Bayesian spatiotemporal models for $PM_{2.5}$ in the Po valley

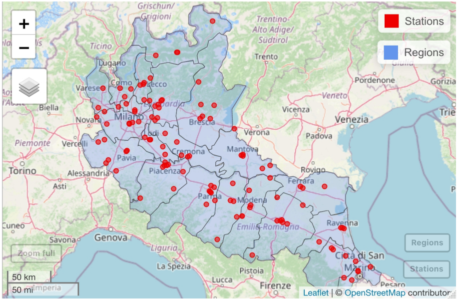
Bayesian Statistics course
Tutor: Alessandra Guglielmi, Michela Frigeri

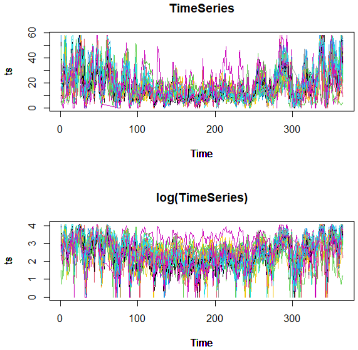
A.Y. 2022/2023

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

Table of contents

Overview





- 62 stations in Lombardia and Emilia-Romagna
- Looking for a model to monitor $PM_{2.5}$ level over space and time

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

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

Fourier:

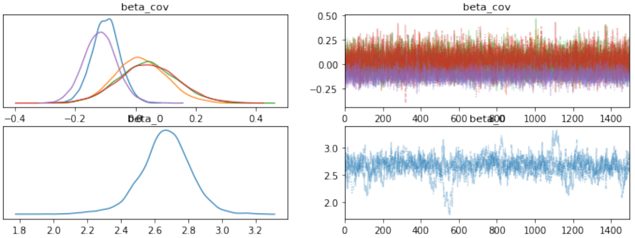
$$f(t) = \sum_{j \in J} a_j \; sin(j\omega t) + b_j \; cos(j\omega t)$$
$$a_j, b_j \overset{iid}{\sim} \mathcal{N}(0, 1)$$

ARIMA:

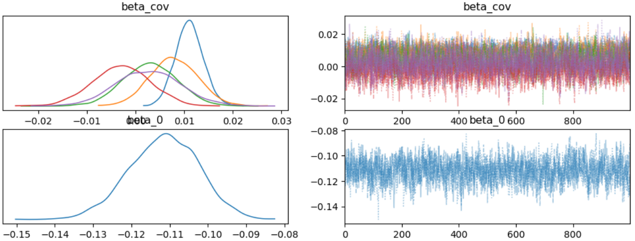
$$f_i(t) = (1 + \phi_1) y_i(t-1) + (\phi_2 - \phi_1) y_i(t-2)$$
$$+ (\phi_3 - \phi_2) y_i(t-3) - \phi_3 * y_i(t-4)$$
$$+ \theta_1 \epsilon_i(t-1) + \theta_2 \epsilon_i(t-2) + \mathbf{x}_i^T \beta$$
$$\epsilon_i = y_i(t) - f_i(t)$$
$$\phi_1, \phi_2, \phi_3, \theta_1, \theta_2 \overset{iid}{\sim} \mathcal{N}(0,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

Posterior: Covariates

Fourier

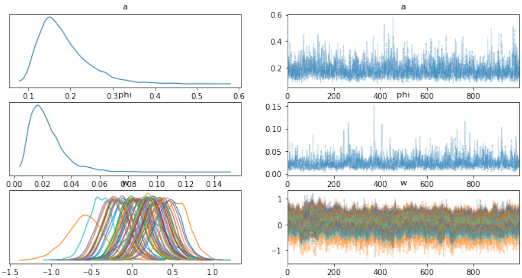


ARIMA

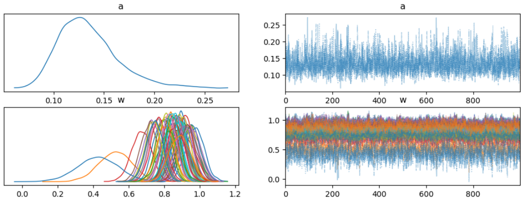


Posterior: Spatial Residuals

Fourier

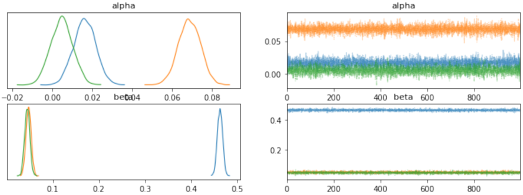


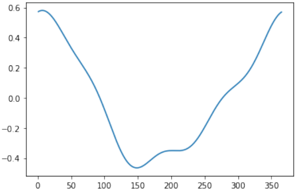
ARIMA



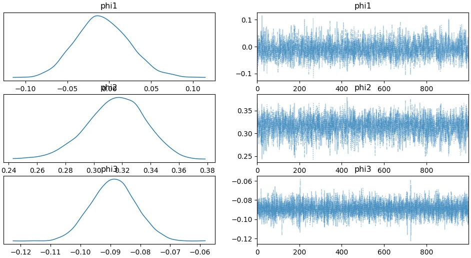
Posterior: Coefficients

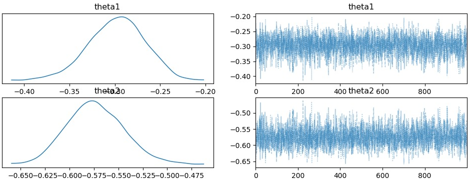
Fourier Coefficients





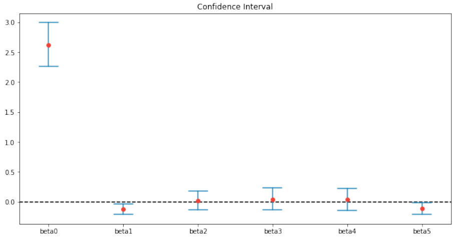
ARIMA Coefficients



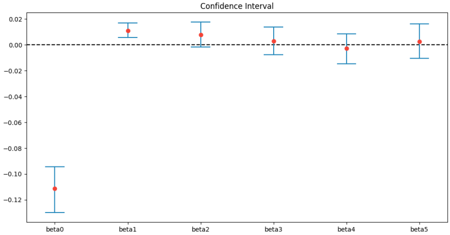


Credibility Intervals: Covariates and Spatial Residuals

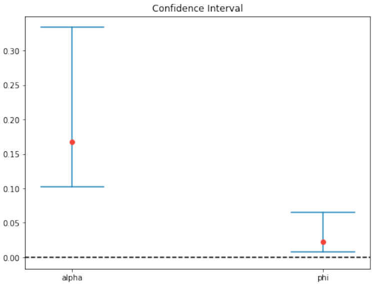
Fourier: Covariates



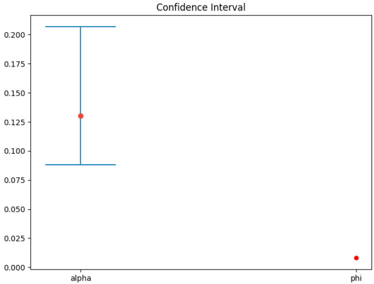
ARIMA: Covariates



Fourier: Spatial Residuals

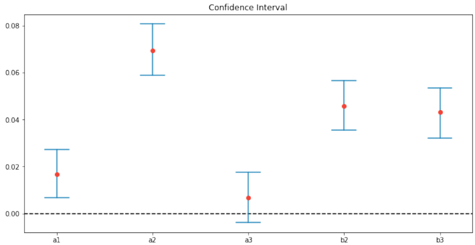


ARIMA: Spatial Residuals

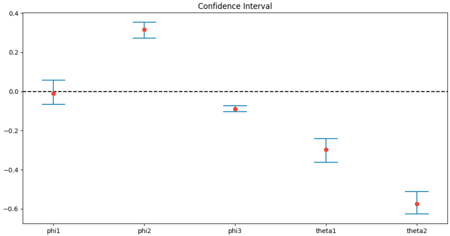


Credibility Intervals: Coefficients

Fourier: Coefficients



ARIMA: Coefficients

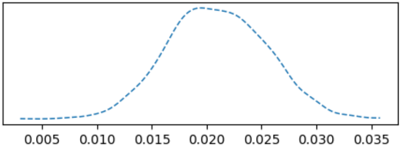


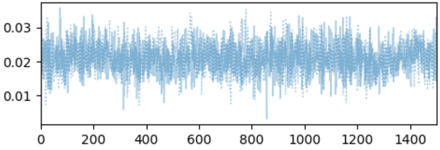
1. Combining the models: ARIMA with Fourier basis

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

$$f_i(t) = (1+\phi_1)y_i(t-1) + (\phi_2 - \phi_1)y_i(t-2) + (\phi_3 - \phi_2)y_i(t-3)$$
$$-\phi_3 * y_i(t-4) + \theta_1 \epsilon_i(t-1) + \theta_2 \epsilon_i(t-2) + \mathbf{x}_i^T \beta + a \cos(\omega t)$$

$$a \sim \mathcal{N}(0,1)$$

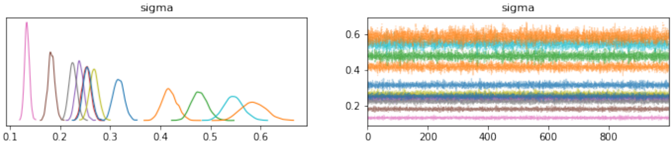




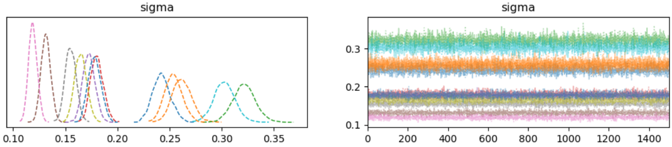
2. Monthly sigma

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

$$\sigma_m^2 \overset{iid}{\sim} InvGamma(3, 2) \quad m = 1, \ldots, 12$$



(a) Fourier



(b) Arima

5 indexes used for model choice:

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

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

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

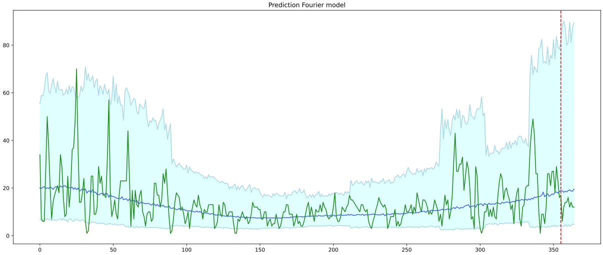
$$Measure_2 = \sum_{j=1}^{62} \sum_{t=356}^{365} \frac{|y_j(t) - \mathbb{E}(\hat{y}_j(t))|}{y_j(t)}$$

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

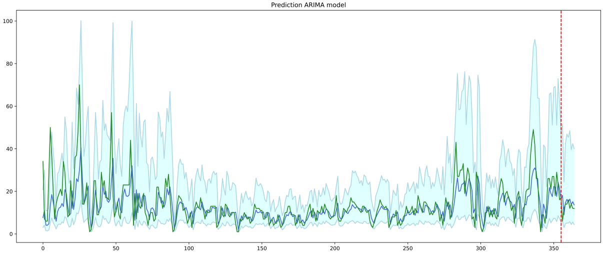
Model	LOO	WAIC	Measure 1	Measure 2	Measure 3
Basic Fourier	-18927.23	-19270.48	0.4213	99.2539	0.9709
Monthly Fourier	-18017.56	-18369.24	0.3882	101.7612	0.9854
Basic ARIMA	-14650.20	-14444.92	0.2781	58.9248	0.9838
ARIMA with cosine	-14638.57	-14437.04	0.2743	58.6859	0.9887
Monthly ARIMA with cosine	-14099.58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

Model choice

10 days prediction:



(a) Fourier prediction

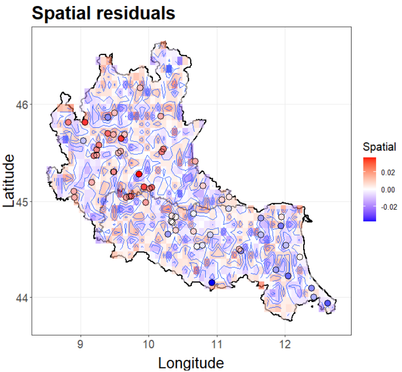


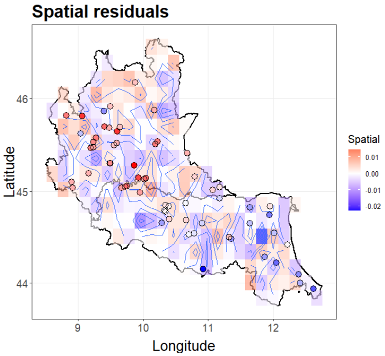
(b) ARIMA prediction

POLITECNICO MILANO 1863

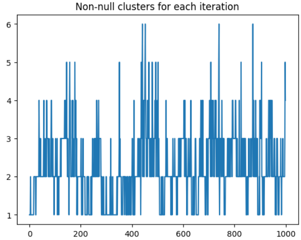
820 points

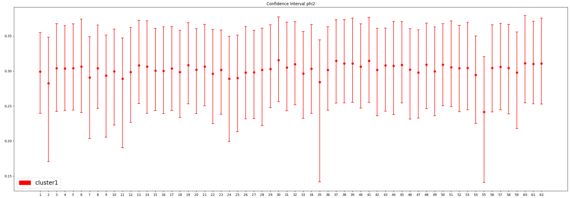
185 points





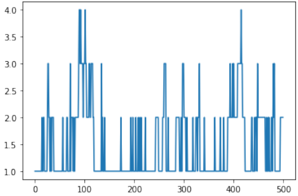
Cluster on ARIMA(3,1,2) model

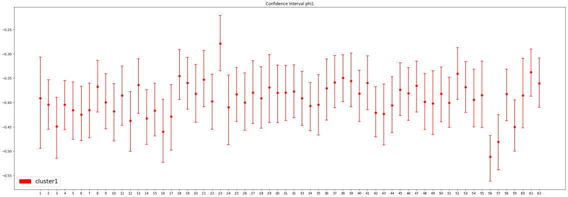


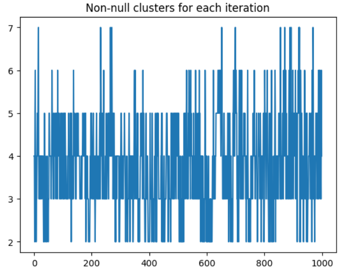


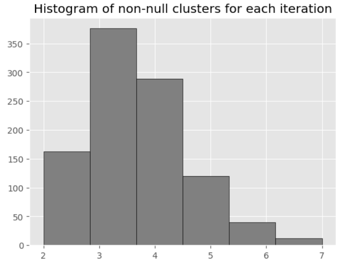
(a) AR(1)

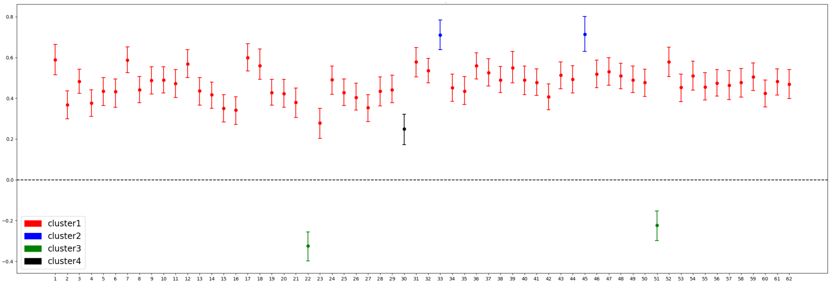
Cluster on AR(1) model





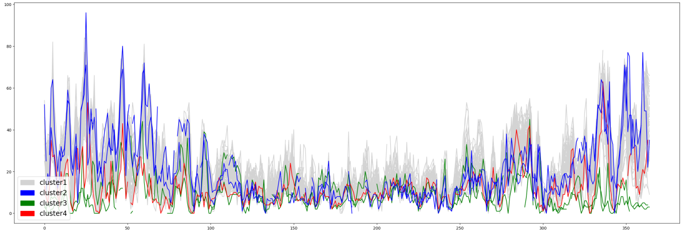

Non-null clusters for each iteration
7
6
5
4
3
2
0
200
400
600
800
1000


Histogram of non-null clusters for each iteration
350
300
250
200
150
100
50
0
2
3
4
5
6
7

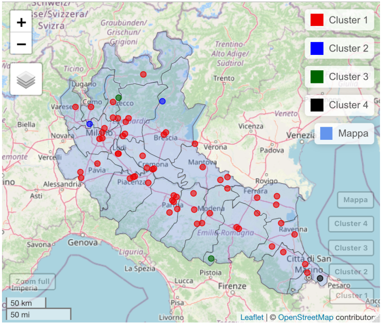

0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
cluster1
cluster2
cluster3
cluster4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62

Plot of the stations

Plot of the PM2.5 levels of the 62 stations, coloured according to the respective cluster



Map of the 62 stations, coloured according to the respective cluster



THANKS FOR YOUR ATTENTION!

We look forward to answering your questions

Extra: Finite Mixture Clustering

Finite Mixture Clustering for a general parameter z:

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

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

$$\eta_1 = v_1$$

with:

- $\mu_k \sim N(0, 10\sigma_k)$
- $\sigma_k^2 \sim InvGamma(a, b) \rightarrow \begin{cases} AR(1) : \sigma_k^2 \sim InvGamma(4.5, 0.001) \\ ARIMA(3,1,2) : \sigma_k^2 \sim InvGamma(3, 0.001) \\ Fourier : \sigma_k^2 \sim InvGamma(3, 0.003) \end{cases}$
- $v_k \sim Beta(1, 2)$
- C=10

Posterior expectation of the Binder's loss function:

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