Project report: Bayesian statistics
2022-2023

Bayesian spatio-temporal models for PM_{10} in the Po valley

Arrigoni Francesca, **Baracchi Federica**, **Cantalini Costanza**, **Gjyli Eno**, **Ferrara Stefano**, **Ursino Bruno**

Tutors: I.Epifani, M.Frigeri

MSc students in Statistical learning, Mathematical Engineering, Politecnico di Milano, Italia

Abstract

Abstract This project aims at building a complete Bayesian spatio-temporal model to better predict and understand PM_{10} pollution phenomena in Emilia Romagna. Our analysis focused on ARIMA models to treat particle concentrations in time, then enriched the model with geographical and topographical covariates, and lastly included spatial correlation effects among measuring stations by a Gaussian Process. Our project extends existing research on the topic and provides comparisons among competing strategies. Our findings support the commonly-held belief describing PM_{10} pollution as a phenomenon with short memory and scarce diffusion capability, peaking at emission sites and in short-term occurrences. We also found that regressive components and spatial dependencies compete in explaining such small-scale variability. Lastly, we provide evaluation on our models and suggest possible future developments.

Keywords: *Bayesian ARIMA; Bayesian Gaussian Process; Bayesian spatio-temporal model; Bayesian Time Structural Models; PM_{10}*

E-mail address:
francesca2.arrigoni@mail.polimi.it
federica.baracchi@mail.polimi.it
costanza.cantalini@mail.polimi.it
eno.gyjili@mail.polimi.it
stefano2.ferrara@mail.polimi.it
bruno.ursino@mail.polimi.it

Contents

1 Introduction

The Po Valley is one of the most polluted areas of Italy, especially when we consider air pollution. This is a consequence of the high level of industrialization in the area, but also its peculiar topography: low altitude and wide plains, hardly reached by long-distance winds. Our project aims at analyzing the concentration of PM_{10}, one of the most relevant air pollutants, in the Po Valley and in particular in the region of Emilia Romagna.
Airborne particulate matter (PM) is a complex mixture of particles that varies in size and chemical composition. For regulatory purposes, particles are defined by their diameter measured in microns. We are going to study PM_{10}, which has a diameter of fewer than 10 microns.

PM may be either directly emitted from sources or formed in the atmosphere through chemical reactions. The most commonly recognized sources are combustion processes, construction sites, wildfires, and urban traffic, responsible for the frequent scratching of tyres and consequent dust suspension.
PM_{10} particles are a health concern as they can be inhaled and deposited on the surfaces of the larger airways of the upper region of the lung inducing tissue damage and lung inflammation. Short-term exposure to PM_{10} has been associated primarily with the worsening of respiratory diseases leading to hospitalization and emergency department visits. The effects of long-term exposure to PM_{10} are less clear, although several studies suggest a link between long-term exposure and respiratory mortality.

For this reason, both the World Health Organization (WHO) and the European Environment Agency have set legal limit values for particulate matter concentrations on their territory.

Table 1. Limit values for PM_{10} given by Europe and Who

	Europe	WHO
Annual limit	$40\mu g/m^3$	$20\mu g/m^3$
Daily limit	$50\mu g/m^3$ max 35days/year	$50\mu g/m^3$ max 3days/year

Our first goal is to model the concentration of PM_{10} in the Emilia Romagna region taking into account its temporal trend (dependence on time), as well as spatial correlations across stations (the closer, the more similar we expect their PM_{10} concentration profiles) and other relevant geographical information (altitude, type, zoning and area). We will further discuss the interpretation and relevance of these variables to the problem at hand.
Our second goal is to develop a prediction method for daily and yearly average concentration, therefore enabling timely corrective action by policymakers.
Lastly, we aim at building several concurring models, at different levels of complexity and use available information, to be then able to decide which offers us the best understanding of this phenomenon.

2 Dataset presentation

The data employed in our study was gathered by ARPA Emilia Romagna, a public environmental monitoring institution, through 49 measurement stations scattered across the region. These cover several altitude levels, and were chosen to represent different degrees of anthropic pollution and activities. Specifically, stations are classified with three separate systems:

- Zoning: translates region-specific geographical cores. This category divides stations according to a merely geopolitical classification. Our zoning labels consist in: East plain, West plain, Agglomerate, Appennini.
- Type: The station can be of Traffic, Industrial or Background type. This characterization describes the specific anthropological footprint in the recording site, that must account for traffic pollution or industrial smoke when in presence of higher PM_{10} concentrations.

- Area: The station can be located in a Rural, Suburban or Urban area, according to the characteristics (e.g. population density) associated to its neighbouring sites.

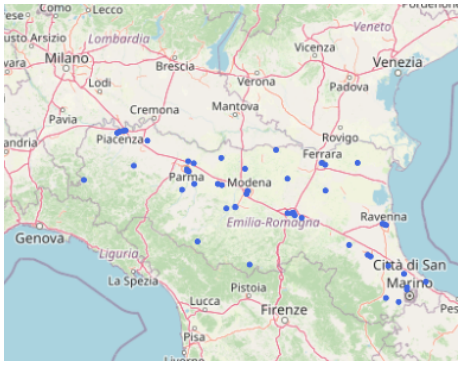


Figure 1. Geographical locations of the 49 stations in Emilia Romagna.

Moreover, each station is associated to its altitude (expressed in meters a.s.l).

Data for PM_{10} is typically registered daily and is never expected to disappear completely; however measures are always bounded from below by a minimal measurement threshold, under which the available machinery is not able to detect PM_{10} presence.

Although PM_{10} concentrations were available starting from 2014, our analysis focused on data collected in 2018. Our choice was motivated by our intention of creating a general model, whose predictive capabilities would not be hindered by post-Covid anomalies, or by considerable amounts of missing or unreliable data. Moreover, as environmental pollution data are known to display annual periodicity related to seasons, one year is conceptually enough to collect observations for all the relevant climatic conditions, albeit preventing further considerations on annual and inter-annual periodicity.

3 Data preprocessing

One of the first challenges presented by our project was to correctly treat, and successively reconstruct by Bayesian techniques, the missing observations in our dataset. Our exploratory analysis suggested that such missing observations do not follow a particular trend or pattern, and are mostly distributed unevenly across stations and seasons, rarely resulting in a distinguishable sequence of missing observations.

Moreover, we decided to log-transform PM_{10} concentration, backed by consistent scientific research on log-normality of environmental data. This naturally implied setting to 1 the values of PM_{10} that had been censored to 0 because of the minimum measurement tolerance mentioned above.

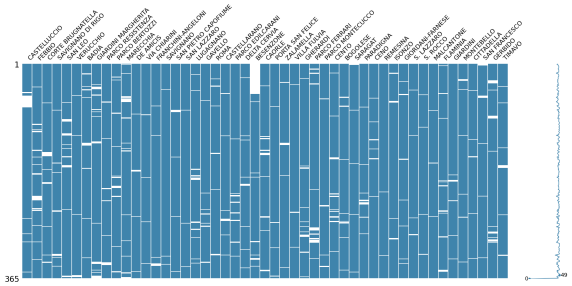


Figure 2. Missing data across time, by station.

4 Preliminary data exploration

Preliminary observations highlight the relevance of Area classification in explaining annual median observations of PM_{10} pollution, neatly overshadowing the classification capabilities of Zoning and Type. Our analysis will consequently refer to Area as the stations' preferential classification system.
Specifically, our explorations show rural areas tend to have a different behaviour over time (possibly a higher variance among temporal observations), and a lower overall PM_{10} concentration level. We will use this information to help us interpret the results of the analyses that will follow.

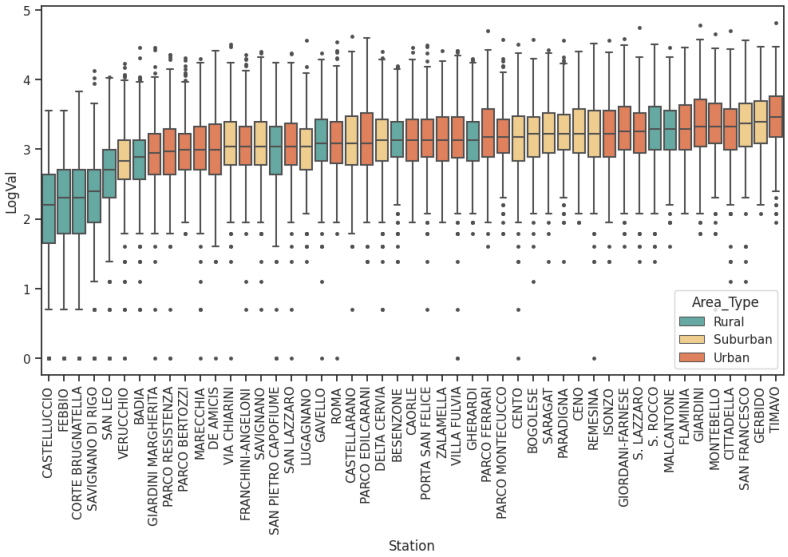


Figure 3. Time series boxplot by station, colored by area.

In addition to this, we noticed that almost all stations present a gradual attenuation of PM_{10} concentrations in spring, which is followed by consistently low recordings in summer, before measurements rise up again at the beginning of the colder season. This sort of temporal trend is well described by a loose U-shape in the time series of the data for almost all stations, and is in accordance with previous findings available in literature. Indeed, PM_{10} is normally removed from the atmosphere during spontaneous chemical processes that take place at soil level during warmer days. The only few stations (6 in total) that did not exhibit a visible U-shaped trend, but somewhat of a reversed trend, all belong to the Appennini (mountain chain) categorization, and possibly share some peculiar atmospheric characteristics. We will try to adapt our general model to be able to describe also this diverging trend for Appenninic stations.

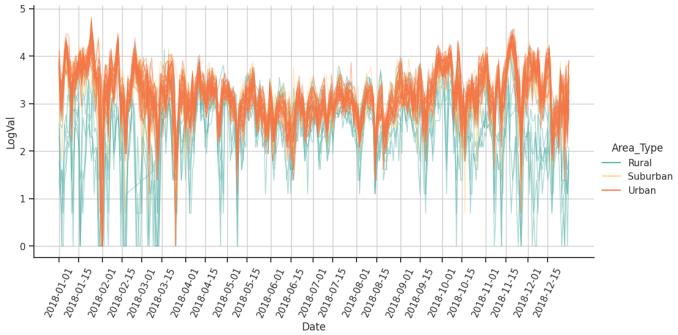


Figure 4. Logarithmically transformed data, coloured by area.

5 Our general model: the starting point

We therefore proceed to write our model in first approximation as:

$$\begin{aligned} Y_{s,t}|\text{params} &= f_s(t) + g_s(t) + \underline{x_s}^T \underline{\beta_0} + w_s + \varepsilon_{s,t} && s = 1, \dots, N \\ \varepsilon_{s,t}|\text{params} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) && t = 1, \dots, T \end{aligned}$$

Our model parcels out the log-transformed PM_{10} concentrations into three components, which will be progressively added to the model, so as to manage complexity in successive steps. Specifically, our model comprises:

- a function of time, $f_s(t)$, that will capture the general trend of the data.
- a function of time, $g_s(t)$, that will capture the periodic fluctuations induced by seasonality.
- a regression component that will take into account the additional information regarding altitude, intensity and type of anthropic activities.
- a spatial function that will enforce similarity of closer stations and independence of further ones (underlying smooth phenomenon).

For the temporal part, we tried three different kinds of models: an ARIMA model, a model based on Bayesian Structural Time Series (BSTS) and a model based on the forecasting procedure prophet developed by Facebook.

6 Temporal model

6.1 ARIMA: AutoRegressive Integrated Moving Average

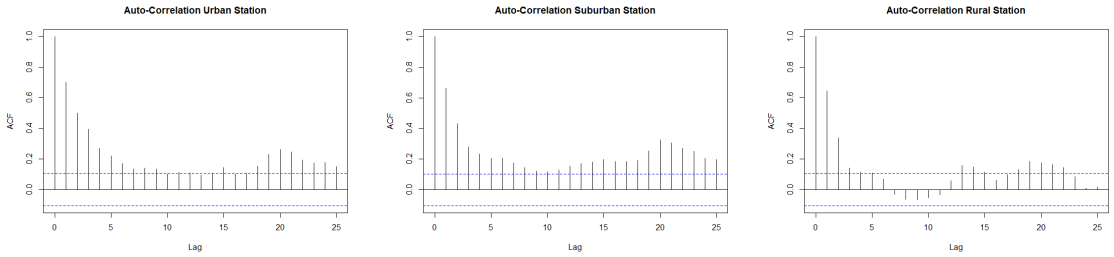
ARIMA (AutoRegressive Integrated Moving Average) models are universally considered to be milestones for time series analysis and are consistently and successfully used in a variety of applications and contexts.
The standard notation of this kind of model is ARIMA(p, d, q), and it can be viewed as an extension of a more widely known ARMA(p, q), applied to a time series that has already been differenced d times.
In short, an ARMA(p, q) model describes the behaviour of a time series through an AutoRegressive component (polynomial of order p) - which models observations at time t as a linear function of previous observations up to time $t - p$ - and a Moving Average component (polynomial of order q), which incorporates a linear combination of previous and actual uncorrelated errors, which are often unobserved, up to time $t - q$. An ARMA(p, q) model is completely described by assigning coefficients (which can potentially be null) to every degree of such polynomials, in order to obtain a complete and time-invariant description of phenomenon at hand.
Estimation of an ARMA(p, q) therefore implies statistical inference of its p + q polynomial coefficients, and can naturally embed a model selection criterion by reducing the order of the resulting model in case the highest order (p or q) is assigned a null coefficient with a certain credibility level.

An ARIMA(p, d, q) is thus an extension of an ARMA(p, q) which consists of a preprocessing of the time series under exam to ensure it has time-invariant statistical properties, such as mean and temporal covariance. Specifically, ARIMA(p, d, q) models study the d-differenced, stationary series obtained from the original by d successive 1-step differentiations ($y_t - y_{t-1}$). For instance, ARIMA$(p, 1, q)$ models are used to eliminate linear trends in data, such as increasing or decreasing observations mean over time.
A further extension consists in recovering time-series stationarity by also removing periodical trends: this, for an observed time period h, is achieved by differentiation by a step h, which translates to $y_t - y_{t-h}$ for each generic observation time t. Applying an h-step differentiation and then proceeding as above results in a SARIMA (Seasonal AutoRegressive Integrated Moving Average) model, which is commonly denoted by SARIMA$(p, d, q) \times (P, D, Q, h)$.

6.1.1 Setting of the parameters

As a first rough estimate, we tried to get a first hint at the appropriate maximum degrees to employ in our Bayesian analysis. It is important to point out that this a priori specification does not exclude model and order selection methods from our Bayesian model, but constitutes a great advantage from the computational point of view in both time and complexity.
Our procedure consisted in fitting with frequentist techniques an ARIMA model to each of the stations separately, so as to allow convergence and explore the possibility of fitting different temporal models to subsets of stations. Our primary output of interest would then be the fitted maximum orders across stations.
The seasonal trend was not clearly identifiable from our data and trying to fit seasonality into the model did not display any apparent benefits. Therefore, we continued this line of work through a non-seasonal ARIMA, keeping in mind that we might need to account for seasonal effects through other techniques.

Although not all stations displayed a clear non-stationarity pattern (which we investigated mostly by means of the ACF) (see Figure 5), we decided to set $d = 1$ for our analyses to have a zero-centered, comparable time series for all stations.



(a) **(b)** **(c)**

Figure 5. Partial autocorrelation functions for stations representative of the three areas. High and periodic autocorrelations for distant time lags suggests possible non-stationarity of the series.

In particular, we observed that every station was fine with an AR order of 2, even though most of them suggested being extracted from an AR of order 1. We decided to stick to an autoregressive order of 2 for all observations for the sake of generality, and draw further conclusions (if any) from the eventual zeroing out of such order 2 coefficients on a subset of stations.
It is to be noted that, in order for the resulting ARMA(p, q) to describe a stationary system, all zeros of the autoregressive polynomial should be strictly contained within the complex unit circle. Enforcing this condition corresponds to requiring both polynomial coefficients (which to us will be real numbers) separately fall into the real interval $(-1, 1)$. Such condition becomes more difficult to impose on a third-degree polynomial as there is no general known relation between boundedness of roots as functions of coefficients.

The results concerning the appropriate MA order were less clear, pointing at orders 1 or 2 as being more likely for our data. However, when Bayesian modelling techniques started to be involved, trying to increase the order of the MA part caused severe problems in both convergence and identifiability. This encouraged us to proceed with the fitting of a moving average component of order 1.

This preliminary discussion therefore motivated our choice of further investigating an ARIMA$(2, 1, 1)$ model.

To get our model started, we needed 3 observations per station referring to days preceding the beginning of our time frame: indeed, 1 is needed to compute the preliminary time differencing,

while the remaining 2 are necessary to compute second-order autoregressive effects on the first day of our time series.
Even though these data were available to us (with the exception of the very last day of 2017, which is conventionally not recorded in Emilia Romagna), we decided to opt for a more general reconstruction method that still provides a sensible starting point for our time series.
In the following, we will refer to these "missing" observations as y_{start}.

$$\begin{aligned} y_{start}|\mu_{start}, \sigma_{start} &\sim \mathcal{N}\left(\mu_{start}, \sigma_{start}\right) \\ \mu_{start} &\sim \mathcal{N}\left(y_{\text{yearly average}}, 1\right) \\ \sigma_{start} &\sim \mathcal{IG}\left(3, 2\right) \end{aligned} \tag{1}$$

Missing data inside the analysed time span are also treated separately in order to reconstruct them:

$$y_{missing} \sim \mathcal{N}\left(y_{\text{yearly average}}, 1\right) \tag{2}$$

6.1.2 First step: univariate ARIMA

As a first attempt, fitting was performed on only one station at a time, resulting in a univariate ARIMA model, whose estimates on different stations are respectively independent.
In order to impose the $(-1, 1)$ bound on AR polynomial coefficients, we selected a sigmoid function, which maps continuously the information on such coefficients to the interior of the stable interval. This also affords us to exclude truncated-support distributions, which might hinder convergence and, in addition, provide less robust estimates.
Our sigmoid reads as:

$$\Sigma(z) = \frac{e^z - 1}{e^z + 1} \tag{3}$$

then:

$$\begin{aligned} \mathbf{y}|\,\gamma_\theta, \underline{\gamma_\phi}, \sigma &\sim \text{ARIMA}_{2,1,1}\left(\Sigma(\gamma_\phi[1]), \Sigma(\gamma_\phi[2]), \Sigma(\gamma_\theta), \sigma\right) \\ \gamma_\phi[j] &\sim \mathcal{N}\left(0, 1\right), \qquad j \in \{1, 2\} \\ \gamma_\theta &\sim \mathcal{N}\left(0, 1\right) \\ \sigma &\sim \mathcal{H}\text{alf-Cauchy}\left(0, 5\right) \end{aligned} \tag{4}$$

where we denote by ϕ the vector containing the two AR fitted coefficients and θ the MA coefficient.

6.1.3 Model (A): multivariate ARIMA

Our results were promising and motivated us to undertake the next step, which consisted in fitting a multivariate ARIMA process, considering all stations at once. This way, we could properly share information across stations and also draw some conclusions on the similarity of stations' temporal models.

$$\begin{aligned} \mathbf{y}_s|\,\gamma_\theta, \underline{\gamma_\phi}, \sigma, \underline{\mu_\phi}, \underline{\sigma_\phi} &\sim \text{ARIMA}_{2,1,1}\left(\Sigma(\gamma_{\phi,s}[1]), \Sigma(\gamma_{\phi,s}[2]), \Sigma(\gamma_\theta), \sigma\right) \\ \gamma_{\phi,s}[j]|\,\underline{\mu_\phi}, \underline{\sigma_\phi} &\sim \mathcal{N}\left(\mu_\phi[j], \sigma_\phi[j]\right), \qquad j \in \{1, 2\} \\ \mu_\phi[j] &\sim \mathcal{N}\left(0, 5\right) \\ \sigma_\phi[j] &\sim \mathcal{IG}\left(2.1, 1.1\right) \\ \gamma_\theta &\sim \mathcal{N}\left(0, 1\right) \\ \sigma &\sim \mathcal{H}\text{alf-Cauchy}\left(0, 5\right) \end{aligned} \tag{5}$$

We were able to compute posterior densities for this model, and convergence diagnostics looked good. We will refer to this model in the following as model (A): our first building block on the

road to a complete regressive spatio-temporal model.
In critically evaluating model (A), we took in great account the produced predictive posteriors and we represented them using 95% credible intervals. In Figure 6, the blue time series represents the true trajectory, whereas the red line is the predicted mean.

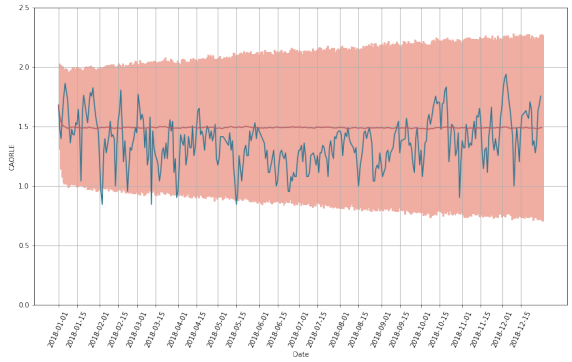


Figure 6. Time series for Caorle station, with predicted mean and 95% credible interval, under model (A).

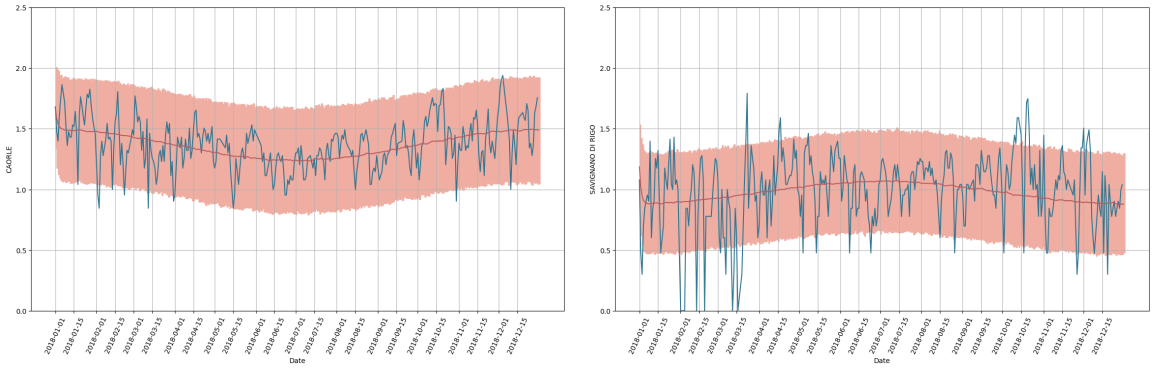
It is very apparent that this model does not replicate the trend of the time series on the predicted mean. We explain this phenomenon by the mild, yet unmodelled, periodicity that characterises the problem at hand.

6.1.4 Model (A.1): multivariate ARIMA with periodicity

To obtain a more accurate result we then understood we needed a time function which would capture the loosened U-shape most of our data display. The next step then consisted in adding a sinusoidal component to the model, which could be capable of modelling both the dominant U-trend that is shown above, and the inverse, "bell" trend that is typical of Appenninic stations. We will further refer to this model as model (A.1).

$$
\begin{aligned}
\mathbf{y}_s|\gamma_\theta, \underline{\gamma_\phi}, \sigma, \underline{\mu_\phi}, \underline{\sigma_\phi} &\sim \text{ARIMA}_{2,1,1}\left(\Sigma(\gamma_{\phi,s}[1]), \Sigma(\gamma_{\phi,s}[2]), \Sigma(\gamma_\theta), \sigma\right) + c_s \cdot cos\left(\frac{2\pi}{365}t\right) \\
c_s &\sim \mathcal{N}(0,1) \\
\gamma_{\phi,s}[j]|\,\underline{\mu_\phi}, \underline{\sigma_\phi} &\sim \mathcal{N}\left(\mu_\phi[j], \sigma_\phi[j]\right), \qquad j \in \{1,2\} \\
\mu_\phi[j] &\sim \mathcal{N}\left(0,5\right) \\
\sigma_\phi[j] &\sim \mathcal{IG}\left(3,2\right) \\
\gamma_\theta &\sim \mathcal{N}\left(0,1\right) \\
\sigma &\sim \log\mathcal{N}\left(\mu_\sigma, \sigma_\sigma\right) \\
\mu_\sigma &\sim \mathcal{N}\left(0,1\right) \\
\sigma_\sigma &\sim \mathcal{IG}\left(3,2\right)
\end{aligned}
\tag{6}
$$

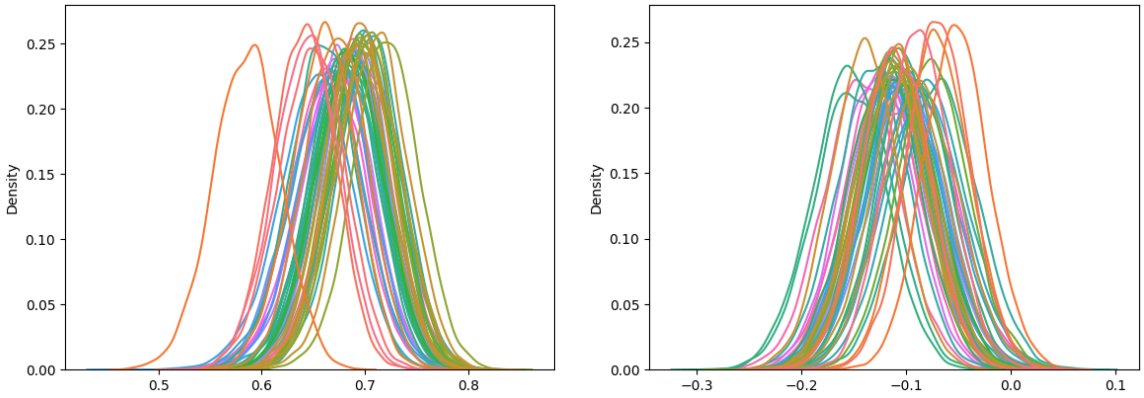
To demonstrate the effect of this model extension, we report in Figure 7a the newly predicted mean and 95% credible intervals on Caorle station. We invite readers to compare this plot with Figure 6, which depicts model (A) estimates on this same station.



(a) **(b)**

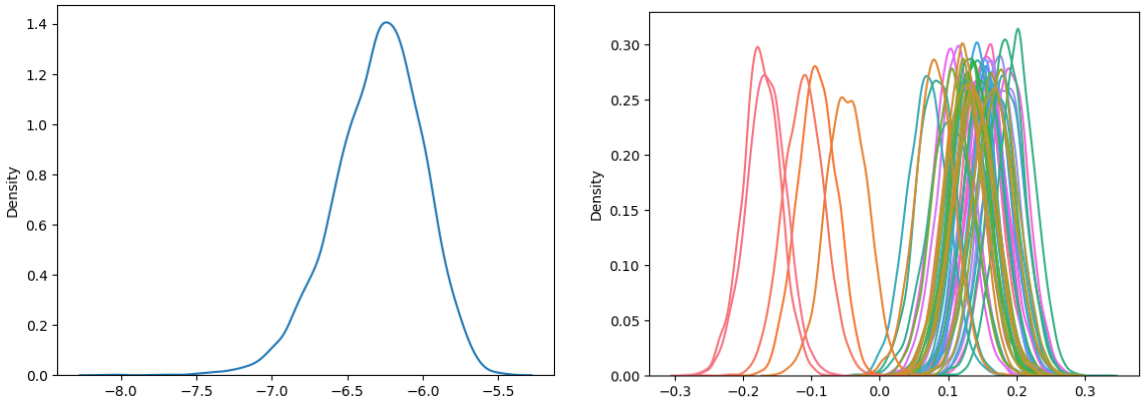
Figure 7. 7a Is the time series for Caorle station, while 7b is an example of an inverse-trend Appenninic station, (Savignano di Rigo), both with predicted mean (red line) and 95% credible interval computed with model (A.1).

To account for across-stations variability, we chose to let the sinusoidal coefficient c_s vary with the stations s in our dataset.
Posterior inference on cosine amplitude c_s (see 9b) reflects very clearly the presence of two complementary temporal trends by clearly separating the majority of observations from a smaller cluster of negative estimates (the reverse trend).



(a) **(b)**

Figure 8. 8a First and 8b second order autoregressive coefficients



(a) **(b)**

Figure 9. 9a Posterior distributions for the logit-transformed moving average coefficient
9b Posterior distributions (by station) for cosine amplitudes c_s

6.2 BSTS: Bayesian Structural Time Series

BTSTS (Bayesian Structural Time Series) models are very simple and modular methods that allow to disentangle various effects on a time series in an additive manner. Their formulation therefore naturally encomprises the estimation of trend, seasonal and regression components.
Our efforts on Structural Time Series analysis in a Bayesian framework were also guided by the desire to develop an alternative, concurrent time series model to compare to our ARIMA temporal model.

In its most general form, a Bayesian Structural Time Series is a state-space model, which relates the observations Y_t to a vector of latent (and possibly unobserved) state variables α_t.

$$\begin{aligned} Y_t &= Z_t^T \alpha_t + \epsilon_t & \epsilon_t &\sim \mathcal{N}(0, H_t) \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t & \eta_t &\sim \mathcal{N}(0, Q_t) \end{aligned} \tag{7}$$

The first of the equations presented above is called the observation equation, because it links explicitly the observed data Y_t with unobserved latent variables α_t, with the addition of some random measurement noise.
The second equation presented above is instead called transition equation because it defines how the latent state evolves over time: this in itself can be viewed as an ARMA(1,1) process whose coefficients are contained in the matrices T_t and R_t respectively.
Typically, all model matrices Z_t, T_t, and R_t contain a mix of known values, and unknown parameters.

More specifically, a full BSTS model reads as follows:

$$Y_t|\text{params} \sim \mu_t + \tau_t + \beta^T \underline{x} + \varepsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2_{obs}) \tag{8}$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t \qquad u_t \sim \mathcal{N}(0, \sigma^2_{level}) \tag{9}$$

$$\delta_t = \delta_{t-1} + v_t \qquad v_t \sim \mathcal{N}(0, \sigma^2_{slope}) \tag{10}$$

$$\tau_t = -\sum_{h=1}^{S-1} \tau_{t-h} + w_t \qquad w_t \sim \mathcal{N}(0, \sigma^2_{per}) \tag{11}$$

The parameters in these equations are the variances $\sigma^2_{obs}, \sigma^2_{level}, \sigma^2_{slope}, \sigma^2_{per}$ and the regression coefficients β.

In our context, a BSTS model aims at explaining the current level of the trend through Equation (9) and its current "slope" (linear approximation) through Equation (10).
Equation (11) focuses on seasonal effects and can be thought of as a set of S dummy variables with dynamic coefficients constrained to have zero expectation over a full cycle of S seasons.

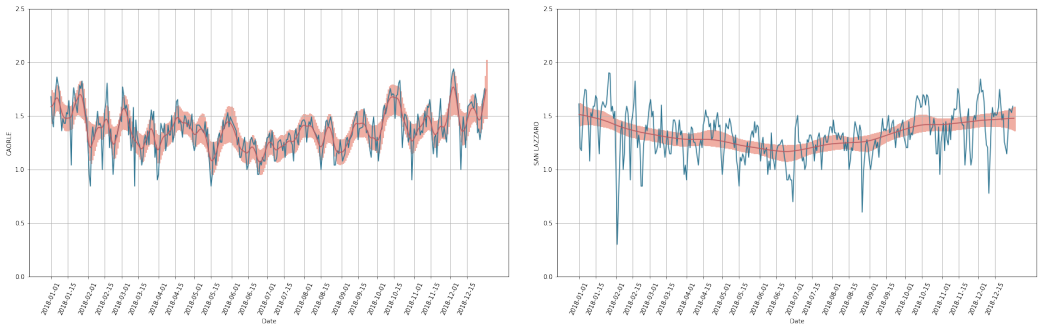
6.2.1 Model (B): BSTS station by station

At first, we decided to ignore regression components and focus on the purely temporal components of the model. It is to be understood that our application of such BSTS model was carried out station by station, and therefore returns independent estimates for each time series.
As priors for the different σ's we employed Scale-Dependent priors automatically obtained using the sdPriors package ([10]).
After a first attempt, since our results indicated that periodicity did not seem to have a relevant impact, and as its results were unclear, Equation (11) was removed from our BSTS model.

What follows is therefore our simplified BSTS model applied separately for each station s. We will refer to the following as model (B):

$$\begin{aligned} Y_{s,t} &\sim \mu_{s,t} + \varepsilon_{s,t} & \epsilon_{s,t} &\sim \mathcal{N}(0, \sigma^2_{obs}) \\ \mu_{s,t} &= \mu_{s,t-1} + \delta_{s,t-1} + u_{s,t} & u_{s,t} &\sim \mathcal{N}(0, \sigma^2_{level}) \\ \delta_{s,t} &= \delta_{s,t-1} + v_{s,t} & v_{s,t} &\sim \mathcal{N}(0, \sigma^2_{slope}) \end{aligned} \tag{12}$$

We report two examples of posterior distributions, superimposed on the real observed data. Surprisingly, BSTS seems to return two different types of model (see Figure 10), but the reasons behind this differing behaviour remain unclear.



(a) Caorle station. **(b)** San Lazzaro station.

Figure 10. Comparison of model (B) fit on two different stations.

One reasonable hypothesis attributes the seemingly overfitting behaviour that occurs in some stations (as in Figure 11a) to a poor choice of prior distributions, which the library we employed did not give us much control on.
On the other hand, it is to be noted that as BSTS models were initially deployed on nowcasting problems (i.e. settings in which a whole time series is used to predict the next single observation), our final outputs (as shown in Figure 10) do not represent posterior predictive distributions for the whole time series, but rather a sequence of mean estimates computed using all available observations up to that time. Therefore, they cannot be exactly compared to the posterior predictive distributions obtained through ARIMA (see Figures 6, 7a), and as such they do not provide a completely adequate answer to our research questions.

After some thought, we decided to not continue investigating Bayesian Structural Time Series in our work, but rather to consider model (B) as a tentative alternative which did not eventually outperform our currently available methods.
Indeed, despite its problems in capturing weekly obscillatory trends, ARIMA models keep being more interpretable and able to render an appropriate description of the available data.

6.2.2 Facebook Prophet

A new attempt was made with another univariate model, using a decomposable time series model: Prophet. It is a procedure for forecasting time series data where non-linear trends are fit with yearly, weekly, and daily seasonality. This method is robust to outliers, missing data, and dramatic changes in the time series.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2) \qquad t = 1, ..., T \tag{13}$$

$g(t)$ is the trend component, $s(t)$ is the seasonality (models periodic changes), ϵ_t is the error term. A linear growth trend is used for forecasting problems showing a non-saturating growth trend:

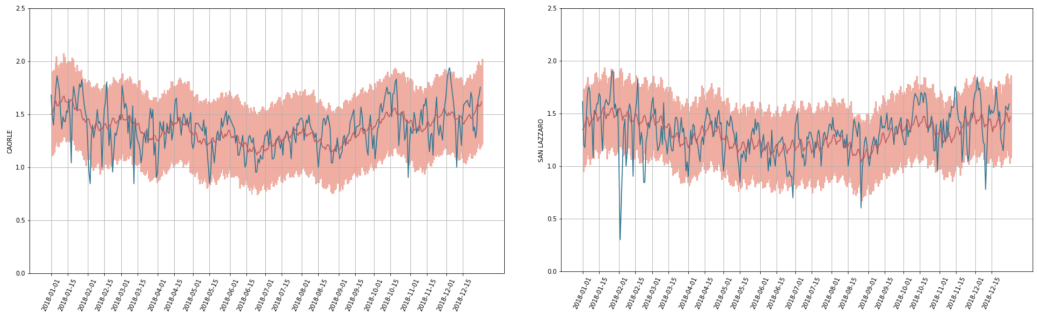
$$g(t) = k't + m', \quad k' = k + \underline{a}(t)^T\underline{\delta}, \quad m' = m + \underline{a}(t)^T\underline{\gamma}, \quad \gamma_j = -s_j\delta_j$$

Defining S changepoints at times s_j (Prophet will automatically detect these changepoints): k' represents the growth rate and m' the offset parameter (both change over time).
a_j(t)=1 , $t \leq s_j$ (0 otherwise). $\underline{\delta}$ is a vector of rate adjustments where
$\delta_j \sim Laplace(0, \tau)$ is the change in rate that occurs at time s_j, and the correct adjustment at changepoint j γ_j :

$$\gamma_j = (s_j - m - \sum_{l<j} \gamma_l)(1 - \frac{k + \sum_{l<j} \delta_l}{k + \sum_{l \leq j} \delta_l})$$

The arbitrary smooth seasonal effects are approximated with a standard Fourier series:

$$s(t) = \sum_{n=1}^{N} (a_n cos(\frac{2\pi nt}{P}) + b_n sin(\frac{2\pi nt}{P})), \quad \beta = [a_1, b_1, .., a_N, b_N], \quad \beta \sim N(0, \sigma^2)$$



(a) Caorle station. **(b)** San Lazzaro station.

Looking at the predictive posterior represented using a 95% credible interval, the model seems to fit the data well and follow the annual seasonality. Being a univariate model it is fitting one station at a time. Since the ARIMA model was well suited for a multivariate implementation, we chose to proceed along that route.

7 Regression model

In order to embed the relevant topographical information into our model, we add to the temporal model (A.1) - which we believe is the best among the 3 concurrent temporal explanations - a regressive component.
Our effort in such sense culminates in the construction of a temporal model with additional station-level covariates, which we denote in the following section by model (C).

7.1 Setting of the parameters

As we already mentioned in the preliminary steps to building model (A) (which were also inherited by model (A.1)) (see Section 6.1.1), our temporal ARIMA model needed three observations per station referring to days preceding the beginning of our time frame. In the set up of those models, we referred to these "missing" observations as y_{start}, and modelled them as station-specific parameters of our problem. By extending model (A.1), we were confronted with the necessity of rethinking this assumption, as mingling station-specific intercepts with regressive covariates presented us with drifting and non-identifiability effects. In other words, our model was overparametrized and our estimates failed to converge.

Our only choice was then to decrease the number of parameters in our model and give up on some of the modelling freedom we had allowed in previous steps.
Specifically, our y_{start} parameters collapsed in a single y_0 model parameter, describing the starting point of our time series for all stations and for all the three required observations.
Some alternative approaches, which led to similar issues, were explored. This delicate matter is further discussed as a suggested possible development (see Section 11).

7.2 Model (C): results and discussion

Our regressors consist in 4 variables:

- Station altitude,
- Station zoning: the part of the region the station is in (East plain, West plain, Agglomerate, Appennini),
- Station type (Traffic, Industrial, Background),
- Station area (Rural, Urban, Suburban).

Our expectations could be summed up as follows:

- Industrial and Traffic regressors having positive regression coefficients, as opposed to less polluted Background stations;
- Rural areas displaying lower pollution levels, reflecting on a negative coefficient;
- Altitude having a negative impact on pollution levels.

The model we obtained is the following:

$$\begin{aligned} \mathbf{y}_s|\gamma_\theta, \underline{\gamma_\phi}, \sigma, \underline{\mu_\phi}, \underline{\sigma_\phi}, \underline{\beta} &\sim \text{ARIMA}_{2,1,1}\left(\Sigma(\gamma_{\phi,s}[1]), \Sigma(\gamma_{\phi,s}[2]), \Sigma(\gamma_\theta), \sigma\right) + c_s \cdot cos\left(\frac{2\pi}{365}t\right) + \underline{x_s}^T \underline{\beta} \\ \underline{\beta} &\sim \mathcal{N}(0,1) \end{aligned} \tag{14}$$

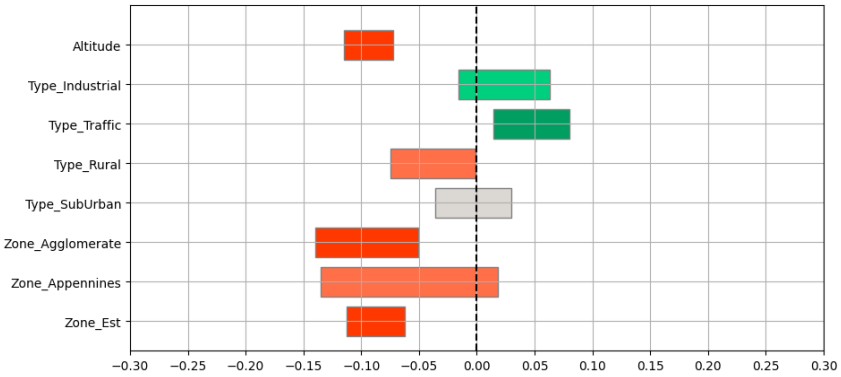


Figure 12. Covariates' 95% credibility intervals

The results show that Traffic stations stand out on PM10 pollution and that altitude is relevant. As was already discussed during exploratory analysis of the data, there is no difference between Urban and Suburban areas, while Rural areas are less polluted. Besides, stations on the West plain zone appear to be more polluted.

8 Spatial model

As a last step to complete our model, finally using all the information provided in the dataset, we took on the task to include spatial dependence across stations by means of their coordinates. Therefore, the subsequent steps constitute a further extension of model (C), which is assumed as the starting point in the following.
The resulting regressive, spatio-temporal model is denoted as model (D).

8.1 General description of a spatial Gaussian process

First of all, we provide a brief introduction to spatial Gaussian processes from a theoretical point of view.
A spatial process is *Gaussian* if for any $n > 1$ and for any set of locations $\{\underline{s_1}, .., \underline{s_n}\}$, the multivariate response variable is Gaussian; in our case, we focused on stationary isotropic spatial processes, also called *homogeneous*, characterized by a covariance matrix that only depends on the norm of the separation vector $\underline{h}$.
For these models the *variogram* is defined as:

$$2\gamma(\underline{h}) = Var(Y(\underline{s} + \underline{h}) - Y(\underline{s})) \tag{15}$$

Concretely, if a homogeneous Gaussian process is assumed, simple parametric formulations can be proposed as feasible candidates for the semivariogram $\gamma(\underline{h})$. A very popular formulation, which is the one selected in Michela Frigeri's Master Thesis and the one we decided to pick as well, is the *exponentiated quadratic covariance function*:

$$C(h) = \begin{cases} \alpha^2 + \sigma^2 & h = 0 \\ \alpha^2 \exp\left(-\frac{h^2}{2\rho^2}\right) & h > 0 \end{cases} \tag{16}$$

Hyperparameter ρ is the *length-scale* parameter, regulating the impact of spatial correlation in the Gaussian process. Small values of this parameter lead the Gaussian process to have nearly null covariance between different locations, while higher values bring higher correlation. Hyperparameter α represents the *marginal standard deviation*, controlling the magnitude of the variability assumed to characterize the Gaussian process.
Given the generic spatial model:

$$Y(\underline{s}) = \mu(\underline{s}) + w(\underline{s}) + \epsilon(\underline{s}) \tag{17}$$

$\mu(\underline{s})$ typically stands for the regression part of the model, while the residual portion is split between two terms, one spatial and one non-spatial.The spatial residual term $w(\underline{s})$ is a realization from a zero-mean Gaussian spatial process. The non-spatial residuals $\epsilon(\underline{s})$ are, instead, uncorrelated pure error terms.
Relying on the above characterization of residual terms and covariance function, the partial sill α^2 and range ϕ are given by the spatial residual term $w(\underline{s})$, while the non-spatial residual term $\epsilon(\underline{s})$ is responsible for the nugget effect σ^2. To be more specific, we consider the range parameter ϕ as proportional to the ρ parameter adopted in covariance function above.
In the field of Gaussian spatial prediction, assuming to have covariates affecting the process' outcome, the simplest possible model for observed data is given by:

$$\begin{aligned} \underline{Y} &= X\underline{\beta} + \underline{\epsilon} \\ \underline{\epsilon} &\sim \mathcal{N}(\underline{0}, \Sigma) \\ \Sigma &= \sigma^2 H(\rho) \\ H_{ij}(\rho) &= g(\rho; \|\underline{s_i} - \underline{s_j}\|) \end{aligned} \tag{18}$$

where, in our situation, $g()$ will be the exponentiated quadratic covariance function.
Finally, introducing the Bayesian approach in this framework, we model the response variable as:

$$\underline{Y}|\text{params} \sim \mathcal{N}(X\underline{\beta}, \sigma^2 H(\rho)) \tag{19}$$

8.2 Setting of the parameters

As also covered in Sections 6.1.1 and 7.1, the choice of sensible estimates for the first 3 temporal observations of each time series is a crucial point in determining convergence and identifiability of our model. With the adding of covariates in model (C) in Section 7.1, our starting observations were collapsed into a single model paramater y_0, which was used to describe all three previous measurements for all stations. However, spatial effects confronted us yet again with the necessity of reducing our number of model parameters even further, thus driving us to fix y_0 - which retains the same purpose in the model - to the mean annual value of PM_{10} across all stations in our dataset.
For comments and criticism on this choice see Section 11.

8.3 Model (D): results and discussion

As specified above, we decided to employ an exponentiated quadratic covariance function to describe the spatial residuals w_s, thus obtaining the subsequent model:

$$\begin{aligned}
\mathbf{y}_s|\gamma_\theta, \underline{\gamma_\phi}, \sigma, \underline{\mu_\phi}, \underline{\sigma_\phi}, \underline{\beta} &\sim \text{ARIMA}_{2,1,1}\left(\Sigma(\gamma_{\phi,s}[1]), \Sigma(\gamma_{\phi,s}[2]), \Sigma(\gamma_\theta), \sigma\right) + \mathbf{c}_s \cdot cos\left(\frac{2\pi}{365}t\right) + \underline{x}_s^T\underline{\beta} + w_s \\
\underline{w} &\sim \mathcal{N}\left(\underline{0}, \Sigma\right) \\
\Sigma_{ij} &= \alpha^2 \exp\left(-\frac{1}{2\rho^2}\|s_i - s_j\|^2\right) \\
\alpha &\sim \mathcal{IG}(6, 2) \\
\rho &= 0.05
\end{aligned} \tag{20}$$

Due to convergence issues, the ρ parameter was fixed rather than having a prior distribution being assigned - we consider this, however, as an interesting possible development. Moreover, it should also be noted that the value we set is very small, as trying to increase it caused convergence to be unsuccessful; this could be related to the fact that PM_{10} pollution is strongly related to combustion, making it a local phenomenon.

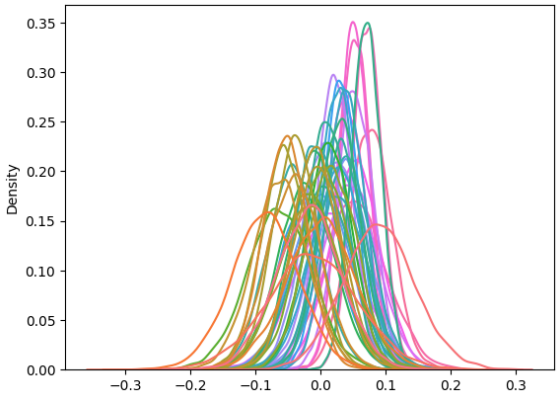


Figure 13. Posterior distributions (by station) for spatial residuals w_s.

9 Model comparison

9.1 Methods

The analysis performed in this section will be based on two predictive model selection criteria:

- Widely Applicable Information Criterion(WAIC).
- Leave One Out Cross Validation(LOO-CV).

These two techniques estimate pointwise out-of-sample prediction accuracy using the log-likelihood evaluated at posterior simulations of the parameters; resulting values are guaranteed to be asymptotically equal. The WAIC for model M_j is defined as:

$$WAIC_j = -2\left(LPPD_j\right) + 2\sum_{i=1}^{n} \mathrm{Var}_{\theta_j|\mathbf{y}}\left[\log\left(p_i\left(y_i \mid \theta_j, M_j\right)\right)\right] \tag{21}$$

where $LPPD_j$ is the *Log-Pointwise Predictive Density* of M_j, expressed, together with its MCMC approximation, as:

$$\begin{aligned} LPPD_j &= \sum_{i=1}^{n} \log\left(m\left(y_i \mid \mathbf{y}, M_j\right)\right) \\ \text{MCMC computed } LPPD_j &= \sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{k=1}^{M} p_i\left(y_i \mid \theta_j^{(k)}, M_j\right)\right) \end{aligned} \tag{22}$$

where $m\left(\cdot \mid \cdot, M_j\right)$ is the posterior marginal under M_j, while $\theta_j^{(k)}$ is the k-th MCMC sample from the posterior of the parameter.
The LOO-CV leans on the *Estimated Log-pointwise Predictive Density*(elpd), outlining the following out-of-sample predictive fit:

$$\begin{aligned} \mathrm{elpd}_{loo} &= \sum_{i=1}^{n} \log\left(p\left(y_i \mid \mathbf{y}_{-i}\right)\right) \\ \text{where } p\left(y_i \mid \mathbf{y}{-}i\right) &= \int_{\Theta} p\left(y_i \mid \theta\right) p\left(\theta \mid \mathbf{y}_{-i}\right) \end{aligned} \tag{23}$$

$p\left(y_i \mid \mathbf{y}{-}i\right)$ is the leave-one-out predictive density, obtained by removing the i-th observed value. The corresponding selection criterion, called PSIS-LOO, will be computed as:

$$\text{PSIS-LOO } = -2elpd_{loo} \tag{24}$$

Assuming asymptotic equality of the two quantities, warranted by the size of our MCMC simulations, we will choose the best performing model as the one associated to the lowest value of WAIC and PSIS-LOO.
We will use these two criteria to compare, in particular, our three ARIMA-based models denoted previously by A.1, C, D and the one developed by our tutor Michela Frigeri in her master's thesis *Spatio-Temporal Models for Particulate Matter in the Po Valley.*

To be more specific, we recall Michela's model (MF):

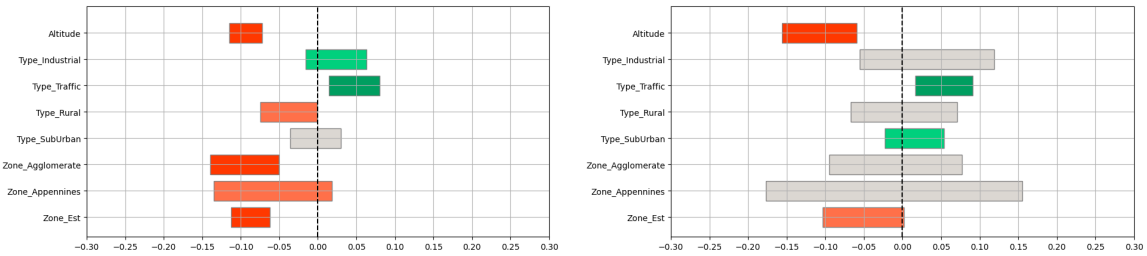
$$
\begin{aligned}
Y_i(t) &\overset{ind}{\sim} \mathcal{N}\left(\mu_i(t), \sigma^2_{m(t)}\right) i = 1, \dots, 49, \quad t = 0, \dots, 364 \\
\mu_i(t) &= f_R(t)\mathbb{1}\{\text{area } (i) = \text{ rural }\} + f_{NR}(t)\mathbb{1}\{\text{area } (i) \neq rural\} + \mathbf{x}i^T\boldsymbol{\beta} + w_i \\
f_R(t) &= a_{R1}\sin(\omega t) + b_{R1}\cos(\omega t) + a_{R2}\sin(4\omega t) + b_{R2}\cos(4\omega t) + c_R \\
f_{NR}(t) &= a_{NR1}\sin(\omega t) + b_{NR1}\cos(\omega t) + a_{NR2}\sin(4\omega t) + b_{NR2}\cos(4\omega t) + c_{NR} \\
\sigma^2_1, \dots, \sigma^2_{12} &\overset{iid}{\sim} \text{Inv Gamma}(3, 2) \\
a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R &\overset{iid}{\sim} \mathcal{N}(0, 1) \\
a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR} &\overset{iid}{\sim} \mathcal{N}(0, 1) \\
\beta_0, \beta_1, \beta_2 &\overset{iid}{\sim} \mathcal{N}(0, 1) \\
\mathbf{w} &\sim \mathcal{N}(\mathbf{0}, \Sigma) \\
\Sigma_{i,j} &= \alpha^2 \exp\left(-\frac{1}{2\rho^2}\|s_i - s_j\|^2\right) \\
\alpha &\sim \mathcal{N}(0.3, 0.1) \\
\rho &\sim \text{Beta}(3, 10)
\end{aligned}
\tag{25}
$$

9.2 Results

Model	WAIC	LOO-CV
A.1	-14528.97	-14506.23
C	-14522.44	-14500.04
D	-14530.08	-14505.76
MF	38284.5	38284.8

Table 2. WAIC and LOO-CV values of the models.

As all models seem to yield comparable goodness-of-fit indicators, our choice falls on model (C), which provides us with a good and easily-interpretable description while keeping the number of parameters limited.
Spatial and regressive components compete in explaining local fluctuations, and this also reflects into the higher variability of regressive coefficient estimates in model (D). It is, in particular, to be remembered that the very small ρ we employ for spatial dependence enforces correlation only among close stations. In addition, zoning seems to lose meaning when applied in concurrence with spatial effects.
By looking at credibility intervals of the regressors in the spatial model, it seems that only altitude and traffic label can be considered as still relevant indicators of the PM_{10} phenomenon.



(a) **(b)**

Figure 14. 14a Covariates' 95% credibility intervals of the regression model. 14b Covariates' 95% credibility intervals of the spatial model

10 Forecasting

We therefore employ our best model (as resulting from our model comparisons, see 9) to forecast which stations will overstep legal PM_{10} thresholds on both annual and daily limits.
Our method is therefore based on the comparison of STAN's generated quantities with the limit values reported in 1 under "Europe".
The final results then represent the probability of trespassing the legal limit (for each station), for annual and daily limits separately.
Our results show that, in spite of the very low (and statistically irrelevant) probability of exceeding the average annual threshold, probabilities associated to surpassing the daily limits are noticeably high. In Figure 15 we show our computed probability of registering values higher or equal than $50\mu g/m^3$ for more than 35 days per year, ordered by decreasing median annual concentrations. It is evident that, with many of the averagely highly-polluted stations approaching the extreme limit of 1, not all of the high-risk stations also correspond to extreme median annual values.

Our findings therefore support the introduction of stricter short-term environmental policies to counteract high PM_{10} concentration peaks, rather than the reinforcement of long-term acts on annual emissions - which do not seem particularly worrying.
As a last comment, we would like to point out that, if we were to refer our analysis to WHO suggested limit values instead, current PM_{10} concentration levels would be considered at very high (and almost certain for most stations) risk of exceeding the prescribed thresholds.

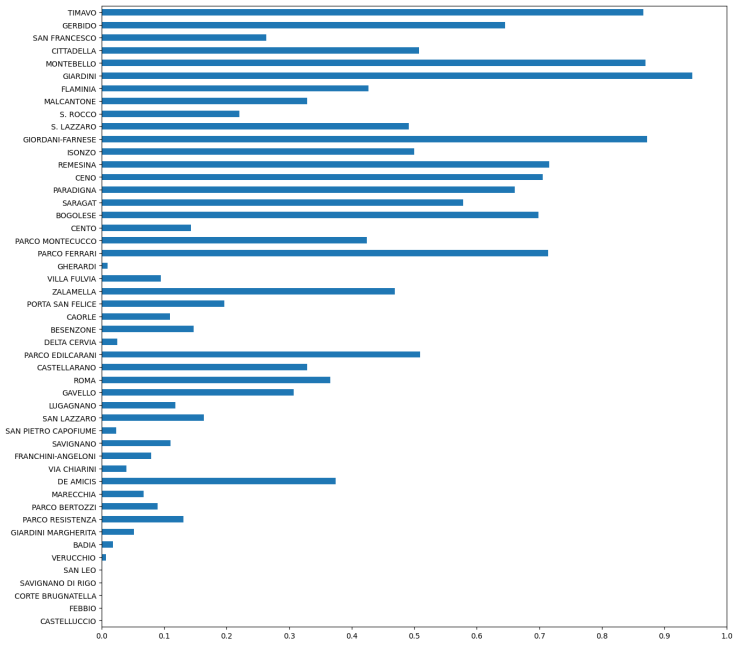


Figure 15. Forecasts on daily limits as outputted by predictions from model (C)

11 Conclusions and possible developments

Our findings support the commonly-held belief describing PM_{10} pollution as a phenomenon with short memory and scarce diffusion capability, peaking at emission sites and in short-term occurrences. We also found that regressive components and spatial dependencies compete in explaining such small-scale variability.
To refine our model, we would suggest to further investigate a covariates-only model, which might possibly involve more detailed information (e.g. on weather conditions). Also, relevant classifications to take into account could distinguish Rural stations with respect to Urban and Suburban ones, or Traffic stations with respect to Industrial and Background. Moreover, second-order AR terms could be recollapsed in a single estimate, valid for all stations.
To further account for phenomenon specifics, it might be interesting to distinguish workdays from weekends, as to better predict probabilities of trespassing a threshold for a certain number of consecutive days (which is part of current policymaking action, but not subject to regulatory EU acts).
We remark that our model failed to explain the seeming heteroschedasticity that affected rural station time series with respect to Urban and Suburban ones.
Moreover, trying different priors on length-scale ρ could provide more insight with respect to fixing it to a small constant, as in our current spatial analysis.
Lastly, it became clear to us that modelling choices on "starting" observations (i.e. pertaining to the 3 days previous to our time frame) proved of paramount importance to ensure convergence and reasonable computation times for our MCMC.
As we discussed in Sections 6.1.1, 7.1 and 8.2, such modelling choices varied in the course of our analysis to adapt to our subsequent models' rise in complexity.
Specifically, we noticed that Rhat convergence diagnostic on the very first "starting" day was often the most problematic among all parameters and required up to 1000 additional simulations to be brought back to acceptable values. This, however, is not surprising if we consider that the first starting day of every time series is the furthest from the observed data, and therefore the one we are feeding the model less information on.
However, the Bayesian models we built were all very dependent on our choice of starting points for our time series, indeed computation times varied greatly across the trials that experimented on such competing options.
Hence, we believe that a natural possible development of this works consists in a comprehensive investigation and discussion of this phenomenon, with the aim of facilitating model convergence, while conserving as much model elasticity as possible.

12 Acknowledgements

We would like to express our most heartfelt thanks to Prof. Ilenia Epifani and PhD student Michela Frigeri, in appreciation of their availability and precious advice.

References

[1] Andrio Adwibowo. "ARIMA forecasting of PM2. 5 and PM10 trends: effects of continuing social distancing on air quality in a Southeast Asian urban area". In: (2020).

[2] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[3] Peter J Brockwell et al. "ARMA models". In: *Introduction to Time Series and Forecasting* (2016), pp. 73–96.

[4] Michela Frigeri. "Spatio-temporal models for particulate matter in the Po valley". In: (2022).

[5] M Gianella, A Guglielmi, G Lonati, et al. "A Bayesian spatio-temporal model of PM10 pollutant in the Po Valley". In: *Book of Short Papers-SIS 2022*. Pearson, 2022, pp. 883–888.

[6] Yiannis Kamarianakis, Poulicos Prastacos, et al. "Spatial time series modeling: A review of the proposed methodologies". In: *The Regional Economics Applications Laboratory* (2003).

[7] PJ Garcia Nieto et al. "PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study". In: *Science of the Total Environment* 621 (2018), pp. 753–761.

[8] Istituto Superiore per la Protezione e Ricerca Ambientale. "Stima della concentrazione di PM10 in Italia con modelli statistici spazio-temporali: esempi applicativi". In: (2022).

[9] Steven L Scott and Hal R Varian. "Predicting the present with Bayesian structural time series". In: *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1-2 (2014), pp. 4–23.

[10] Steven L Scott et al. "Package 'bsts'". In: (2022).

[11] Sean J Taylor and Benjamin Letham. "Forecasting at scale". In: *The American Statistician* 72.1 (2018), pp. 37–45.

[12] Weiqiang Wang and Ying Guo. "Air pollution PM2. 5 data analysis in Los Angeles long beach with seasonal ARIMA model". In: *2009 international conference on energy and environment technology*. Vol. 3. IEEE. 2009, pp. 7–10.

A Some posterior predictive

A.1 Caorle station

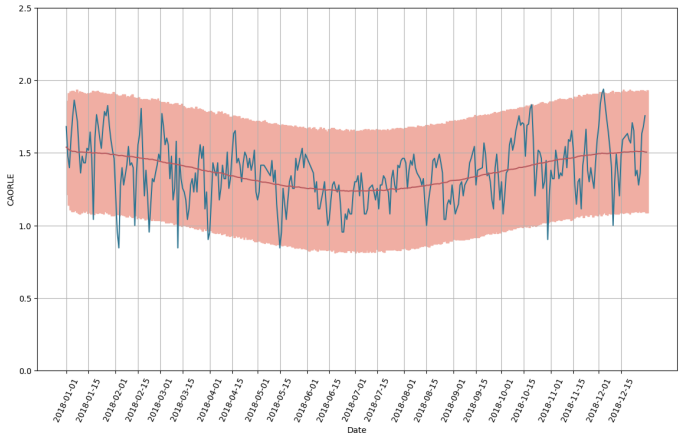


Figure 16. 95% C.I. ARIMA model

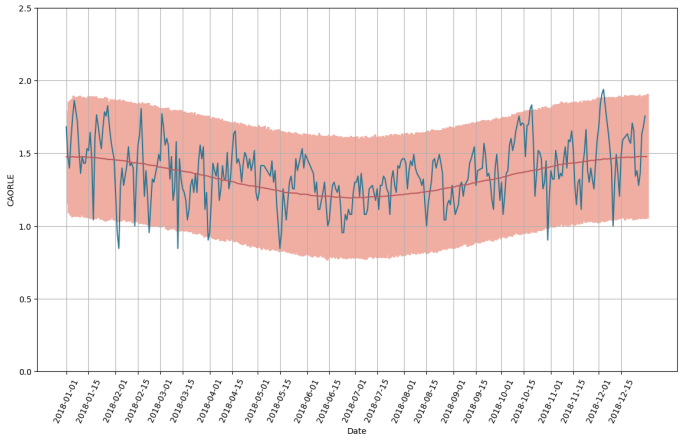


Figure 17. 95% C.I. regression model

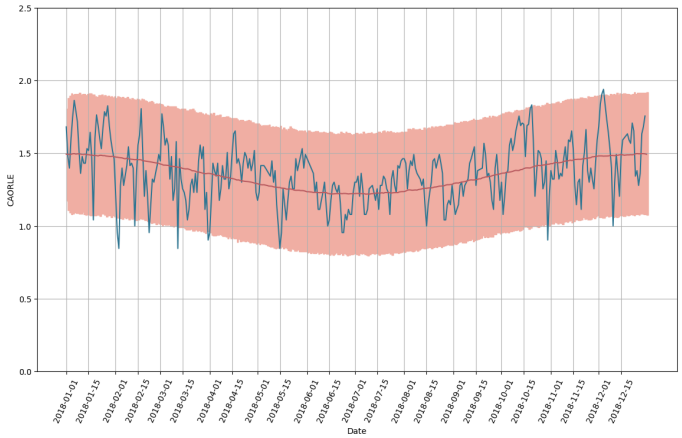


Figure 18. 95% C.I. spatial model

A.2 Savignano di Rigo station

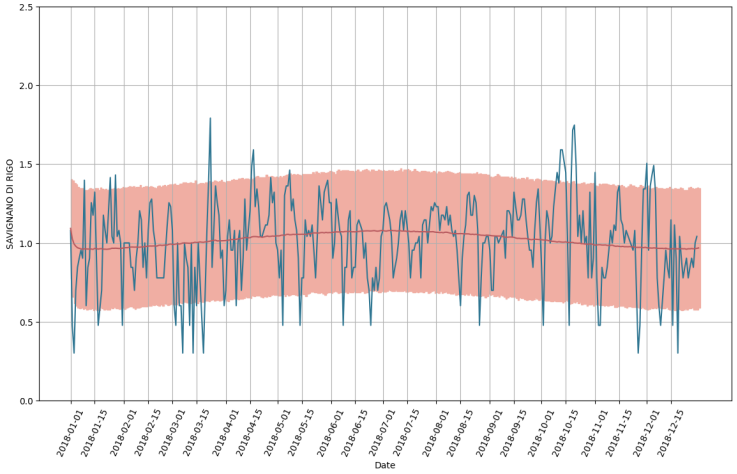


Figure 19. 95% C.I. ARIMA model

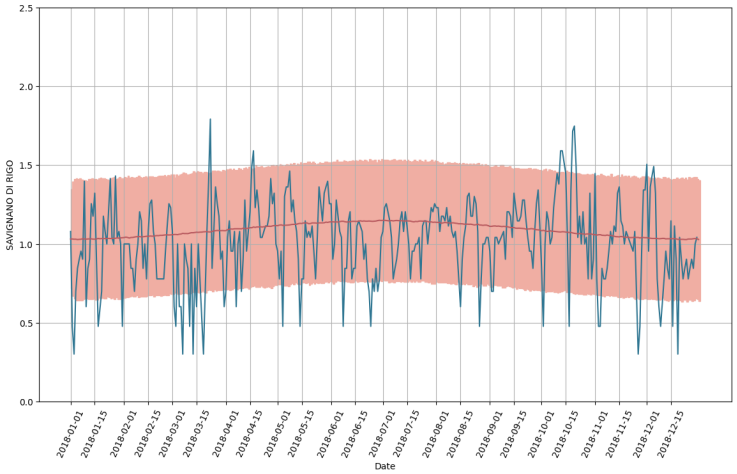


Figure 20. 95% C.I. regression model

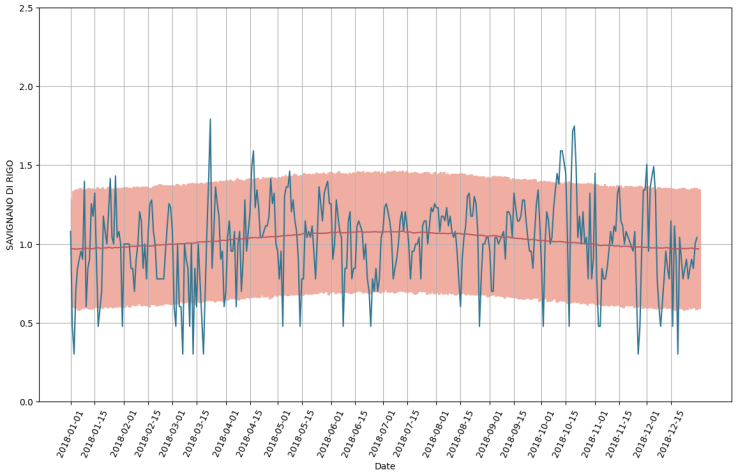


Figure 21. 95% C.I. spatial model

A.3 Febbio station

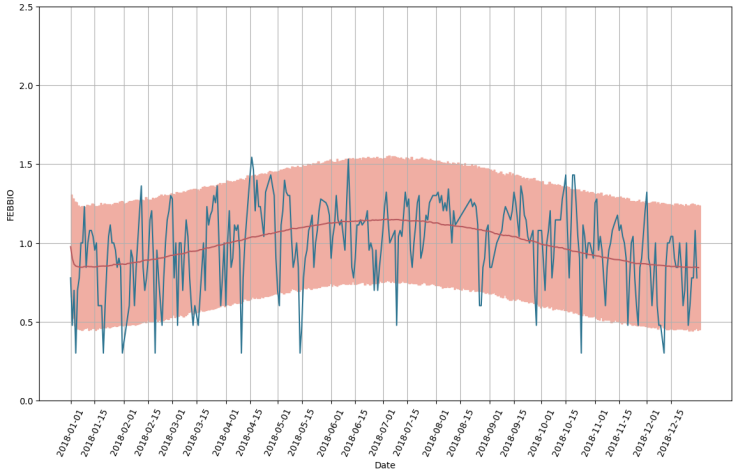


Figure 22. 95% C.I. ARIMA model

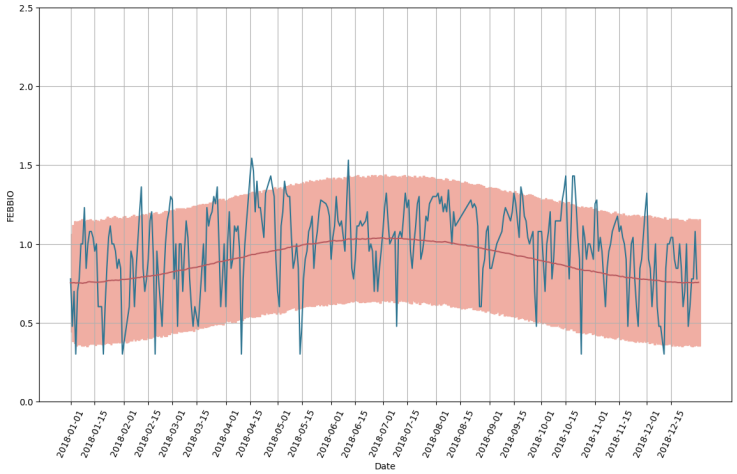


Figure 23. 95% C.I. regression model

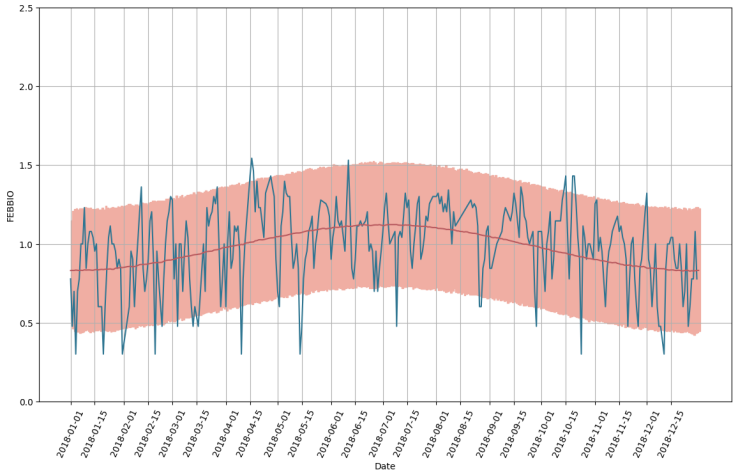


Figure 24. 95% C.I. spatial model

A.4 Parco Edilcarani station

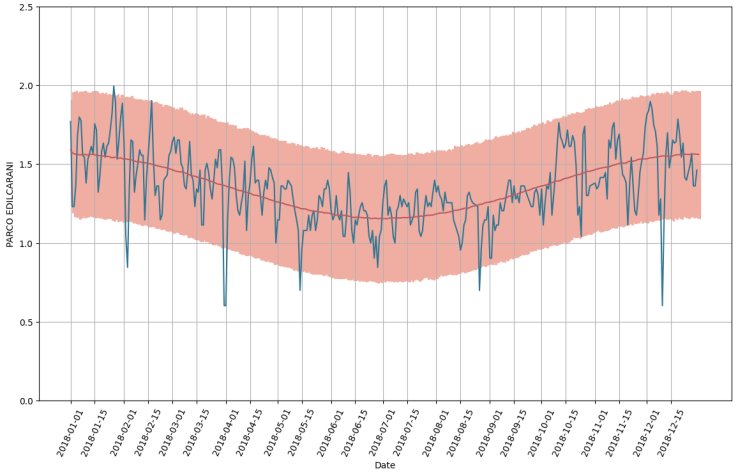


Figure 25. 95% C.I. ARIMA model

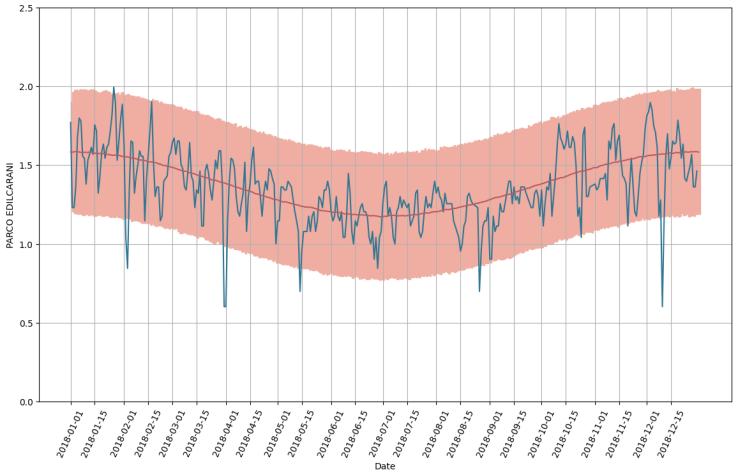


Figure 26. 95% C.I. regression model

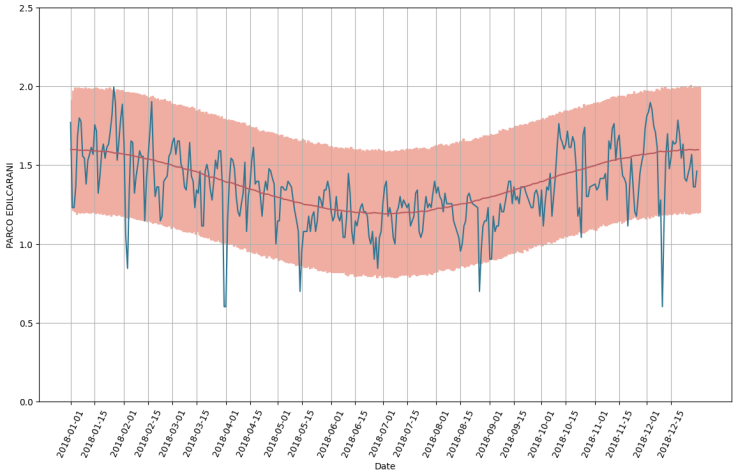


Figure 27. 95% C.I. spatial model

B Model Diagnostics

B.1 Multivariate ARIMA

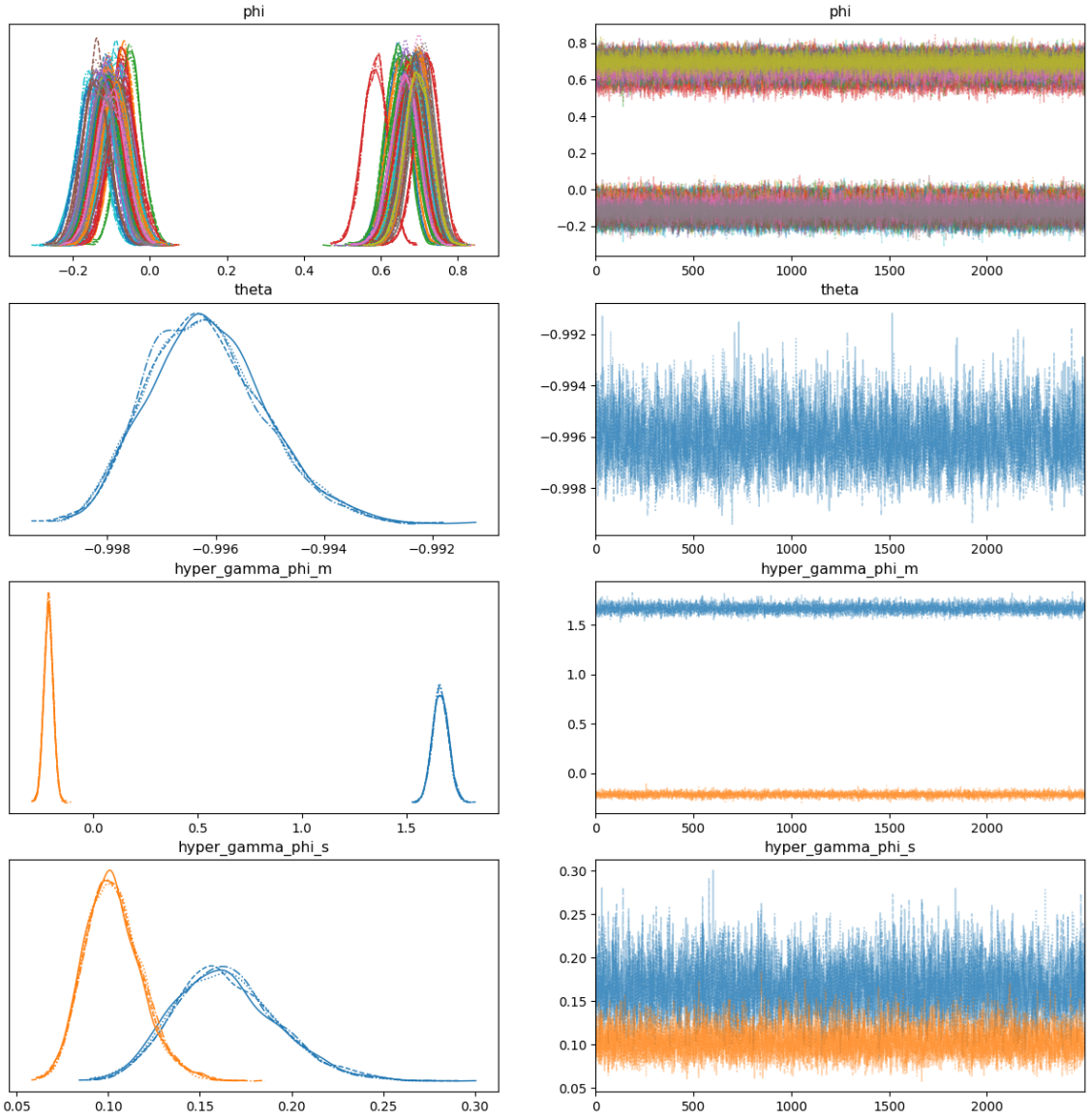


Figure 28. ARIMA parameters traceplot

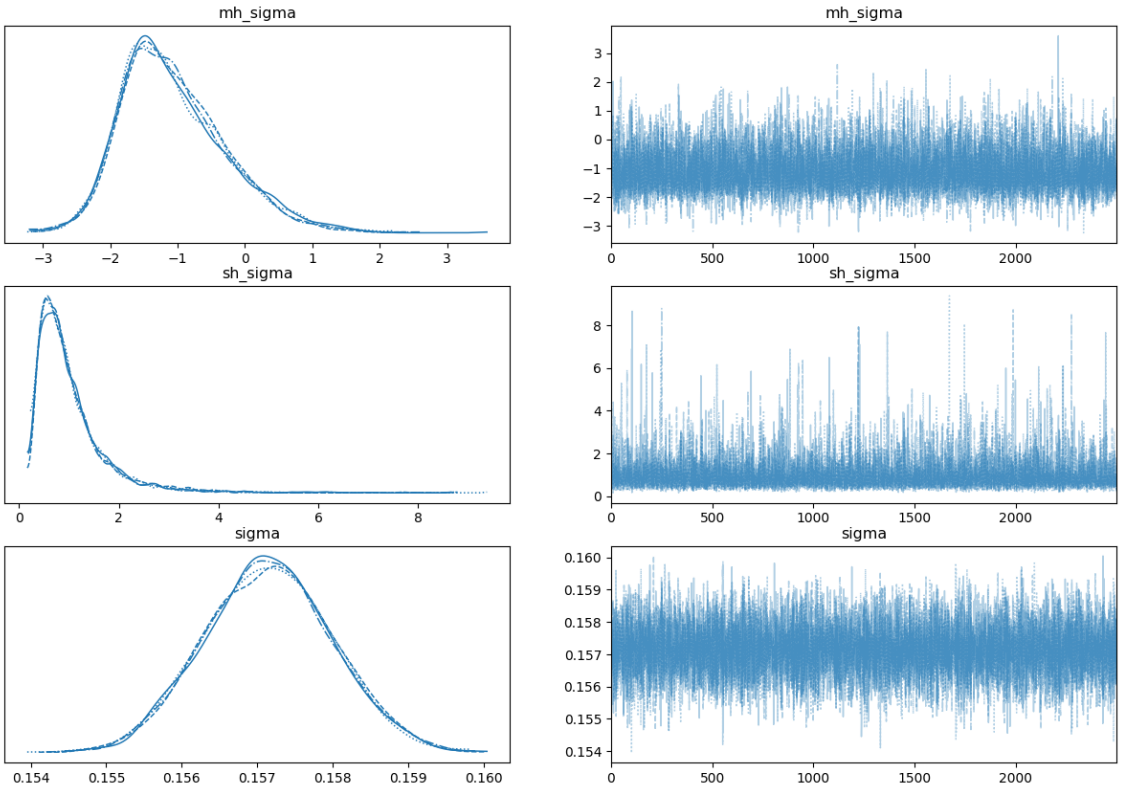


Figure 29. ARIMA residual traceplot

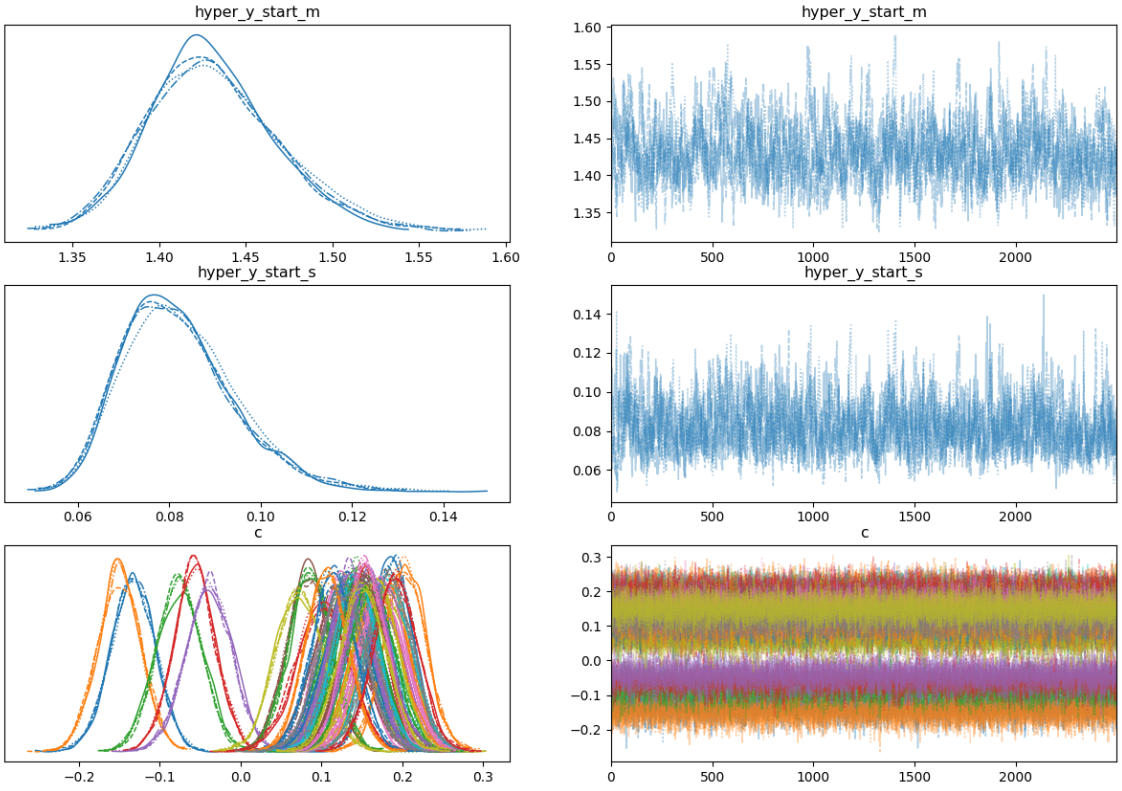


Figure 30. Starting conditions and cosine coefficient traceplot

B.2 Multivariate ARIMA with covariates

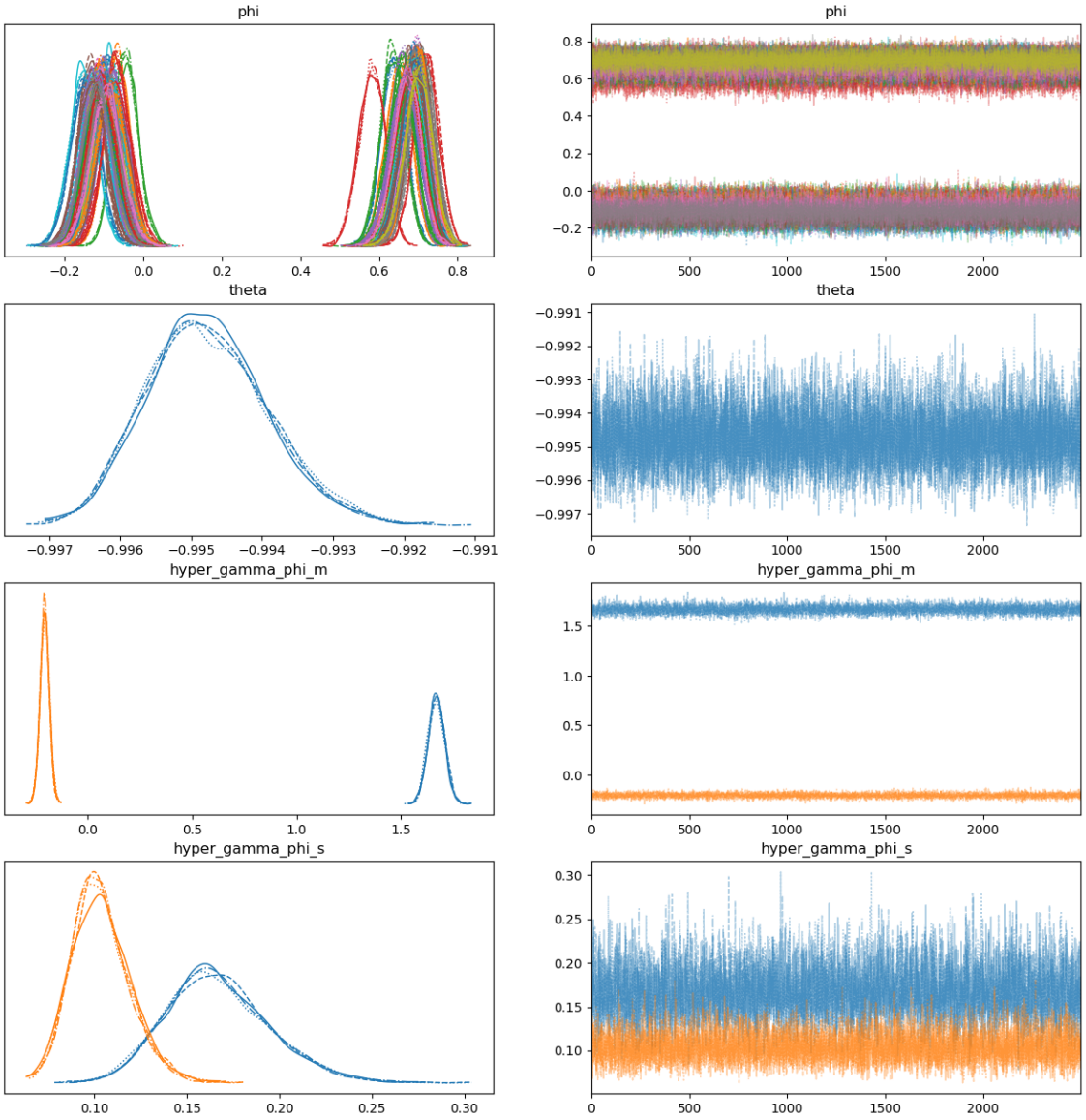


Figure 31. ARIMA parameters traceplot

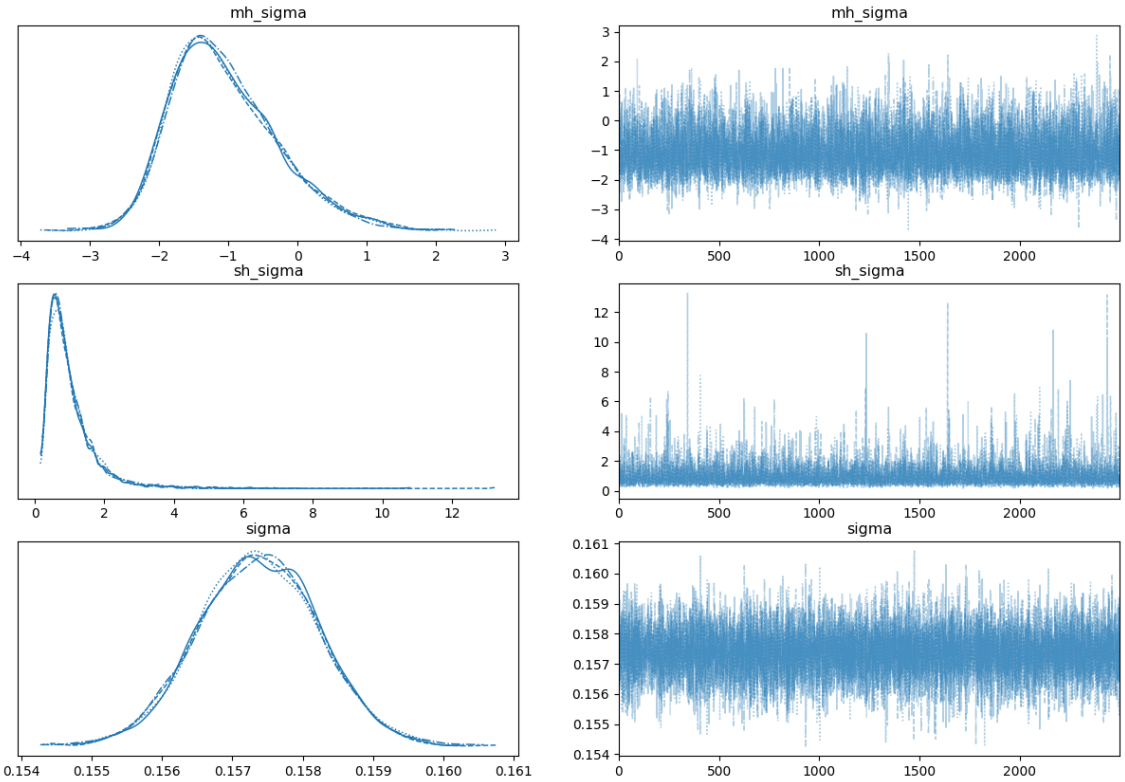


Figure 32. ARIMA residual traceplot

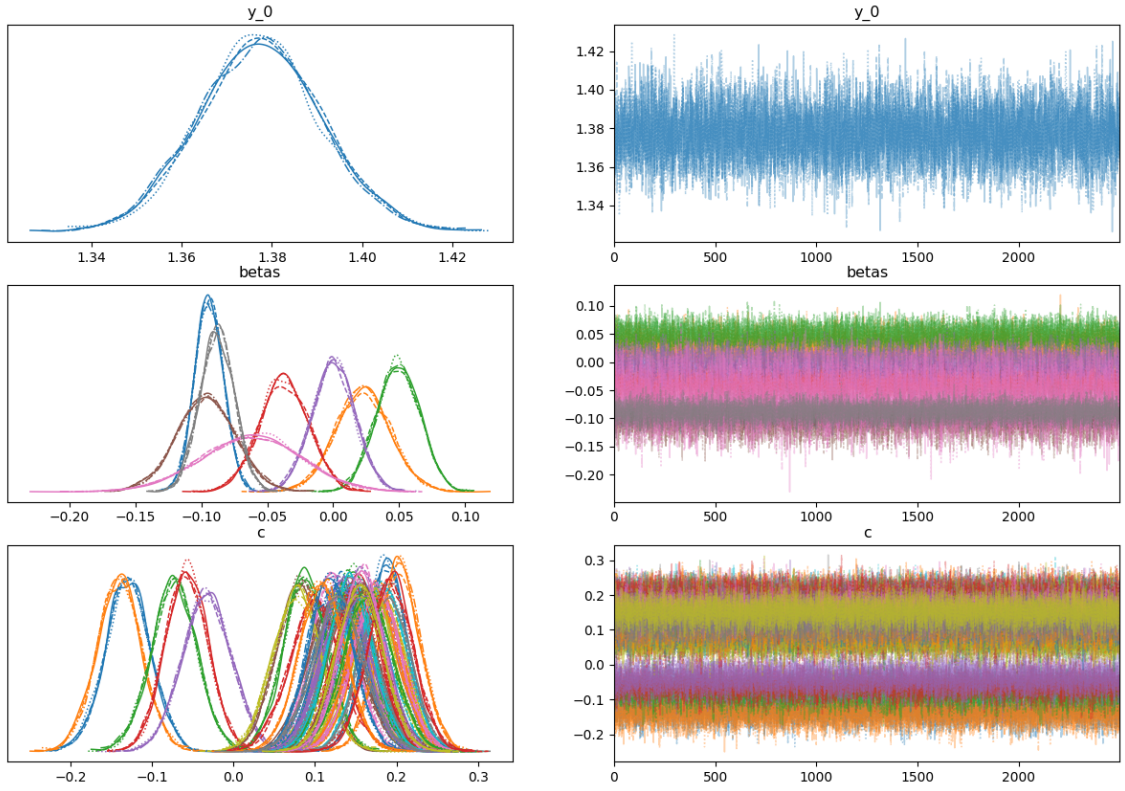


Figure 33. Starting condition, cosine and regression coefficient traceplot

B.3 Multivariate ARIMA with spatial Gaussian process

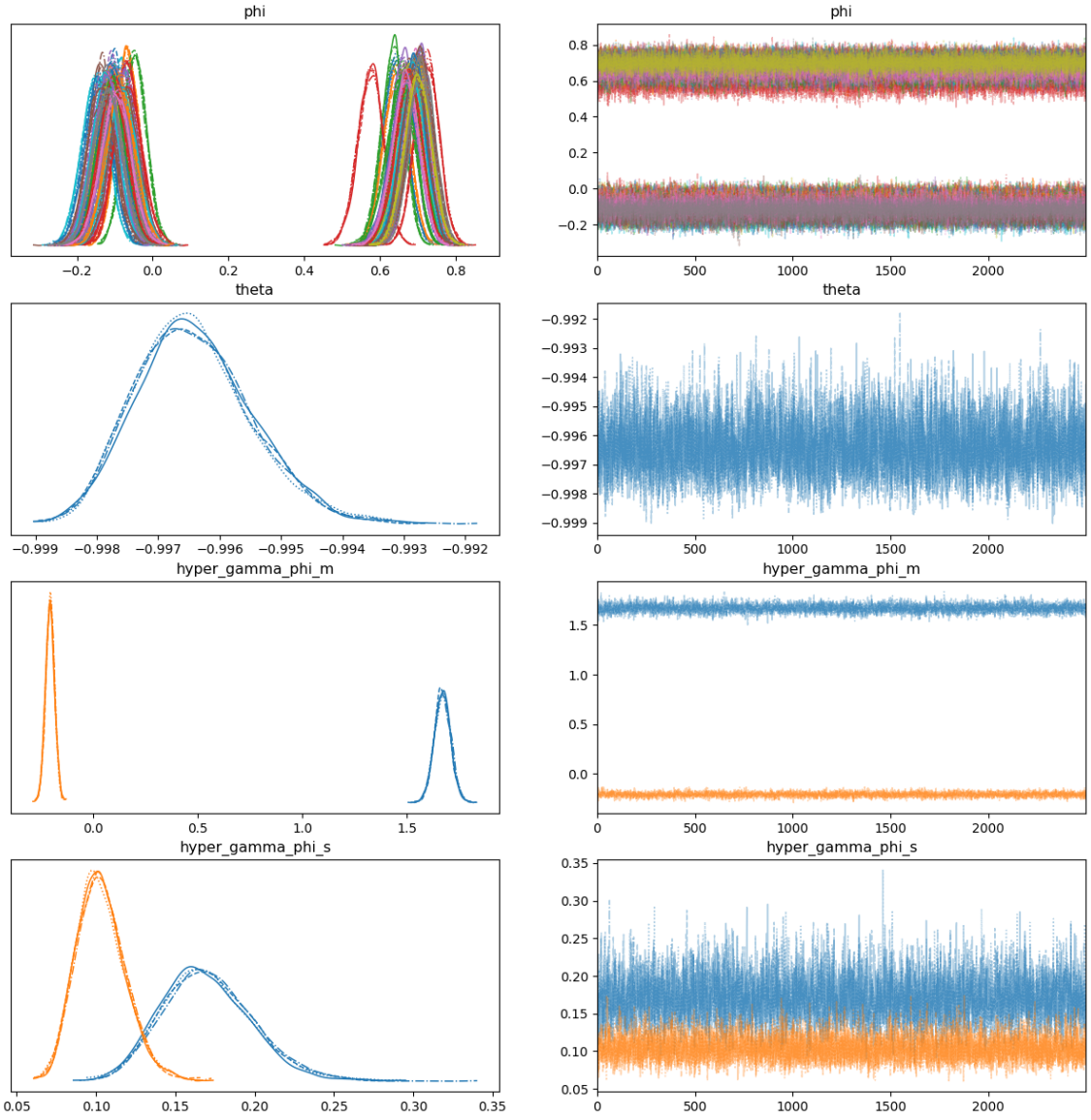


Figure 34. ARIMA parameters traceplot

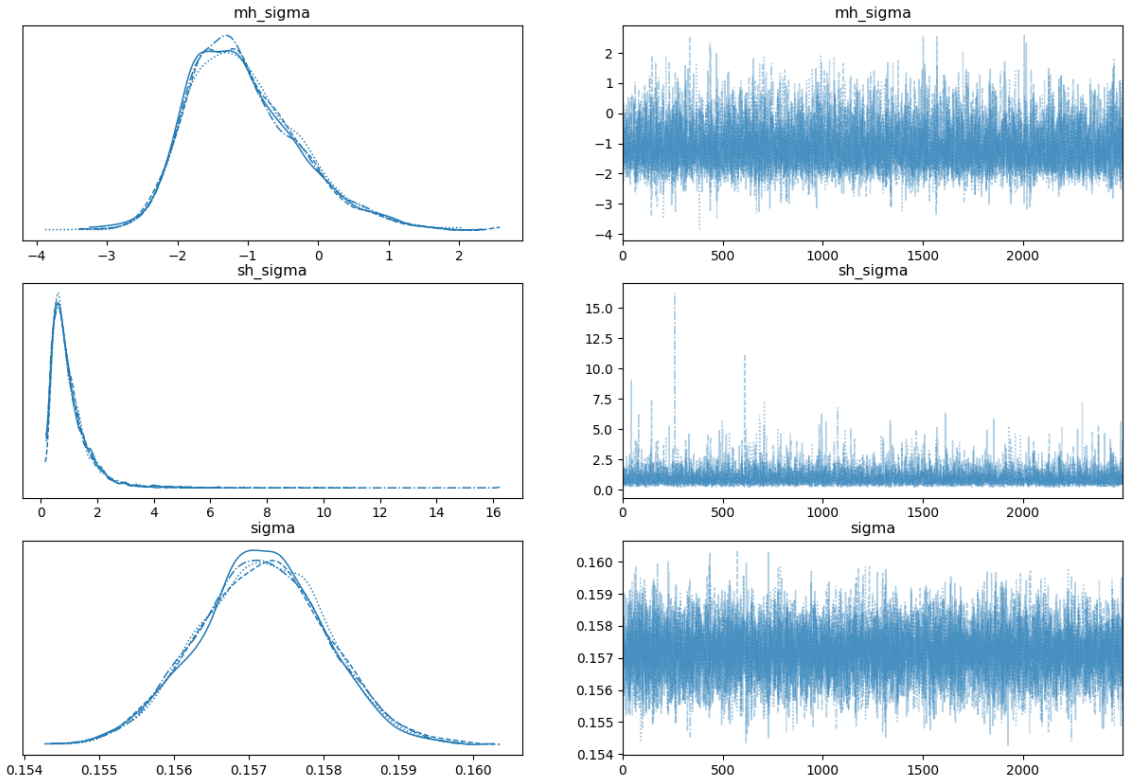


Figure 35. ARIMA residual traceplot

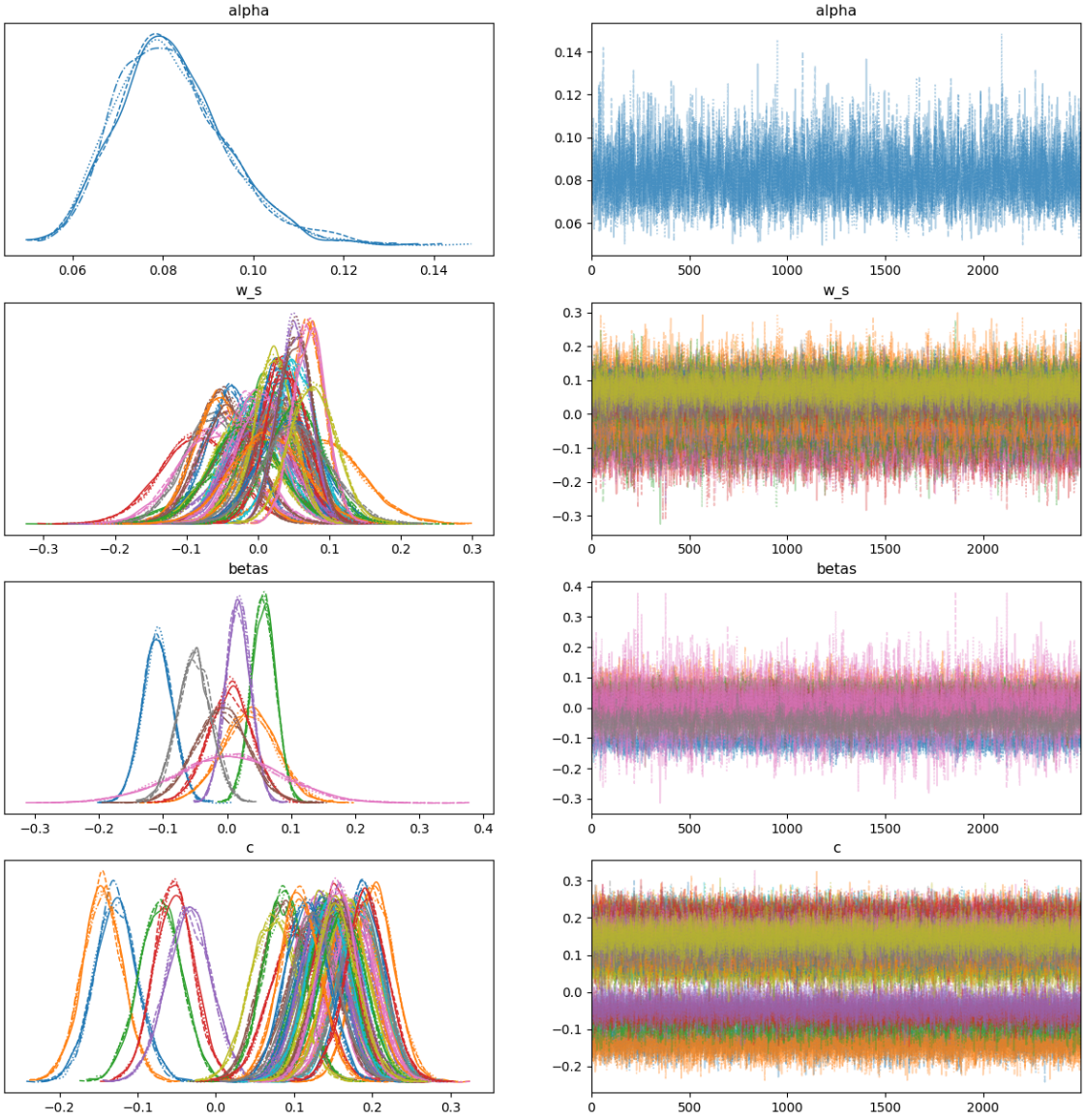


Figure 36. Spatial residuals, cosine and regression coefficient traceplot

C Missing data reconstruction example: 31 December 2018

As this datum is missing from all stations, it's well suited for showing the reconstruction across the various models:

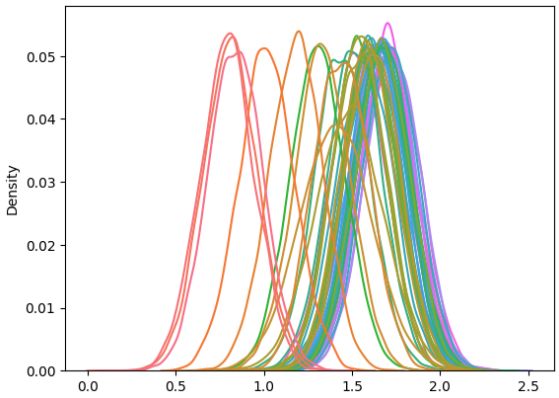


Figure 37. Multivariate ARIMA model

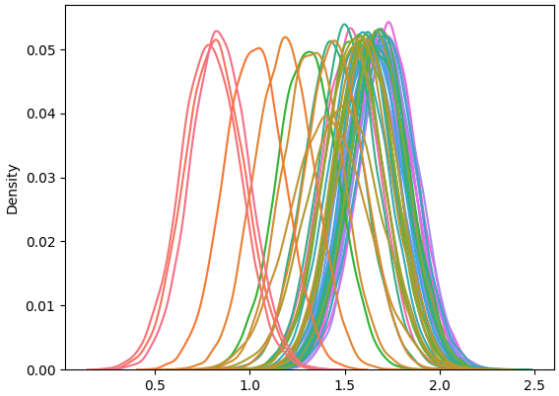


Figure 38. Multivariate ARIMA with covariates

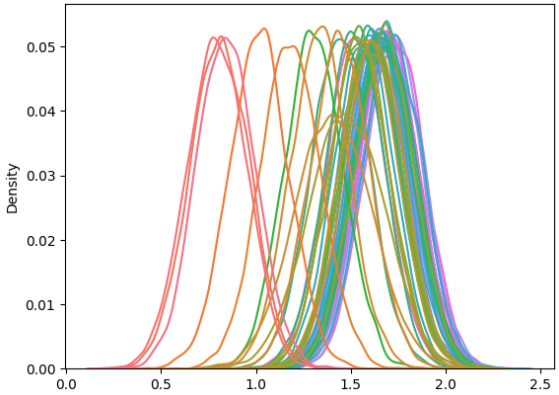


Figure 39. Multivariate ARIMA with spatial Gaussian process

D *STAN* Code

D.1 Model D

```
data {

  int<lower=1> T; // la lunghezza dell'intervallo temporale
  int<lower=1> S; // il numero di stazioni
  int<lower=1> reg; //numero di regressori utilizzati

  int<lower=1> p; // il grado della porzione autoregressiva
  int<lower=1> q; // il grado della porzione a media mobile

  matrix[T,S] y; // i dati osservati (al posto dei NaN si passa 1.0 per compatibilità,
                         //ma il valore non viene realmente utilizzato)
  matrix[S,reg] X;

  matrix<lower=0,upper=1>[T,S] is_missing;
                  // contiene un 1 se il dato corrispondente è mancante, altrimenti 0

  int<lower=0> missing_size; // contiene il numero di dati mancanti

  array[S] vector[2] coord;

}

transformed data {

  array[T+1] int u;
  array[missing_size] int<lower=1,upper=S> v;

  u = csr_extract_u(is_missing);
  v = csr_extract_v(is_missing);

  array[missing_size] int<lower=1,upper=T> u_mine;

  int index = 1;
  for (t in 1:T) {
    if (u[t+1]-u[t] > 0) {
      for (j in 1:(u[t+1]-u[t])) {
u_mine[index] = t;
index += 1;
      }
    }
  }

  row_vector[S] media_stazione;
  for (s in 1:S) {
    media_stazione[s] = mean(y[1:T,s]);
  }

  real rho = 0.05;

  vector[p+1] y_start = rep_vector(mean(media_stazione),p+1);

}
```

```
parameters {

  // Regressione
  vector[reg] betas;

  // ARMA
  vector[q] gamma_th; // "logit"(theta)
  array[p] row_vector[S] gamma_phi; // "logit"(phi)
  vector[p] hyper_gamma_phi_m;
  vector<lower=0>[p] hyper_gamma_phi_s;

  // deviazione standard di err e iperparametri associati
  real<lower=0> sigma;
  real mh_sigma;
  real<lower=0> sh_sigma;

  // dati mancanti
  vector[missing_size] w;

  // Coefficiente moltiplicativo del coseno
  vector[S] c;

  // Residui spaziali e deviazione standard del residuo spaziale
  vector[S] w_s;
  real<lower=0> alpha ;

}

transformed parameters {

  matrix[T,S] cos_of_day;
  for (s in 1:S) {
    cos_of_day[:,s] = c[s]*cos( 2*pi()*(cumulative_sum(rep_vector(1,T)) )/365 );
  }

  matrix[T,S] regres = rep_matrix((X*betas)',T);

  cov_matrix[S] H = gp_exp_quad_cov(coord,alpha,rho);
  matrix[T,S] spatial = rep_matrix(w_s',T);


  // THETA e phi sono ottenuti mediante:
  //            exp(gamma_th) - 1
  //   theta = --------------------
  //            exp(gamma_th) + 1
  // analogamente per phi, nelle righe a seguire
  vector<lower=-1,upper=1>[q] theta;
  array[p] row_vector<lower=-1,upper=1>[S] phi;
  for (j in 1:q) {
    theta[j] = (exp(gamma_th[j]) - 1)/(exp(gamma_th[j]) + 1);
  }
  for (j in 1:p) {
    phi[j] = (exp(gamma_phi[j]) - 1)./(exp(gamma_phi[j]) + 1);
  }

  // produce un vettore che contiene il dato
```

```
  // in corrispondenza dei dati noti, e contiene una variabile w oppure y_start
  // in corrispondenza dei dati mancanti
  // poi lo differenzia per produrre il vettore D (parte integrativa dell'ARIMA)
  matrix[T+p+1,S] right_y;
  right_y[1:(p+1), 1:S] = rep_matrix(y_start,S);
  right_y[(p+2):(T+p+1), 1:S] = y;
  for (k in 1:missing_size) {
    right_y[p+1+u_mine[k],v[k]] = w[k];
  }
  right_y[(p+2):(T+p+1), 1:S] = right_y[(p+2):(T+p+1), 1:S]
                                - cos_of_day - regres - spatial;

  matrix[T+p, S] D = right_y[2:(T+p+1), 1:S] - right_y[1:(T+p), 1:S];


  // ARIMA calcolo dei residui
  matrix[T,S] nu = rep_matrix(0,T,S);
  matrix[T,S] err;

  // AR VETTORIALIZZATO
  for (j in 1:p) {
    nu += rep_matrix(phi[j],T).*D[(1-j+p):(T-j+p), 1:S];
  }

  // MA
  err = D[(p+1):(p+T), 1:S] - nu;
  for (t in 1:T) {
    for (j in 1:q) {
      if ((t-j) > 0) {
err[t, 1:S] -= theta[j]*err[t-j, 1:S];
      }
    }
  }


}

model {


  // priors
  for (j in 1:p) {
    gamma_phi[j] ~ normal(hyper_gamma_phi_m[j], hyper_gamma_phi_s[j]);
  }
  gamma_th ~ normal(0,1);
  hyper_gamma_phi_m ~ normal(0,5);
  hyper_gamma_phi_s ~ inv_gamma(3,2);


  sigma ~ lognormal(mh_sigma, sh_sigma);
  mh_sigma ~ normal(0,1);
  sh_sigma ~ inv_gamma(3,2);


  w ~ normal(mean(media_stazione),1);
```

```
  c ~ normal(0,1);



  betas ~ normal(0,1);

  alpha ~ inv_gamma(6, 2);
  w_s ~ multi_normal(rep_vector(0,S), H);

  // likelihood
  for (s in 1:S) {
    err[1:T,s] ~ normal(0, sigma);
  }

}


generated quantities {

  matrix[T+p+1,S] y_post_pred_aux;
  matrix[T,S] err_post_pred;
  for (s in 1:S) {
    err_post_pred[1:T,s] = to_vector(normal_rng(rep_vector(0,T), sigma));
  }
  y_post_pred_aux[1:(p+1), 1:S] = rep_matrix(mean(media_stazione),(1+p),S);

  for (t in (p+2):(T+p+1)) {

    row_vector[S] mean_val = rep_row_vector(0,S);

    for (j in 1:p) {
      mean_val += phi[j].*(y_post_pred_aux[t-j, :] - y_post_pred_aux[t-j-1, :]);
    }

    for (j in 1:q) {
      if ((t-p-1)-j > 0) {
mean_val += theta[j]*err_post_pred[(t-p-1)-j, :];
      }
    }

    y_post_pred_aux[t,:] = y_post_pred_aux[t-1,:] + mean_val + err_post_pred[t-p-1,:];

  }


  matrix[T,S] y_post_pred = y_post_pred_aux[p+2:T+p+1,:] +
                            cos_of_day[:,:] + regres[:,:] + spatial[:,:];

  vector[S] annual_mean;
  vector[S] annual_max;
  vector[S] annual_median;
  vector[S] annual_days_over_threshold;
  array[S] int is_over_daily_limit;
  array[S] int is_over_annual_limit;
```

```
  for (s in 1:S) {
    annual_mean[s] = mean(y_post_pred[:,s]);
    annual_max[s] = max(y_post_pred[:,s]);
    annual_median[s] = quantile(y_post_pred[:,s],0.5);
    annual_days_over_threshold[s] = 0;
    for (t in 1:T) {
      annual_days_over_threshold[s] += (y_post_pred[t,s] > log10(50));
    }
    is_over_daily_limit[s] = (annual_days_over_threshold[s] > 35);
    is_over_annual_limit[s] = (annual_mean[s] > log10(40));
  }



  vector[T*S-missing_size] log_lik;
  int index_gen = 1;
  for (t in 1:T) {
    for (s in 1:S) {
      if (is_missing[t,s] == 0) {
log_lik[index_gen] = normal_lpdf(err[t,s] | 0, sigma);
index_gen += 1;
      }
    }
  }


  array[missing_size] int<lower=1,upper=T> missing_index_time = u_mine;
  array[missing_size] int<lower=1,upper=S> missing_index_station = v;

}
```