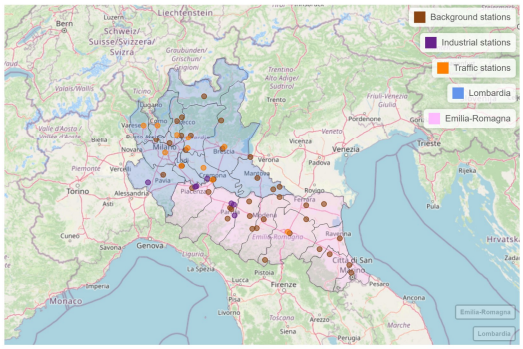


Figure 5: Localization of stations divided by type

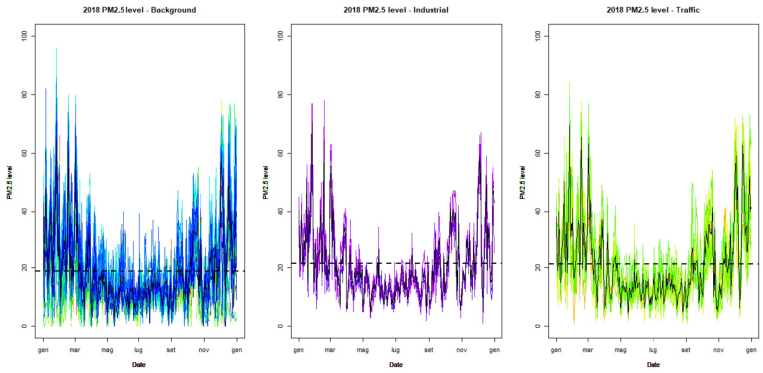


Figure 6: Concentration of PM2.5 in Traffic, Industrial and Background types

3 Spatio-Temporal model

In order to pursue the goal of describing the concentration of $PM_{2.5}$ we will make use of a Bayesian spatio-temporal model. This will allow us to monitor the level measured in the 62 stations of Emilia-Romagna and Lombardia exploiting their spatial correlation. The time information will also play a key role in estimating $PM_{2.5}$ throughout 2018. The formulation of the model is the following

$$Y_i(t)|parameters \sim \mathcal{N}(\mu_i(t), \sigma^2) \tag{1}$$

$$\mu_i(t) = f(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \tag{2}$$

$$i = 1, \dots, 62 \qquad t = 1, \dots, 365$$

where

- $Y_i(t)$: log-transformed $PM_{2.5}$ values
- $f(t)$: Temporal mean process
- $\mathbf{x}_i$: Covariates associated to each station
- w_i: Spatial residuals

We will dedicate a separate section for each of these components in order to better explain their meaning.

3.1 Modelling $f(t)$

To construct the backbone of the model we resort to two popular techniques in the context of time series analysis. The first one is smoothing through a truncated Fourier expansion, able to capture the seasonality of the data and tackle periodicity in an ad hoc way. The other approach we will try is ARIMA, predicting the current value based on the previous ones and their relative errors. Our work consists in estimating the model specific parameters in a Bayesian framework, instead than a more classical OLS or frequentist fashion. Moreover, this setting also allows to take into account information as the spatial correlation and the characteristics of each station.

3.1.1 Fourier Basis

The general formula we will make use of is the following

$$f(t) = \sum_{j \in J} a_j \, sin(j\omega t) + b_j \, cos(j\omega t) \tag{3}$$

where $\omega = \frac{2\pi}{T}$ with $T = 365$ days.
Note that the constant term c which is usually added in the expansion is not present. The reason will be clear later. Also, the set J was manually chosen in order to select the basis with a seasonality that had a relevant meaning. In general one could simply set $J = \{1, 2, 3, \dots, N\}$, and the periodicity of the corresponding harmonic function will be determined by the formula

$$\text{Periodicity} = \frac{2\pi}{\omega j} = \frac{T}{j}$$

Instead of truncating this expansion to an arbitrary N we picked out the terms we thought have a relevant contribution in estimating the shape of the curve: in particular

$$J = \{1, 2, 4\}$$

where $j = 1$ plays the most important role, modeling the annual periodicity, while $j = 2$ and $j = 4$ help in estimating the seasonal behaviour.

3.1.2 ARIMA

General form of an ARIMA(p, d, q) model:

$$\Phi(B)(1-B)^d \mathbf{y}_t = c + \Theta(B)\mathbf{w}_t$$

where

- $\mathbf{y}_t$ is the time series
- $\mathbf{w}_t$ is the white noise: $\mathbf{w}_t \sim N(0, \sigma_w^2)$
- B is the backshift operator, defined as: $B\mathbf{y}_t = \mathbf{y}_{t-1}$
- $\Phi(B)$ is the autoregressive operator: $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p$
- $\Theta(B)$ is the moving average operator: $\Theta(B) = 1 + \Theta_1 B + \Theta_2 B^2 + ... + \Theta_q B^q$

The final model we chose in an ARIMA(3,1,2), where p, q and d were picked by analyzing the mean time series and so, proceeding in this way, we were able to generalize and adopt a single model that could describe all the 62 time series. Another generalization attempt was to take p and q as the maximum values found by fitting an independent ARIMA model for each station, but since in this way we we should have introduced 2 extra parameters we have decided to keep the model as simple as possible in view of future unfoldments. Also in this case, similarly to what already said in the Fourier's part, we take c=0, for a reason that will be explained later.

3.2 Covariates

We have 7 covariates, of which only one can be thought of as continuous (Altitude), while the others are all categorical. In particular, we have the distinction of the Areas in Urban, Suburban and Rural, and the distinction of the Types in Background, Traffic and Industrial. For each of these two we associated n-1 categorical variables, therefore the resulting β are 5, plus the constant β_0. We decided to keep all the available covariates at first and discard the ones resulting not significant with posterior Confidence Intervals. Altitude has been standardized. In details

Constant	Altitude	Urban	Suburban	Background	Traffic
β_0	β_1	β_2	β_3	β_4	β_5

We report two examples to clarify how to obtain the expression for each station:

$$\mu_i(t) = f(t) + \beta_0 + \beta_1 \cdot \text{Altitude}_i + \beta_2 + \beta_5 + w_i \qquad \text{Urban Traffic station}$$
$$\mu_i(t) = f(t) + \beta_0 + \beta_1 \cdot \text{Altitude}_i + \beta_4 + w_i \qquad \text{Rural Background station}$$

The constant term will be the only one we add in the analysis: for this reason we removed the c term in both the Fourier and ARIMA expansion. Indeed, it is not necessary to include different constant terms as in the end we would estimate only the sum of those.

3.3 Spatial residuals

This term is used to model the spatial correlation between the station, as it represents a location-specific term which is constant over time. In particular we will make use of a Matèrn kernel to model the covariance function in a GP framework. In particular

$$\mathbf{w} \sim \mathcal{N}(0, \Sigma) \tag{4}$$
$$\Sigma_{i,j} = \alpha^2 \, exp(-\phi \, ||s_i, s_j||) \tag{5}$$

where ϕ is a length-scale parameter, α controls the smoothness of the function and s_i is the location of the i-th station. The idea is to control the covariance of two stations using only the distance between them. For this reason the process is stationary. Moreover, we will make use of the Euclidean distance in UTM coordinates, resulting in an isotropic process. In other words, if the distance between two stations is high, the resulting exponential will be close to 0 modeling little to no interaction. On the other hand, a small distance would result in a greater corresponding covariance element. The parameter ϕ can be set to a constant value, for example $3/d_0$ where d_0 is the maximum distance between two stations (in our case $3/d_0 \approx 0.008$). Better, we can specify a prior for it. Note that the unit measure for this parameter is km^{-1}.

3.4 Specifying priors and completing the model

The model is to be completed by specifying priors on the parameters of the model. For the regression part we used a $\mathcal{N}(0,1)$ for which more or less all the values between -4 and 4 are covered: considering the range of the data this is more than enough. The same is to be said for the Fourier and ARIMA specific parameters. The variance σ^2 is modeled as an $InvGamma(3,2)$ as it is common in the literature; for the same reason, an InvGamma was also chosen for the shape parameter α. For the scale-length parameter we had to use a different approach in order to avoid problems of identifiability: indeed using a rather uninformative prior, like a $U(0,1)$, caused divergent iterations in the procedure. For this reason we set $\phi = 0.008$ in the first runs and then we tried with very informative priors to run the code. In particular, we used Beta distributions with a mean close to the value of 0.1 and a very small variance. In the end, in the Fourier approach a $Beta(7,70)$ was enough to get a good estimate, while we kept ϕ constant in the ARIMA model. Summing up

$$
\begin{gathered}
Y_i(t)|parameters \overset{ind}{\sim} \mathcal{N}(\mu_i(t), \sigma^2) \\
\mu_i(t) = f(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \\
\mathbf{w} \sim \mathcal{N}(0, \Sigma)
\end{gathered}
$$

$$
\begin{gathered}
\sigma^2 \sim InvGamma(3,2) \\
\beta_0, \beta_1, \ldots, \beta_5 \overset{iid}{\sim} \mathcal{N}(0,1) \\
\Sigma = \alpha^2 \exp(-\phi \, ||s_i, s_j||) \\
\alpha^2 \sim InvGamma(3,2) \\
\phi \sim Beta(7,70)
\end{gathered}
$$

Fourier:

$$
\begin{gathered}
f(t) = \sum_{j \in J} a_j \, sin(j\omega t) + b_j \, cos(j\omega t) \\
a_j, b_j \overset{iid}{\sim} \mathcal{N}(0,1)
\end{gathered}
$$

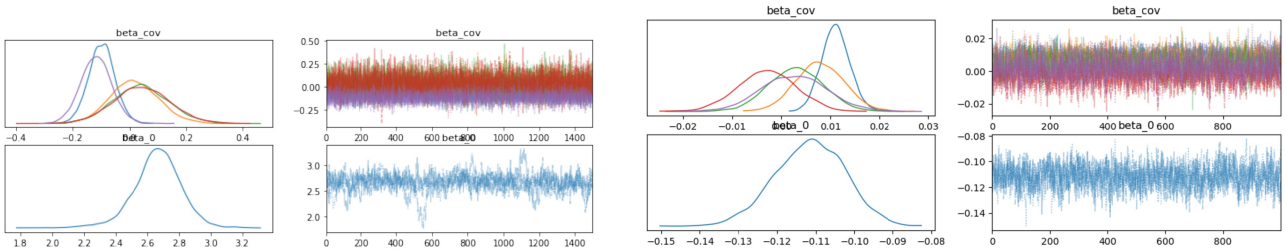
ARIMA:

$$
\begin{gathered}
f_i(t) = (1+\phi_1) y_i(t-1) + (\phi_2 - \phi_1) y_i(t-2) + (\phi_3 - \phi_2) y_i(t-3) - \phi_3 * y_i(t-4) + \\
+ \theta_1 \epsilon_i(t-1) + \theta_2 \epsilon_i(t-2) + \mathbf{x}_i^T \boldsymbol{\beta} \\
\epsilon_i = y_i(t) - f_i(t) \\
\phi_1, \phi_2, \phi_3, \theta_1, \theta_2 \overset{iid}{\sim} \mathcal{N}(0,1)
\end{gathered}
$$

4 STAN

First of all we look at the quality of the chains we set up. All of the distributions we reached seem to be stationary, therefore the chains have reached convergence. The only exception might be in the β_0 for the Fourier model, which has a more spread out distribution. Let us look at the common part of the two approaches and then delve into details for the peculiarities.

4.1 Covariates



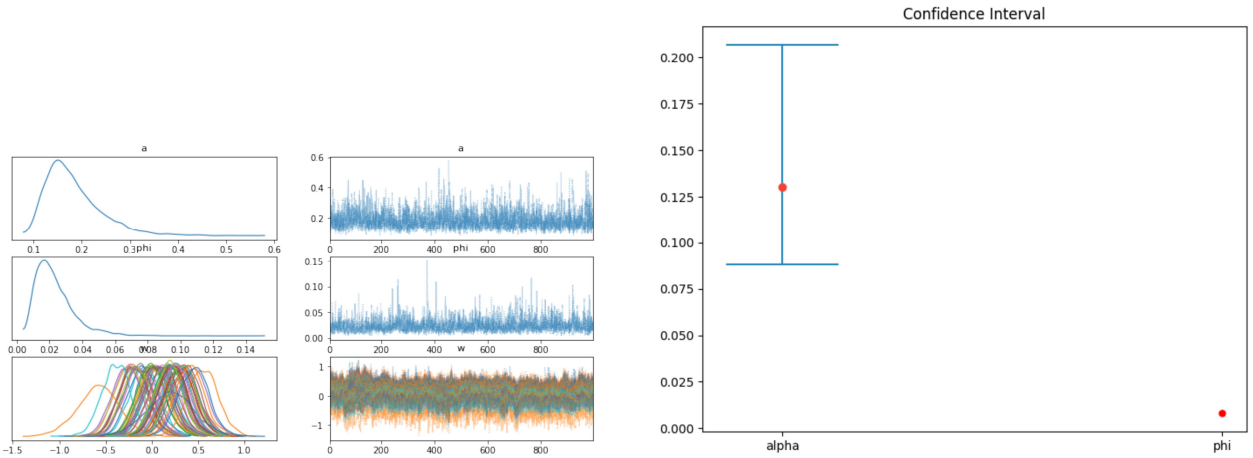
(a) Fourier: covariates

(b) ARIMA: covariates

Figure 7: Covariates

Both the models show that the covariates are well distributed around their mean, with a Gaussian-like density. In the Fourier model, the β_1 associated to altitude is negative, coherent with what we expect: indeed usually pollution in high altitude is lower. Moreover, as shown in figure 7a, β_0 is centered around the value 2.7 which is close to the temporal mean log-pollution level. On the other hand, the results regarding the ARIMA model are more difficult to interpret, since the biggest contribution to the value related to a certain station at each time t is given by a linear combination of the previous values and errors of the same station, for which covariates like altitude, area and zone remain obviously the same. For this reason the only slightly significant β are β_1 (altitude) and β_2 (the dummy associated with the urban area), in addition to the intercept $beta_0$. The other β associated with the categorical instead do not seem significant, or at least do not contribute very much. To provide evidence to this intuition we computed posterior credibility intervals in section 4.4 to determine which coefficients are significant, and eventually remove the others from the analysis.

4.2 Spatial residuals



(a) Fourier: spatial part

(b) ARIMA: spatial part

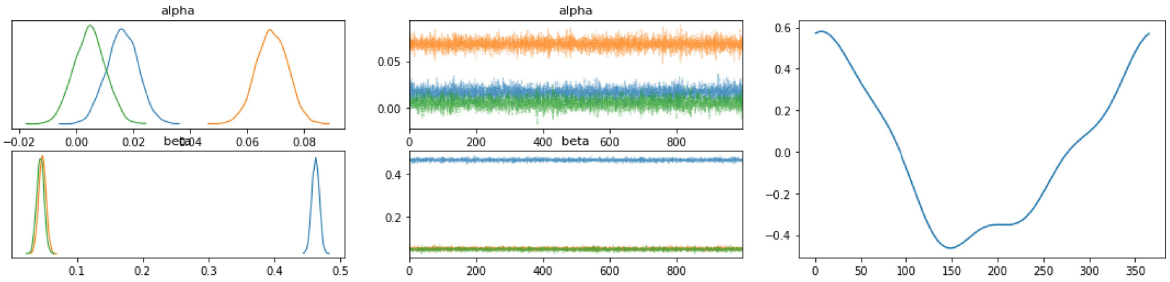
Figure 8: Spatial residuals

All the coefficients reach stationarity; in the Fourier case, we are able to effectively estimate both the smoothness parameter and the scale-length. A good estimate for ϕ, the posterior median, is around 0.018. For these parameters the distribution is not symmetric, since it has a heavier tail on the right. The spatial residuals are well distributed around their mean, showing a Gaussian-like posterior distribution. They position around the 0, with a good spread, covering values between -1.0 and 1.0. The interpretation is clear: the time series associated to each station will move up or down according to each of these values. Bear in mind that when smoothed with the logarithm the $PM_{2.5}$ values are between 1 and 5, so a factor of 0.5 is very significant. In the ARIMA approach, choosing both ϕ and w randomly led to problems of identifiability, and for this reason we decided to keep ϕ constant. The value was chosen to be 0.008, which is equal to 3 over the maximum distance between all the stations, as suggested by the literature. As for the spatial residuals we can make a similar reasoning, with the only exception that they are centered around 0.8 and cover values in the range 0.6 - 1.

4.3 Coefficients

Now, we will proceed to analyze the coefficients of both models individually.

4.3.1 Fourier coefficients



(a) Fourier coefficients: blue(annual), orange(half-yearly), green(seasonal)

(b) Fourier smoothing

Figure 9: Fourier

All of the coefficients reach stationarity and are well distributed around their mean, reaching a Gaussian-like behaviour (figure 9a). The annual cosine is the one whose associated coefficient is by far greater, indeed it is the one explaining the U-shape of the data. Again, we will compute posterior credibility intervals to determine if all these coefficients are significant. In figure 9b we can see the resulting smoothing using posterior mean as pointwise estimator for the coefficients: we are quite satisfied by the resulting shape, since it well captures the temporal annual trend of the pollution level. The point estimates for the coefficients (posterior means) are:

- $\alpha_1 = 0.018$
- $\alpha_2 = 0.067$
- $\alpha_3 = 0.0041$
- $\beta_1 = 0.47$
- $\beta_2 = 0.0048$
- $\beta_3 = 0.0046$

4.3.2 ARIMA coefficients

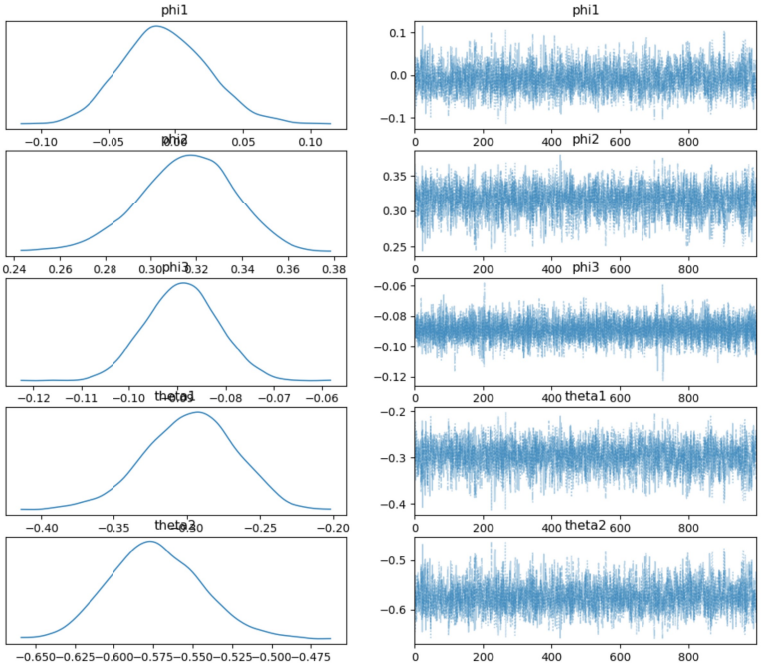


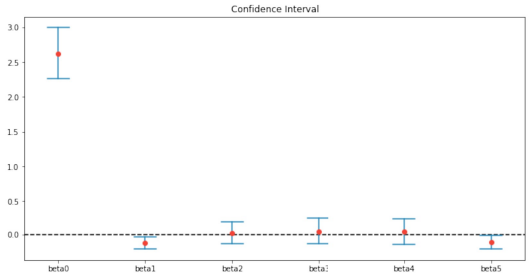
Figure 10: ARIMA coefficients: *phi*1, *phi*2, *phi*3, *theta*1, *theta*2

Also in this case all the coefficients reach stationarity, and moreover they are within the range (-1,1), even though it was not imposed as a constraint in the model, confirming in this way the stationarity of the time series. We obtained the following median values for each distribution:

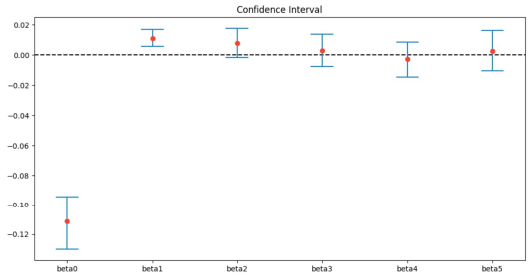
- ϕ_1: -0.01
- ϕ_2: 0.31
- ϕ_3: -0.09
- θ_1: -0.29
- θ_2: -0.57

In general the magnitude of the coefficients provides information about the strength and the direction of the relationships between the variables, so we see that there is a positive relationship between the current value differenced time series and its second lag, while there is a negative relationship with its third lag. *Phi*1 is not significant (as we will better exploit in the next session) but, keeping in mind that we are dealing with the differenced time series, this does not mean that y_t is not influenced by y_{t-1}, but the reasoning is applied to $y_t - y_{t-1}$ and $y_{t-1} - y_{t-2}$. Finally, there is a strong negative relation between the current values of the differenced time series and its past errors: generally negative MA coefficients can indicate that there is a corrective mechanism in the time series that brings it back towards its mean after a deviation from the mean.

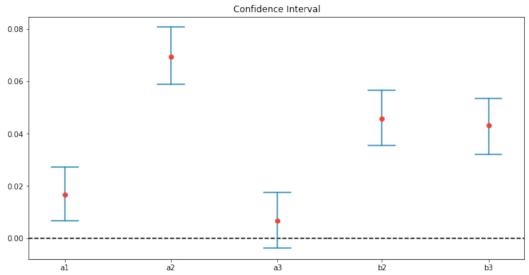
4.4 Posterior CI



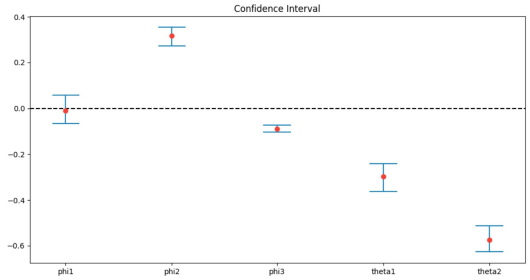
(a) Fourier: CI for the β of the covariates



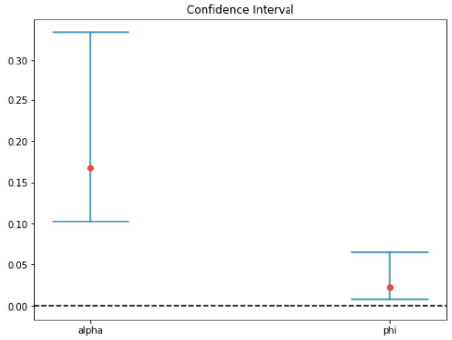
(b) ARIMA: CI for the β of the covariates



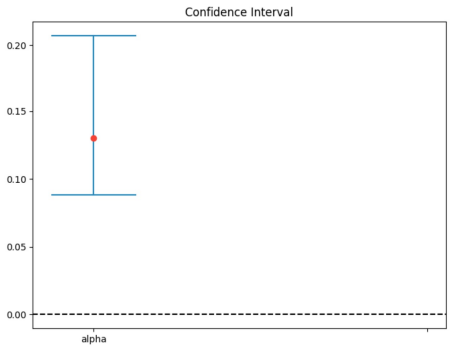
(c) CI for the Fourier coefficients (to have a better view the annual cosine coefficient is not plotted)



(d) CI for the ARIMA coefficients



(e) Fourier: CI for the spatial coefficients



(f) ARIMA: CI for the spatial coefficients

Figure 11: Posterior CI

To determine which coefficients are not significant we computed posterior credibility intervals using quantiles of the posterior distribution. If the 0 is well inside these intervals it means we can safely remove the coefficient from the analysis. This is indeed what happens for almost all the β associated to the categorical variables, except for the one associated to the Traffic stations in the Fourier model, and the one associated with the Industrial Stations in the ARIMA one. The same could be said for the seasonal coefficient associated to the sine function in the Fourier part and the first autoregressive coefficient in the ARIMA one, but in this case we decided to keep them for readability of the model.

5 Further improvements and model choice

The next step in our analysis is to improve as far as possible the models and in the end compare them using some GOF techniques in order to find the best one, or at least the best with the choices we made.

5.1 Combining the models: ARIMA with Fourier basis

A possible solution to incorporate the best characteristics of both models could be to add in the ARIMA model an annual cosine to the mean function μ in order to obtain a better shape of it. Indeed the cosine with yearly periodicity is the one causing the U-shape: in this sense we can retain the flexibility of the ARIMA model while still providing a robust mean to build upon. Therefore in formulas we have the following model:

$$
\begin{gathered}
Y_i(t)|parameters \overset{ind}{\sim} \mathcal{N}(\mu_i(t), \sigma^2) \\
\mu_i(t) = f(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \\
\mathbf{w} \sim \mathcal{N}(0, \Sigma)
\end{gathered}
$$

$$
\begin{gathered}
\sigma^2 \sim InvGamma(3, 2) \\
\beta_0, \beta_1, \ldots, \beta_5 \overset{iid}{\sim} \mathcal{N}(0, 1) \\
\Sigma = \alpha^2 \exp(-\phi \, ||s_i, s_j||) \\
\alpha^2 \sim InvGamma(3, 2)
\end{gathered}
$$

$$
\begin{gathered}
f_i(t) = (1 + \phi_1) y_i(t-1) + (\phi_2 - \phi_1) y_i(t-2) + (\phi_3 - \phi_2) y_i(t-3) - \phi_3 * y_i(t-4) + \\
+ \theta_1 \epsilon_i(t-1) + \theta_2 \epsilon_i(t-2) + \mathbf{x}_i^T \boldsymbol{\beta} + a \, cos(\omega t) \\
a \sim \mathcal{N}(0, 1) \\
\epsilon_i = y_i(t) - f_i(t) \\
\phi_1, \phi_2, \phi_3, \theta_1, \theta_2 \overset{iid}{\sim} \mathcal{N}(0, 1)
\end{gathered}
$$

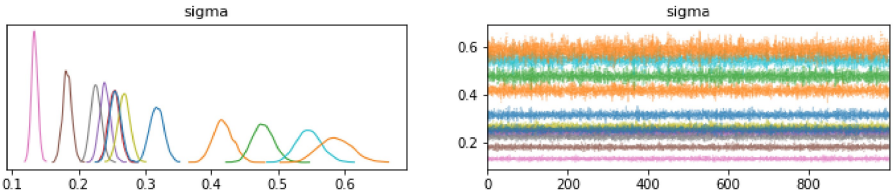
5.2 Monthly sigma

To improve both the models (the Fourier one and the Arima with the addition of the annual cosine one) we decided to impose a monthly σ rather than a single one, common throughout the all year. In formulas

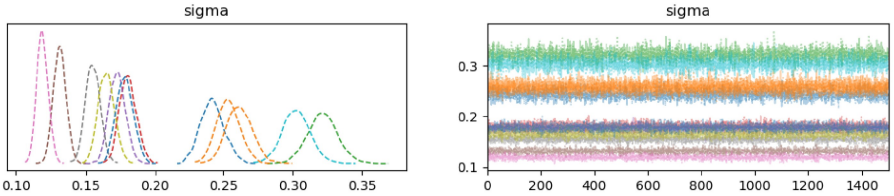
$$
\begin{gathered}
Y_i(t)|parameters \overset{ind}{\sim} \mathcal{N}(\mu_i(t), \boldsymbol{\sigma}^2_{\boldsymbol{m}}) \\
\mu_i(t) = f(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i
\end{gathered}
$$

$$
\sigma^2_m \sim InvGamma(3, 2) \quad m = 1, \ldots, 12
$$

This choice was driven by the exploratory data analysis, in which it became clear that the warmer month do not only experience lower $PM_{2.5}$ levels but also lower variability. Complete STAN results are shown in section 8.2 in the appendix. In figure 12a we report the estimated $\boldsymbol{\sigma}$



(a) Fourier: posterior distributions for $\boldsymbol{\sigma}$: in order July, June, August, May, April, January, September, November, February, March, October, December



(b) ARIMA: posterior distributions for $\boldsymbol{\sigma}$: in order July, June, August, September, May, April, January, November, December, February, October, March

Figure 12: Monthly sigma

The posteriors for both models are well distributed around their mean, reaching a Gaussian-like behaviour. Moreover the chains have successfully reached convergence. As expected in warmer months the estimated σ is lower than in the coldest months. This model allows for a better control of the error and the posterior credibility bands (see figure 21): the prediction will be narrower in summer while providing a broader area in winter.

5.3 Model choice

The first approach that came to our mind to evaluate the best model was to resort to classical predictive goodness of fit measures, like LOO and WAIC. They are both estimates of the log-pointwise predictive density. Smaller values indicates higher out-of-sample predictive fit ("better" model). Of course, all of these measures are based on the hypothesis of IID data, which we cannot assume in this case since we have a clear spatial correlation. However, they can still provide relevant information and we will report them here. The general formula is

$$elpd_{loo} = \sum_{i=1}^{n} log\, p(y_i|y_{-i}) \tag{6}$$

$$elpd_{waic} = \sum_{i=1}^{n} log\, p(y_i|y_{-i}) - \sum_{i=1}^{n} Var_{post}(log\, p(y_i|parameters)) \tag{7}$$

To overcome this problem, we came up with other more empirical error measures that do not rely on the IID hypothesis. To achieve this purpose, we ran again our code, this time leaving out the last 10 days to perform prediction and evaluate the error on those. Then we came up with three different measures:

$$Measure_1 = \frac{1}{62 \cdot 10} \sum_{j=1}^{62} \sum_{t=356}^{365} |y_j(t) - median(\hat{y}_j(t))| \tag{8}$$

$$Measure_2 = \sum_{j=1}^{62} \sum_{t=356}^{365} \frac{|y_j(t) - \mathbb{E}(\hat{y}_j(t))|}{y_j(t)} \tag{9}$$

$$Measure_3 = \frac{1}{62 \cdot 10} \sum_{j=1}^{62} \sum_{t=356}^{365} \mathbb{1}_A(y_j(t)) \quad \text{where} \quad A = CI_{95\%} \tag{10}$$

The first two measures are quite similar, as they rely on the posterior distribution to provide the prediction, with which we compute normalized measures. The differences are in using posterior mean or median (very close in our case since the posterior distribution are quite symmetric). The third one instead counts how many of the real values fall within the 95% posterior confidence interval. Since for the first two measures we are computing a prediction error the lower the value obtained the better, while the for the third one is the opposite, the best value is the closest to 1. Here are the results:

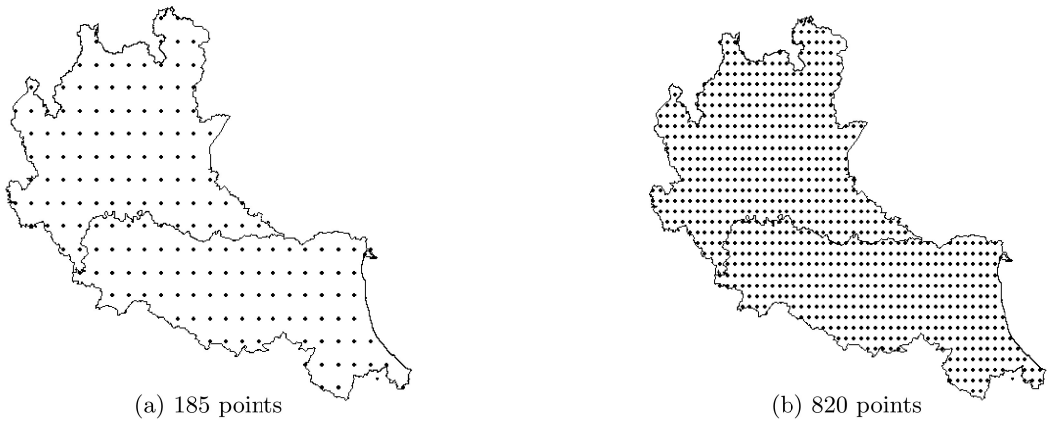
Model	LOO	WAIC	Measure 1	Measure 2	Measure 3
Basic Fourier	-18927.23	-19270.48	0.4213	99.2539	0.9709
Monthly Fourier	-18017.56	-18369.24	0.3882	101.7612	0.9854
Basic ARIMA	-14650.20	-14444.92	0.2781	58.9248	0.9838
ARIMA with cosine	-14638.57	-14437.04	0.2743	58.6859	0.9887
Monthly ARIMA with cosine	-14099.58	-13866.29	0.2749	58.9118	0.9919

Table 1: Comparison between models

As we can see from table 1 the best model according to all criteria is an ARIMA model. In particular, the one with monthly variance and cosine performs better in the LOO and WAIC indexes, as well as in the percentage of predictive points belonging to the posterior 95% CI. As for what concerns Measure 1 and 2 (which, as mentioned before, are very similar) the model with a unique σ seems to perform slightly better but, given the almost imperceptible difference, we classify as best model, according to these indexes, the ARIMA one containing both the cosine and monthly sigma. As we expect, each ARIMA model is much heavier to run in comparison to any Fourier model, as the estimates for the parameters are way slower. A further comment can be done on the predictions realized by the best models for the two approaches, as plotted in figure 21 in the Appendix. The models follow two opposite trends: in the Fourier case the variance is minimized while the bias is higher, while in the ARIMA case the bias is very small, paying with a high variance. Of course this is due to the differences in the models: when we smooth using Fourier basis we can well predict the general trend but the peaks are completely left out. With ARIMA we are able to estimate very well when the time series goes up or down but there is a greater risk of finding wrong predictions especially when going further in time. In the short period we chose to make prediction our suppositions are verified. To sum up, accordingly to the comparison indexes and the performance on the prediction interval, we choose the third ARIMA model as the best one in describing our data.

6 Kriging

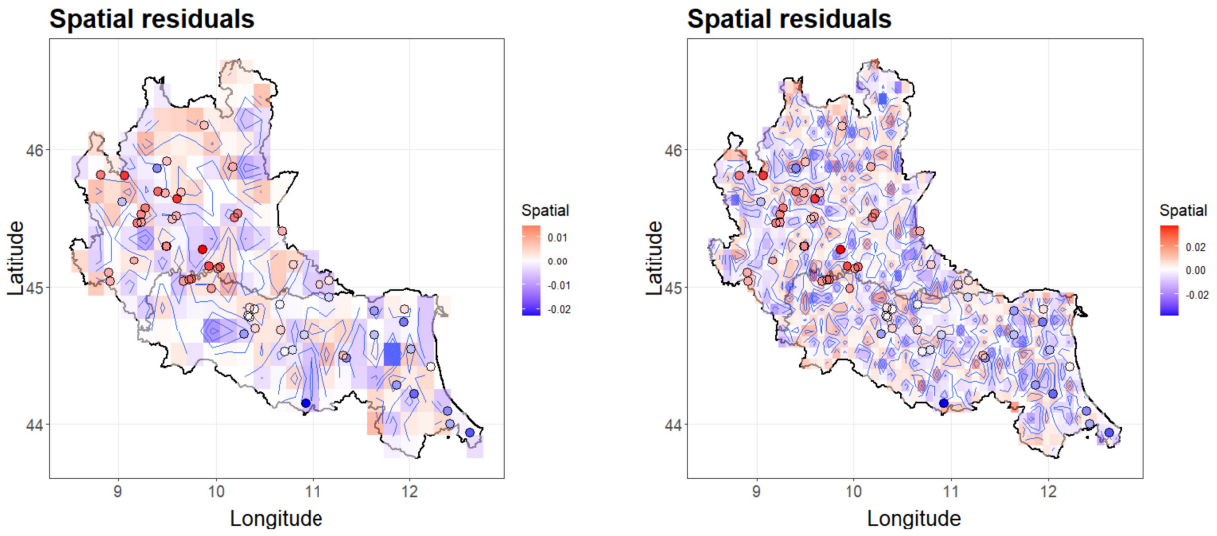
A possible extension to this model would be to spatially predict the pollutant level over the entirety of the Lombardia and Emilia Romagna map. Of course, we do not possess the covariate values for all points in the map, but we can try to predict the value of the spatial residuals based on the correlation model we have used until now. In particular, we would like to perform kriging on the spatial residuals in a GP fashion. We will use the 62x62 distance matrix estimated from the stations to perform out of sample prediction in the points provided to the model. For this reason, we created a grid of roughly 2000 points over the two regions, and we cropped it with the boundaries, obtaining 820 points, as it can be seen from figure 13b.



(a) 185 points (b) 820 points

Figure 13: Grid over Lombardia and Emilia-Romagna

Even if our final best model is the ARIMA one, containing both the cosine and monthly sigma, we decided to perform kriging on the Fourier one. Indeed, the spatial part is better modeled and identified in this second setting, where ϕ is not priorly fixed to a constant value and the obtained spatial residuals centered around zero (instead of being centered in 0.8 as in the ARIMA model). The goal is to predict on each of these 820 points the spatial residual based on the ones estimated on the stations' locations. The estimate we get is shown in figure 14b. In both settings we can notice that Lombardia is associated with positive values of spatial residuals in the stations, as opposed to Emilia Romagna, denoting higher pollution levels in the first region. Unfortunately the prediction is not very accurate, as almost all the estimates are close to 0 and therefore the plot is not very informative. We also tried to perform kriging on a less fine grid of 185 points since the observed values are only 62, each one associated to a single station. The result are quite similar and not informative (see figure 14a). The prediction is able to work best when many points are close together, for example in the area close to Milan or near the Adriatico sea. Instead where there are no stations the prediction is not reliable, outputting values close to 0. The main issue is the dimensionality of the problem: we have a few observations in comparison to the the number of points to be predicted. Morevoer, the presence of stations of different types and areas very close together creates difficulties in identifying recognizable patterns for prediction.



(a) 185 points

(b) 820 points

Figure 14: Kriging result

7 Univariate clustering

At this point of the analysis we use the models we have built to make clustering, which in this case means grouping the 62 stations in such a way that objects in the same group are more similar to each other than to those in other groups. To do it we implement a Finite Mixture Clustering on different parameters of the model in order to capture the best division between stations. The model for a general parameter z is:

$$z_i \overset{\text{iid}}{\sim} \sum_{k=1}^{C} \eta_k N(\mu_k, \sigma_k^2) \qquad i = 1, ..., N$$

$$\eta_k = v_k \prod_{i=1}^{k-1} (1 - v_i) \qquad k = 2, ..., C$$

$$\eta_1 = v_1$$

with:

- $\mu_k \sim N(0, 10\sigma_k)$
- $\sigma_k^2 \sim InvGamma(a, b)$
- $v_k \sim Beta(1, 2)$
- C=10

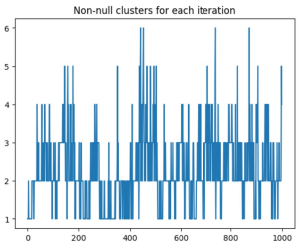
We simulate an MCMC chain with $M = 1000$ iterations and obtain a sample of clustering $c^{(1)}, \ldots, c^{(M)}$ which we summarize with a single cluster estimate c_hat. c_hat can be taken as clustering c* minimizing the posterior expectation of the Binder's loss function:

$$E(L(c, c*)|data) = \sum_{i<j} |\mathbb{1}_{c_i*=c_j*} - \pi_{ij}| \tag{11}$$

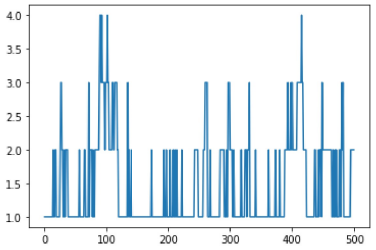
where π_{ij} is the elements at position (i,j) of the posterior similarity matrix, a 62×62 symmetric matrix that contains the pairwise probabilities that two observations belong to the same cluster. With Binder's loss function a loss of 1 is made whenever a pair of observations is treated differently in the estimated clustering c* than in the true c. The loss is thus the sum of disagreements in the treatment of pairs of observations between the estimated and true clustering. We did not want to limit the possible number of clusters, so we kept a high threshold, which has never been reached in any simulation.

7.1 Clustering on ARIMA(3,1,2) and AR(1) models

At first we made clustering on our best model, ARIMA (3,1,2) with $\sigma_k^2 \sim InvGamma(3, 0.001)$, and on its simplified version AR(1) with $\sigma_k^2 \sim InvGamma(4.5, 0.001)$. As for the first one we choose as clustering parameter Phi_2, since Phi_1 proved to be not significant at all, according on the posterior CI, and so it wouldn't make sense to use it to make a division among the stations, while for the second model we obviously used phi_1, which in this context is strongly significant. In both the cases (as we can see from figure 15) the number of non-null clusters varies at each iteration of the MCMC, meaning that the algorithm is working correctly.



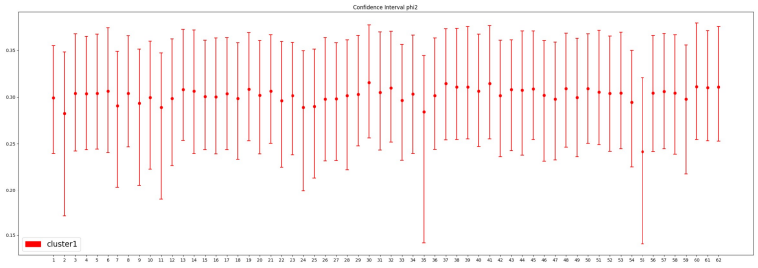
(a) ARIMA(3,1,2)



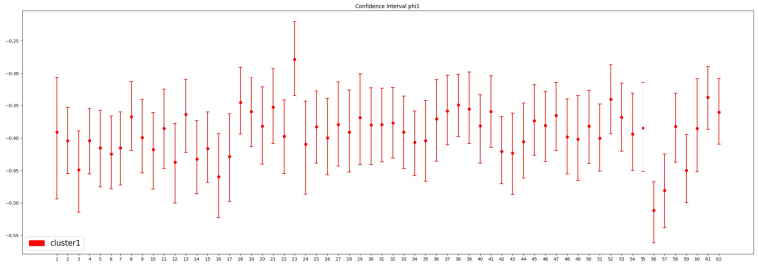
(b) AR(1)

Figure 15: number of non-null clusters found at each iteration of the MCMC for the two models

However, when selecting the best cluster through the proceeding explained before, in almost all the simulations all the points were placed in the same group. As we can see from figure 16 this result is actually reasonable since all the posterior 95% marginal posterior CI of both phi2 in the model ARIMA(3,1,2) and phi1 in the model AR(1) cover very similar range of values and therefore they should be classified as belonging to the same group.



(a) ARIMA(3,1,2)



(b) AR(1)

Figure 16: 95% marginal posterior CI of the parameters $phi2_i$ in the ARIMA model (top) and of $phi1_i$ in the AR model (bottom)

7.2 Clustering on Fourier model

We also present a clustering analysis on the other model we have analyzed, the Fourier one with $\sigma_k^2 \sim InvGamma(3, 0.003)$. In this case we took as discrimination parameter $b1$, the one related to the annual cosine in the Fourier expansion. Once again the selected number of clusters for each iteration of the MCMC varies between 2 and 7, but the most frequent value is 3, as we can see by the histogram in figure 17b.

(a) Plot of non-null clusters per each iteration

(b) Histogram of non-null clusters per each iteration

Figure 17: number of non-null clusters found at each iteration of the MCMC for the Fourier model

The number of estimated clusters which minimizes the posterior expectation of the Binder's loss function is 4. This value is plausible and the division seems to be reasonable if we look at the 95% marginal posterior confidence intervals. Indeed we can identify a big cluster which includes 92% of the values and three more small clusters which share the remaining percentage. In particular the second cluster includes the two stations with the highest values of b1, the third the only two stations with negative values of b1 and the last one is composed by only one station, which could actually be also classified as belonging to the biggest cluster since its values of the posterior CI are very close to the others. In figure 18 each of the 62 stations has been coloured according to depending on the cluster to which it belongs.

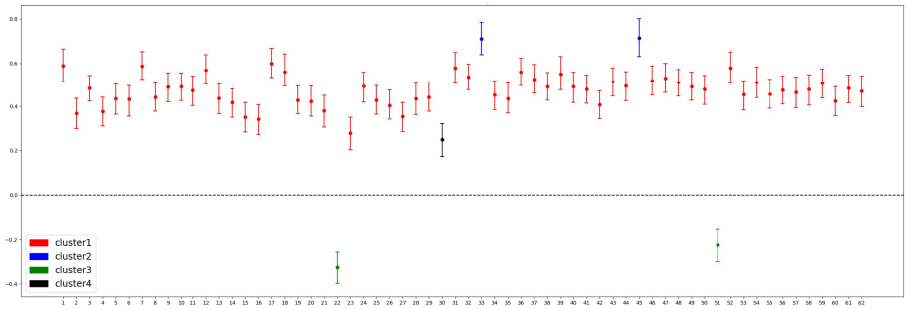


Figure 18: 95% marginal posterior CI of the parameters $b1$ in the Fourier model

We have plotted the time series of all the station, to check if different clusters correspond to different behaviours and we found consistent results. Indeed, as we can see from figure 19, the two stations belonging to the third cluster, which is the most differentiated from the others since it is the only one with negative values of the coefficient b, have the lowest values of $PM_{2.5}$. A same reasoning could be done for the second cluster: at high values of the coefficient $b1_i$ correspond high value of pollution, especially in the first part of the year. To be noted that the colors of the clusters differ from the previous one, just for visual reasons, while the label are of course the same. We have also plotted the stations divided in clusters in the map but here we can't see any particular behaviour.

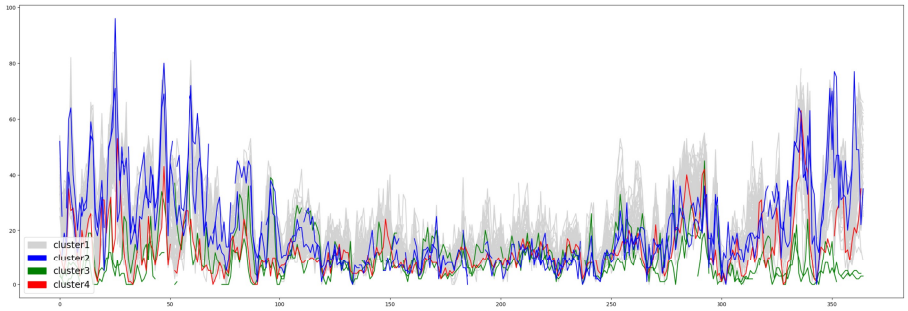


Figure 19: Plot of $PM_{2.5}$ particles for each of the 62 stations, coloured according to the respective cluster

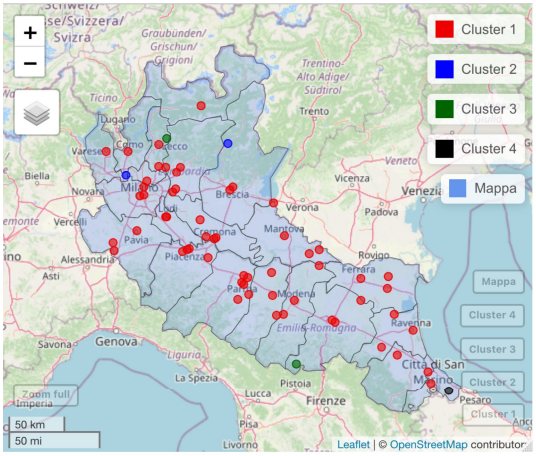
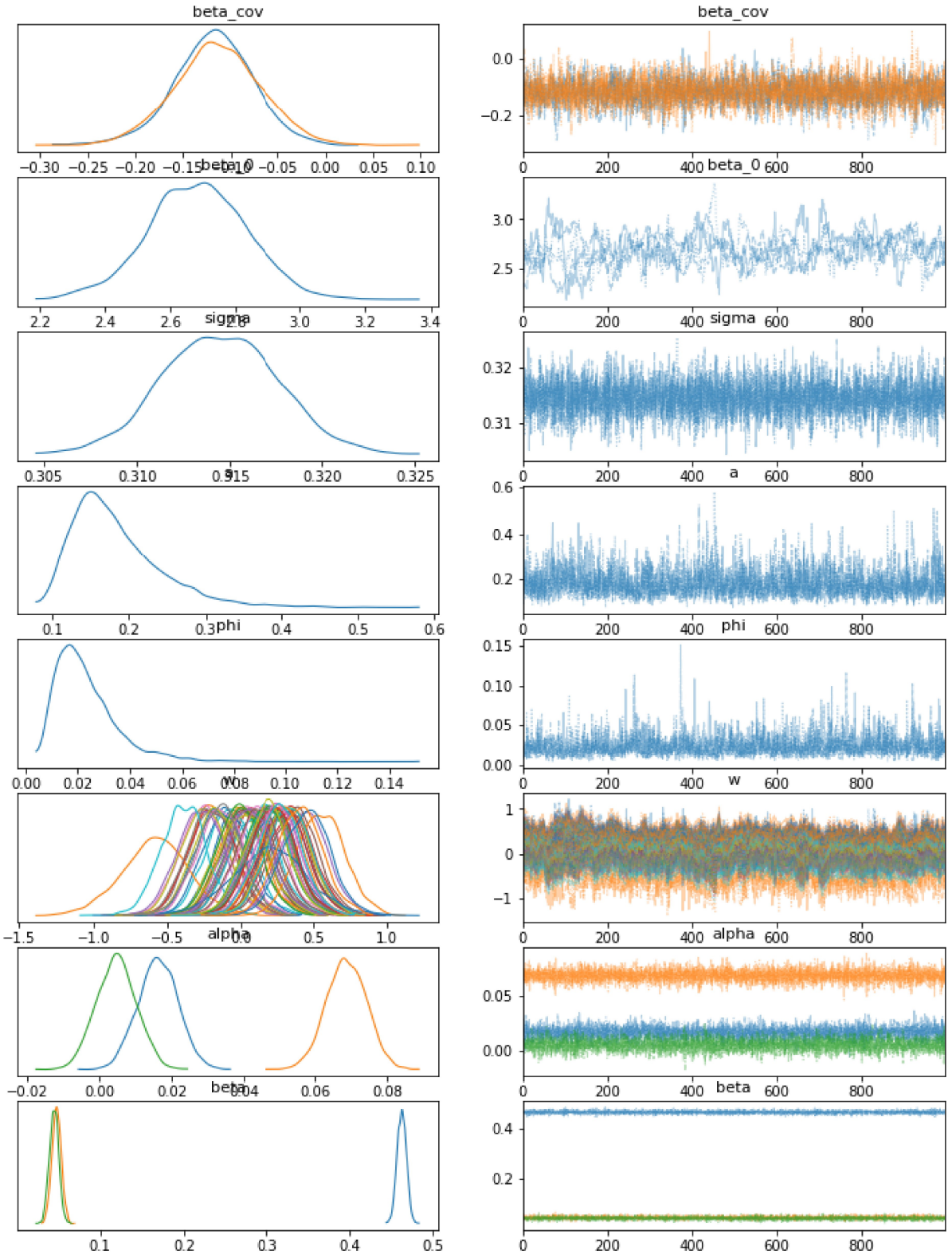


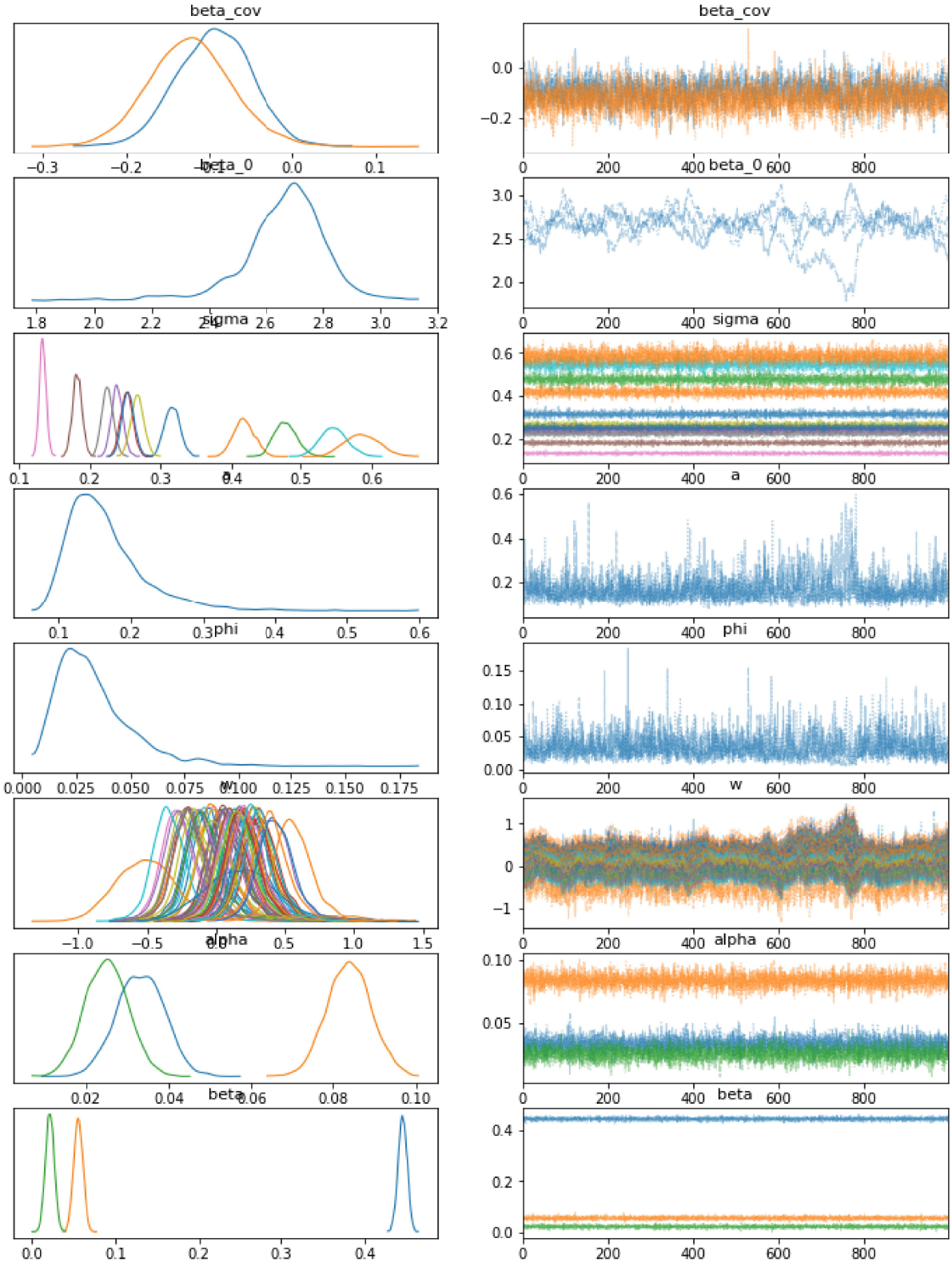
Figure 20: Plot of the clusters in the map

8 Appendix

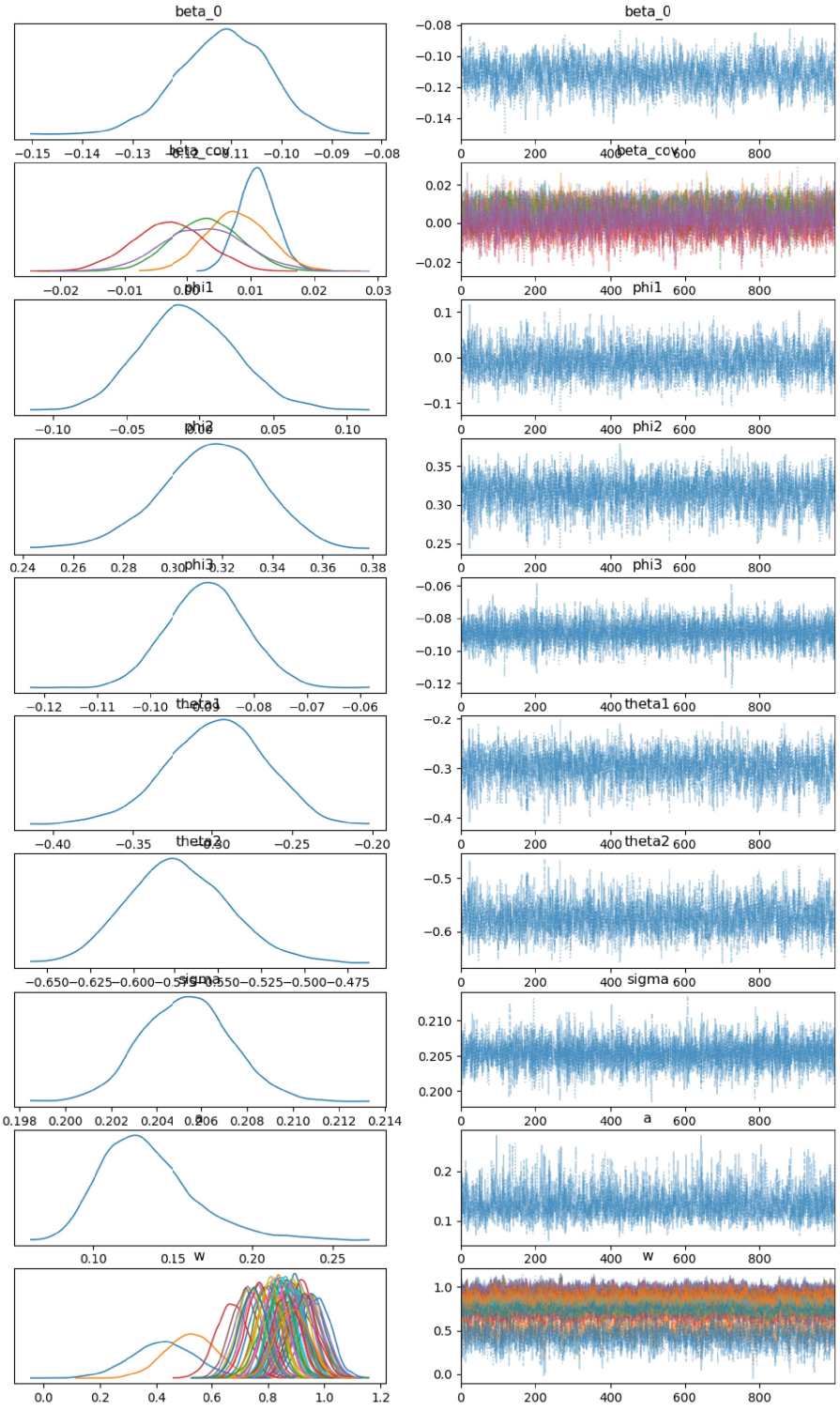
8.1 Fourier



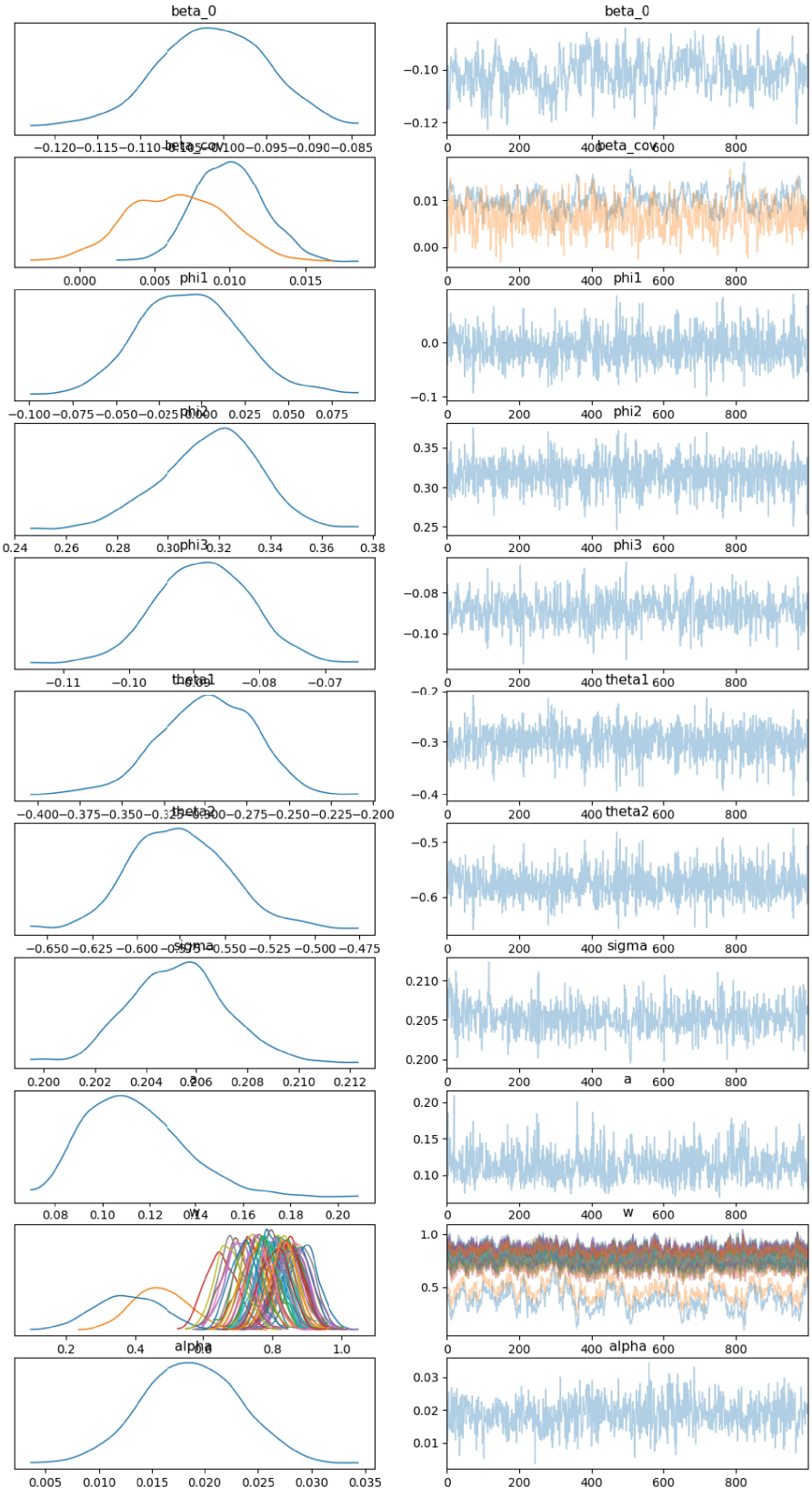
8.2 Fourier with monthly σ



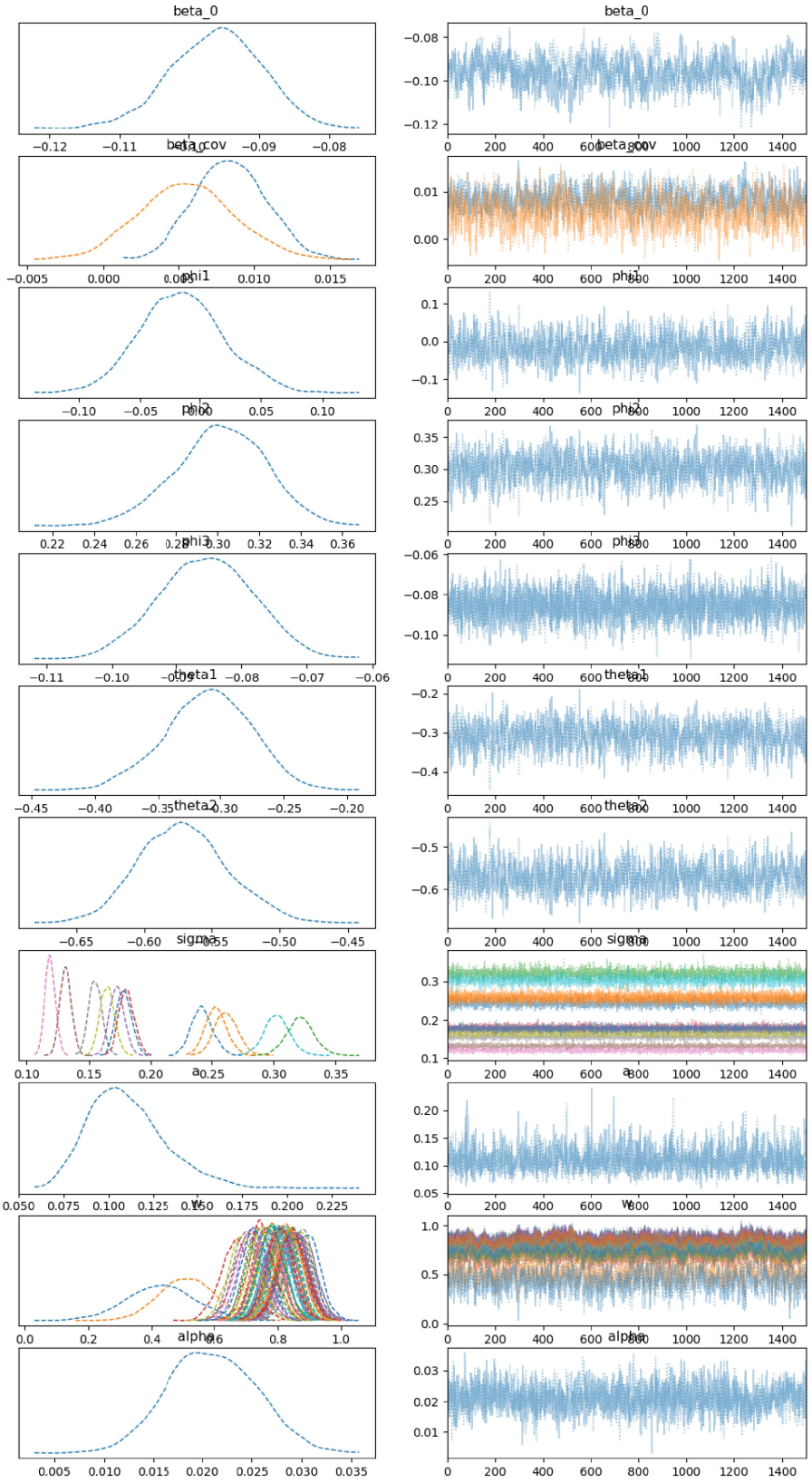
8.3 ARIMA



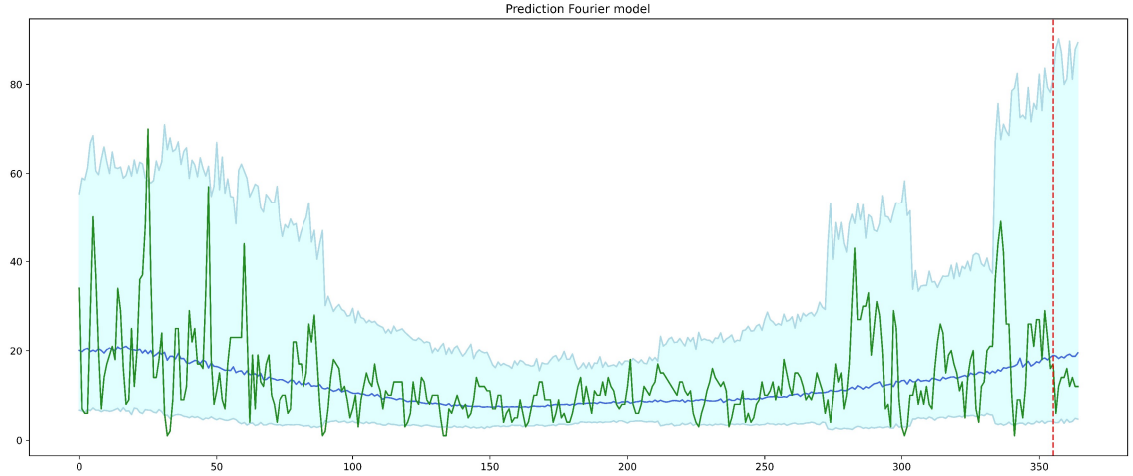
8.4 ARIMA with cosine



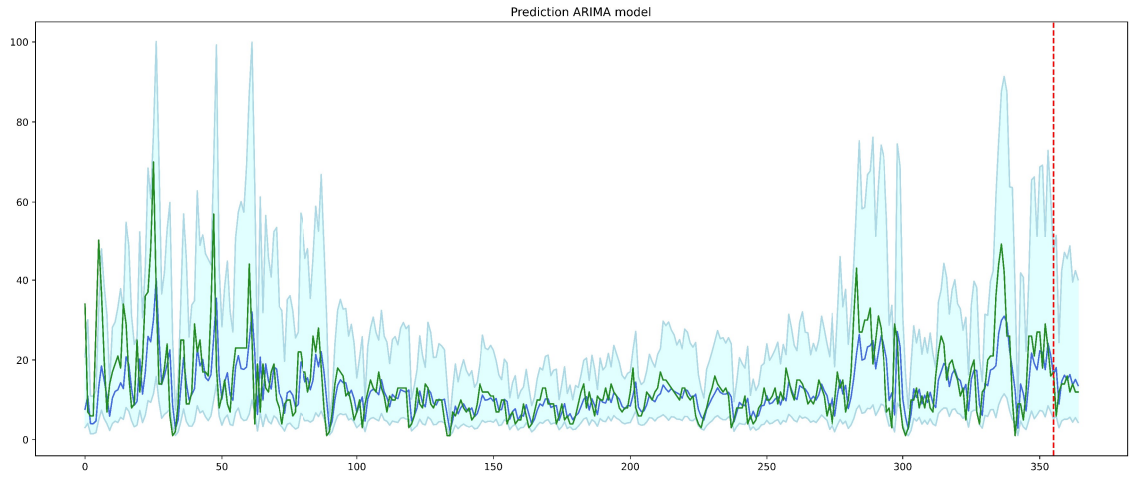
8.5 ARIMA with cosine and monthly sigma



8.6 Prediction



(a) Fourier prediction



(b) ARIMA prediction

Figure 21: Prediction in blue, CI in light blue and true data in green