POLITECNICO MILANO 1863

Data fusion for estimating ambient air pollution with spatial disalignment

Barbato Giulia Cosi Michele Del Basso Martina
Esposito Chiara Frabetti Alessandro Rizzo Marco

Mentors: Francesco Denti, Veronica J. Berrocal
A.Y. 2022/2023

Contents

1 Introduction

Nowadays, researchers in environmental health, climatology and ecology are increasingly facing issues regarding data analysis due to the highly multivariate, spatially referenced and temporally correlated available information. Weather patterns are ones of the most highlighted topics of the last years. For this reason, a lot of modern researches are concerned on air quality monitoring (Berrocal et al. [2010],Warren et al. [2021]). For instance, health professionals who collect data of cervical and lung cancer in a particular region and year, together with life and air/health quality information (Eckel et al. [2016]) , have to tackle some challenges: from the modeling of the correlation structure and the estimation of the model parameters, to the prediction of the variables of interest at unobserved time or place.

Our project focuses on investigating the spatial distribution of $PM_{2.5}$, a particular air pollutant representing a complex mixture of solid and liquid particles that can be generated from anthropogenic sources as vehicle emissions and power generation. High ambient $PM_{2.5}$ have been consistently associated with increased risks of various adverse health outcomes. Hence the need to closely monitor $PM_{2.5}$ evolution.

Recent statistical advances allow to obtain improved estimates of air pollution concentrations and appropriate measures of uncertainty, there have been developed advanced statistical techniques for combining multiple sources of air pollution information at different spatial scales.

The baseline measuring tools, in environmental data collection, are numerical models and monitoring networks. Considering a geographical area divided into uniform cells, the first source provides predictions at the level of grid cells, while the second gives measurements at points. For this reason, the numerical models spatially cover the whole area of interest, while the network of monitors are sparsely collected in space. Thus, even if they are negatively affected by missing data, they provides the 'true' value. Accommodating the spatial misalignment between this two sources is one of the main challenges to tackle with a view to evaluate and to predict the phenomenon under investigation.

Data fusion is a possible way to overcome this problem by modelling and predicting pollutant concentrations at locations and times that lack monitoring measurements. Data fusion techniques typically use directly measured concentrations in combination with alternative pollutant estimates, which have improved spatial coverage but may be biased and produced at different spatial scales (e.g., grid averages).

2 Data

2.1 Theoretical data

Our model is based on two types of data with different spatial scales: pointwise data and areal data.

The former ones are extracted from the air quality system (AQS) maintained by the U.S. Environmental Protection Agency (EPA) and include the latitude and longitude of each monitoring location along with measured concentrations on each day where the monitor was active for $PM_{2.5}$. More specifically, monitor measurements are obtained at fixed monitors. They refer to the daily average concentration of a pollutant at sparse spatial locations.

The latter ones are comparable daily estimates of each pollutant produced by the Community Multiscale Air Quality (CMAQ) modeling system obtained from the U.S. EPA.

Whereas the AQS monitoring network can be sparse in space and time, CMAQ estimates have excellent space-time coverage. They can be beneficial when evaluating different air pollution scenarios (e.g., new emission regulations). We underline that the CMAQ concentrations are outcomes of environmental models and not actual measurements. Therefore, they can be biased.

For further development, it could be useful to include a different type of points measurement data: remote sensing data. They are obtained by instruments aboard satellites. The instruments' measurements, related to light/radar reflections, are then transformed into proxy variables related to pollution levels. Remote sensing data are snapshots across the proxy variable's satellite orbit at a given time. There is also the possibility of accessing different types of remote sensing data generating different proxy variables (from a stationary satellite and an orbiting satellite) leading also to temporal disalignment.

2.2 Simulated Data

In this section, we will provide results on simulated data (designed as we will explain below) for our complete model.

2.3 Real Data

Moreover in this section we will provide results for the complete model on real data. the used dataset is based on two different ones: *Obspm2.5concNe2018training* and *US City Population Densities* merged with *US 2020 Census Cities Populations and Coordinates.*

3 The Model

For our project, we refer to the model written in Berrocal et al. [2010].
Following the aforementioned paper, we denote with $Y(\mathbf{s})$ the square root of the observed ozone concentration at a point $\mathbf{s}$ of the area under consideration. Instead, since the CMAQ output is given in terms of averages over 12 km grid cells, with $x(B)$ we denote the square root of the numerical model output over grid cell B.
Each point $\mathbf{s}$ is associated with the 12 km CMAQ grid cell B in which it lies. So, all the points $\mathbf{s}$ falling in the same 12 km square region are assigned to the same CMAQ output value.
The authors developed a model to relate the observed data to the CMAQ output as follows: for each $\mathbf{s}$ in B, they assume that

$$Y(\mathbf{s}) = \underbrace{\tilde{\beta}_0(\mathbf{s})}_{\hookrightarrow \beta_0 + \beta_0(\mathbf{s})} + \underbrace{\tilde{\beta}_1(\mathbf{s})}_{\hookrightarrow \beta_1 + \beta_1(\mathbf{s})} x(B) + \epsilon(\mathbf{s}), \quad \epsilon(\mathbf{s}) \overset{\text{ind}}{\sim} \mathrm{N}\left(0, \tau^2\right) \tag{1}$$

where $\epsilon(\mathbf{s})$ is a white noise process with nugget variance τ^2.

Moreover, β_0 and β_1 represent the overall additive and multiplicative bias of the CMAQ model, while $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are local adjustments to the additive and multiplicative component, respectively. The spatially-varying coefficients $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are in turn modeled as bivariate mean-zero Gaussian spatial processes using the method of coregionalization. Therefore, they suppose that there exist two mean-zero unit-variance independent Gaussian processes $w_0(\mathbf{s})$ and $w_1(\mathbf{s})$ such that:

$$\mathrm{cov}\left(w_j(\mathbf{s}), w_j\left(\mathbf{s}'\right)\right) = \exp\left(-\phi_j \left|\mathbf{s} - \mathbf{s}'\right|\right), \tag{2}$$

where ϕ_j is the spatial decay parameter for Gaussian process $w_j(\mathbf{s}), j = 0, 1$ (See Appendix B). Moreover,

$$\begin{pmatrix} \beta_0(\mathbf{s}) \\ \beta_1(\mathbf{s}) \end{pmatrix} = \mathbf{A} \begin{pmatrix} w_0(\mathbf{s}) \\ w_1(\mathbf{s}) \end{pmatrix} \tag{3}$$

where the unknown $\mathbf{A}$ matrix in Equation (3) can be assumed to be lower-triangular to tackle the identifiability problem. Moreover, they note that this problem vanishes when they have multiple spatial location $\mathbf{s}$'s within a given B since the associated $Y(\mathbf{s})$'s will vary over the $\mathbf{s}$'s in B.

They complete the specification of the Bayesian hierarchical model with the following prior specifications: they use a bivariate normal distribution for the overall bias terms β_0 and β_1, lognormal distributions for the two diagonal entries a_{00} and a_{11} of the coregionalization matrix A, a normal distribution for the off-diagonal entry, a_{10}, of A, and an inverse gamma distribution for τ^2. Since it is not possible to estimate consistently all of the covariance parameters, under weak prior specifications they find weak identifiability of these parameters in

MCMC chains. Hence, they perform a grid search to estimate the spatial decay parameters ϕ_j , $j = 0, 1$. Therefore, they use discrete uniform priors on m values for the decay parameters ϕ_j, $j = 0, 1$.

Overall, in our project we follow this model. For simplicity, we will assume the coregionalization matrix to be fixed.

4 Road map and goal

Our goal is obtaining estimates of the daily average concentration of a pollutant across a region (e.g., state of California) by combining the data sources.
For better understanding the role of coregionalization matrix we proceed step-by-step using increasingly complicated models. Before using the whole real data – which are computationally expensive – we consider the simulated data.

The following list summarizes the steps we performed during the project:

1. Derive and implement the full conditionals for the model described in Section 2
2. Construct a grid with simulated data
3. Implement a Gibbs Sampler for our parameters
4. Make prediction using Kriging algorithm
5. Fit the model with real data

5 Full conditional distributions

The parameters vector of our model is

$$\boldsymbol{\theta} = (\beta_0, \beta_1, \omega_0(\mathbf{s}), \omega_1(\mathbf{s}), \tau^2, a_{00}, a_{10}, a_{11}, \phi_0, \phi_1)$$

Hence, the posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \mathbf{y}) &\propto \pi(\mathbf{y}, \beta_0, \beta_1, \omega_0(\mathbf{s}), \omega_1(\mathbf{s}), \tau^2, a_{00}, a_{10}, a_{11}, \phi_0, \phi_1) \\ &\propto \pi(\mathbf{y} \mid \beta_0, \beta_1, \omega_0(\mathbf{s}), \omega_1(\mathbf{s}), \tau^2) \cdot \pi(\boldsymbol{\beta} \mid \mathbf{b}, \Sigma_\beta) \\ &\cdot \pi(\tau^2 \mid a, b) \cdot \pi(\omega_0(\mathbf{s}) \mid b_{0s}, \Sigma_{0s}) \cdot \pi(\omega_1(\mathbf{s}) \mid b_{1s}, \Sigma_{1s}) \\ &\cdot \pi(\phi_0 \mid n_0) \cdot \pi(\phi_1 \mid n_1)\end{aligned}$$

It is clear that the joint posterior density is analytically intractable. For this reason, our posterior inferences are based on MCMC simulation and, in detail on a Gibbs Sampler algorithm. In order to implement it, we first derive all the full conditional distributions.

5.1 Full conditional of $\boldsymbol{\beta}$

Given the model

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})x(B) + \epsilon(\mathbf{s}) \quad \text{with}\, \epsilon(\mathbf{s}) \sim N\left(0, \tau^2\right),$$

that can be written in the extended form

$$Y(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s}) + \beta_1 x(B) + \beta_1(\mathbf{s})x(B) + \epsilon(\mathbf{s})$$

Denoting with $\boldsymbol{\theta}_y = (\beta_0, \beta_1, \omega_0(\mathbf{s}), \omega_1(\mathbf{s}), \tau^2)$, we can state that

$$Y(\mathbf{s})|\boldsymbol{\theta}_y \sim N_n\left(\tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})x(B), \tau^2\mathbf{I}_n\right)$$

Thus, defining $\mathbf{y} = Y(\mathbf{s})$ to simplify the notation

$$\mathbf{y}_i = \mathbf{x}_i\tilde{\boldsymbol{\beta}} + \epsilon_i,\ i = 1, \ldots, n \rightarrow \mathbf{y}|\boldsymbol{\theta}_y \sim N_n\left(X\tilde{\boldsymbol{\beta}}, \tau^2\boldsymbol{I}_n\right)$$

where X is the design matrix with rows given by

$$\mathbf{x}_i = (1, x_i(B)),\ i = 1, \ldots, n$$

Moreover, defining $\mathbf{z}$ as

$$\mathbf{z} = \mathbf{y} - \beta_0(\mathbf{s}) - \beta_1(\mathbf{s})x(B),$$

with

$$\begin{aligned}\beta_0(\mathbf{s}) &= a_{00}w_0(\mathbf{s}), \\ \beta_1(\mathbf{s}) &= a_{10}w_0(\mathbf{s}) + a_{11}w_1(\mathbf{s})\end{aligned}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

it is possible to derive the following expression for the likelihood of $\mathbf{y}$

$$L(\mathbf{y} \mid \boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2} \frac{(\mathbf{z} - X\boldsymbol{\beta})^T (\mathbf{z} - X\boldsymbol{\beta})}{\tau^2} \right\}$$

Calling the MLE of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{z}$$

we can notice that

$$(\mathbf{z} - X\boldsymbol{\beta})^T (\mathbf{z} - X\boldsymbol{\beta}) = (\mathbf{z} - X\boldsymbol{\beta} + X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}})^T (\mathbf{z} - X\boldsymbol{\beta} + X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}) =$$
$$\left(\mathbf{z} - X\hat{\boldsymbol{\beta}}\right)^T (\mathbf{z} - X\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T X^T X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

Since the first term does not depend on $\boldsymbol{\beta}$, we can derive a clearer expression of the likelihood

$$\exp\left\{ -\frac{1}{2} \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T X^T X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{\tau^2} \right\}.$$

Furthermore, given the prior

$$\boldsymbol{\beta} \sim N_2\left(\mathbf{b}, \Sigma_\beta\right)$$

$$\pi(\boldsymbol{\beta}) \propto \frac{1}{\sqrt{\det \Sigma_\beta}} \cdot \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{b})^T \Sigma_\beta^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right\}$$

the posterior of $\boldsymbol{\beta}$ given all the rest, is

$$\pi\left(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\theta}_{-\beta}\right) \propto \exp\left\{ -\frac{1}{2} \left[\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T X^T X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{\tau^2} + (\boldsymbol{\beta} - \mathbf{b})^T \Sigma_\beta^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right] \right\},$$

evolving the calculations and defining $\Sigma^{-1} = X^T X / \tau^2$

$$\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}^T \Sigma^{-1} \hat{\boldsymbol{\beta}} - 2\boldsymbol{\beta}^T \Sigma^{-1} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T \Sigma^{-1} \hat{\boldsymbol{\beta}} + \boldsymbol{\beta}^T {\Sigma_\beta}^{-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T {\Sigma_\beta}^{-1} \mathbf{b} + \mathbf{b}^T \Sigma_\beta^{-1} \mathbf{b} \right] \right\}$$

$$\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}^T \underbrace{\left(\Sigma^{-1} + \Sigma_\beta^{-1}\right)}_{M} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \underbrace{\left(\Sigma^{-1} \hat{\boldsymbol{\beta}} + \Sigma_\beta^{-1} \mathbf{b}\right)}_{\mathbf{d}} \right] \right\}.$$

Thus, rearranging the terms

$$\pi\left(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\theta}_{-\beta}\right) \propto \exp\left\{ -\frac{1}{2} \left[\left(\boldsymbol{\beta} - M^{-1}\mathbf{d}\right)^T M \left(\boldsymbol{\beta} - M^{-1}\mathbf{d}\right) - \mathbf{d}^T M^{-1} \mathbf{d} \right] \right\}$$

To conclude, the final form of the posterior is

$$\begin{aligned}\pi\left(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\theta}_{-\beta}, X\right) &\sim N_2\left(M^{-1}\mathbf{d}, M^{-1}\right) \\ &\sim N_2\left(\left(\Sigma^{-1}+\Sigma_\beta^{-1}\right)^{-1}\left(\Sigma^{-1}\hat{\boldsymbol{\beta}}+\Sigma_\beta^{-1}\mathbf{b}\right),\left(\Sigma^{-1}+{\Sigma_\beta}^{-1}\right)^{-1}\right) \\ &\sim N_2\left(\left(\frac{X^TX}{\tau^2}+\Sigma_\beta^{-1}\right)^{-1}\left(\frac{X^TX}{\tau^2}\hat{\boldsymbol{\beta}}+\Sigma_\beta^{-1}\mathbf{b}\right),\left(\frac{X^TX}{\tau^2}+\Sigma_\beta^{-1}\right)^{-1}\right)\end{aligned}$$

5.2 Full conditional of τ^2

Given the prior:

$$\tau^2 \sim IG(a,b)$$

$$\pi(\tau^2) \sim \frac{b^a}{\Gamma(a)}\left(\frac{1}{\tau^2}\right)^{a+1}\exp\left(\frac{-b}{\tau^2}\right)$$

the posterior of τ^2 is proportional to

$$\begin{aligned}\pi\left(\tau^2 \mid \mathbf{y}, \boldsymbol{\theta}_{-\tau^2}\right) &\propto \pi\left(\tau^2\right)\pi\left(\mathbf{y} \mid \tau^2, \beta_0 \ldots\right) \\ &\propto e^{-b/\tau^2}\left(\tau^2\right)^{-a-1}\cdot\left(\tau^2\right)^{-m/2}e^{-\sum_{i=1}^{m}\frac{(y_i-\mu)^2}{2\tau^2}} \\ &\propto \left(\tau^2\right)^{-\left[a+\frac{n}{2}\right]-1}e^{-\frac{1}{\tau^2}\left(b+\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2}\right)}\end{aligned}$$

with $\mu = \beta_0 + \beta_0(\mathbf{s}_i) + \beta_1 x(B) + \beta_1(\mathbf{s}_i)x(B)$, and $\mathbf{s}_i = (long, lat)$ of point i^{th}.

Hence, it is straightforward that

$$\tau^2 \mid \mathbf{y}, \boldsymbol{\theta}_{-\tau^2} \sim IG\left(a+\frac{n}{2}, b+\sum_{i=1}^{n}\frac{\left(\mathbf{y}_i-\mu\right)^2}{2}\right)$$

5.3 Full conditional of $\omega_0(\mathbf{s})$ and $\omega_1(\mathbf{s})$

The complete final model is

$$Y(\mathbf{s}) = \beta_0 + \beta_1 x(B) + \omega_0(\mathbf{s})[a_{00} + a_{10}x(B)] + \omega_1(\mathbf{s})[a_{11}x(B)] + \epsilon(\mathbf{s})$$

proceeding as above, and defining

$$\mathbf{z} = \mathbf{y} - \beta_0\mathbf{1} + \beta_1 x(B) - \text{diag}(a_{11}x(B))\omega_1(\mathbf{s}), \text{ and}$$

$$c = \text{diag}(a_{00} + a_{10}x(B))$$

we can write the likelihood as

$$L(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\frac{(\mathbf{z}-c\omega_0(\mathbf{s}))^T(\mathbf{z}-c\omega_0(\mathbf{s}))}{\tau^2}\right\}$$

Moreover, setting $\hat{\omega_0}(\mathbf{s}) = \left(c^T c\right)^{-1} c^T \mathbf{z}$ and given the prior

$$\omega_0(\mathbf{s}) \sim N\left(\boldsymbol{b}_{0s}, \Sigma_{0s}\right)$$
$$\pi(\omega_0(\mathbf{s})) \propto \frac{1}{\sqrt{\det \Sigma_{0s}}} \exp\left\{-\frac{1}{2}(\omega_0(\mathbf{s}) - \boldsymbol{b}_{0s})^T \Sigma_{0s}^{-1} (\omega_0(\mathbf{s}) - \boldsymbol{b}_{0s})\right\}$$

the posterior $\pi\left(\omega_0(\mathbf{s}) \mid \mathbf{y}, \boldsymbol{\theta}_{-\omega_0(\mathbf{s})}\right)$ is proportional to

$$\exp\left\{-\frac{1}{2}\left[\omega_0(\mathbf{s})^T \underbrace{\left(\Sigma^{-1} + \Sigma_{0s}^{-1}\right)}_{M} \omega_0(\mathbf{s}) - 2\omega_0(\mathbf{s})^T \underbrace{\left(\Sigma^{-1}\hat{\omega_0}(\mathbf{s}) + \Sigma_{0s}^{-1}\boldsymbol{b}_{0s}\right)}_{\mathbf{d}}\right]\right\}$$

where $\Sigma^{-1} = c^T c/\tau^2$. Hence, rearranging the terms

$$\pi\left(\omega_0(\mathbf{s}) \mid \mathbf{y}, \boldsymbol{\theta}_{-\omega_0(\mathbf{s})}\right) \propto \exp\left\{-\frac{1}{2}\left[\left(\omega_0(\mathbf{s}) - M^{-1}\mathbf{d}\right)^T M \left(\omega_0(\mathbf{s}) - M^{-1}\mathbf{d}\right) - \mathbf{d}^T M^{-1}\mathbf{d}\right]\right\}$$

Finally, as shown previously, we obtained

$$\begin{aligned}\pi\left(\omega_0(\mathbf{s}) \mid \mathbf{y}, \boldsymbol{\theta}_{-\omega_0(\mathbf{s})}, c\right) &\sim N\left(M^{-1}\mathbf{d}, M^{-1}\right) \\ &\sim N\left(\left(\frac{c^T c}{\tau^2} + \Sigma_{0s}^{-1}\right)^{-1}\left(\frac{c^T c}{\tau^2}\hat{\omega_0}(\mathbf{s}) + \Sigma_{0s}^{-1}\boldsymbol{b}_{0s}\right), \left(\frac{c^T c}{\tau^2} + \Sigma_{0s}^{-1}\right)^{-1}\right).\end{aligned}$$

An identical reasoning can be followed for $\omega_1(\mathbf{s})$, obtaining that $\pi\left(\omega_1(\mathbf{s}) \mid \mathbf{y}, \boldsymbol{\theta}_{-\omega_1(\mathbf{s})}, c\right)$ is distributed as

$$N\left(\left(\frac{c^T c}{\tau^2} + \Sigma_{1s}^{-1}\right)^{-1}\left(\frac{c^T c}{\tau^2}\hat{\omega_1}(\mathbf{s}) + \Sigma_{1s}^{-1}\boldsymbol{b}_{1s}\right), \left(\frac{c^T c}{\tau^2} + \Sigma_{1s}^{-1}\right)^{-1}\right)$$

and in this case $c = \text{diag}(a_{11}x(B)), \mathbf{z} = \mathbf{y} - \beta_0\mathbf{1} + \beta_1 x(B) - \text{diag}(a_{00} + a_{10}x(B))\omega_0(\mathbf{s})$ and $\hat{\omega_1}(\mathbf{s}) = (c^T c)^{-1} c^T \mathbf{z}$.

5.4 Full conditional of ϕ_0 and ϕ_1

For $i = 0, 1$ we take a discrete support $\mathcal{S}_i = \left\{a_i, \ldots, a_i + \frac{j_i - 1}{n_i - 1}(b_i - a_i), \ldots, b_i\right\}$, with $j_i = 1, \ldots, n_i$ and let n_i the cardinality of $\mathcal{S}_i$.
The prior of ϕ_i is a discrete uniform distribution on $\mathcal{S}_i$

$$\pi(\phi_i) \sim \mathcal{U}([a_i, b_i], n_i)$$

Knowing that the only parameter that depends on ϕ_i is $\omega_i(\mathbf{s})$, one computes the full conditional as follows.

$$\begin{aligned}\pi(\phi_i \mid \mathbf{y}, \boldsymbol{\theta}_{-\phi_i}, X) &\propto \pi(\phi_i)\pi(\omega_i(\mathbf{s})) \\ &\propto \log(\pi(\phi_i)\pi(\omega_i(\mathbf{s}))) \\ &= \log(\frac{1}{n_i}\frac{1}{\sqrt{\det \Sigma_{ij}}} \exp\left\{-\frac{1}{2}(\omega_i(\mathbf{s}) - \boldsymbol{b}_{is})^T \Sigma_{ij}^{-1}(\omega_i(\mathbf{s}) - \boldsymbol{b}_{0s})\right\})\end{aligned}$$

Where Σ_{ij} is the covariance matrix of i-th Gaussian Process $\omega_i(\mathbf{s})$ computed with the j-th elements of the vector of ϕ_i. After transforming and normalizing, we obtain a Categorical distribution on $\mathcal{S}_i$ with probability vector $\boldsymbol{p^{(i)}} = (p_1, ..., p_{n_i})$ as full conditional , where $p_j^{(i)} = P(\phi_i = a_i + \frac{j_i - 1}{n_i - 1}(b_i - a_i))$ for $j_i = 1, \ldots, n_i$

$$\pi(\phi_i \mid \mathbf{y}, \boldsymbol{\theta}_{-\phi_i}, X) \sim \text{Categorical } (\boldsymbol{p^{(i)}})$$

6 Gibbs Sampler

The computed full conditionals are in closed form.
We decide to implement the MCMC Gibbs Sampler algorithm with 2000 iterations. We want to obtain the updated values for the parameters of our model. Taking into account the last 1000 iterations, we compute the sample mean for certain parameters, which will be useful in the next section:

$$\overline{\beta_0} \quad \overline{\beta_1} \quad \overline{\tau^2} \quad \overline{\omega_0(\mathbf{s})} \quad \overline{\omega_1(\mathbf{s})}$$

7 Prediction with Kriging

Defining s^{new} as new locations where we do not have the data provided by the monitors. Following the approach applied in Carl Edward Rasmussen [2006] we make prediction using kriging: a method of interpolation based on Gaussian process governed by prior covariances. Under suitable assumptions of the prior, kriging gives the best linear unbiased prediction (BLUP) at unsampled locations.
We compute the prediction for our gaussian processes with the key equation for gaussian process regression with noisy observations:

$$\omega_0^{new} = \mathcal{K}_0^*[\mathcal{K}_0 + \overline{\tau^2} I]^{-1} \overline{\omega_0(\mathbf{s})}$$
$$\omega_1^{new} = \mathcal{K}_1^*[\mathcal{K}_1 + \overline{\tau^2} I]^{-1} \overline{\omega_1(\mathbf{s})}$$

where :

- $\mathcal{K}_0$ is the covariance matrix of the first GP
- $\mathcal{K}_1$ is the covariance matrix of the second GP
- $\mathcal{K}_i^*$ is the mix covariance matrix for the i-th GP for $i = 0, 1$
- $\overline{\tau^2}$ and $\overline{\omega_0(\mathbf{s})}$, $\overline{\omega_1(\mathbf{s})}$ are those obtained from the Gibbs Sampler algorithm.

Mixing the results of the Gibbs Sampler algorithm and the estimate of our GPs we are able to make prediction.
We compute :

$$\overline{y} = X\overline{\boldsymbol{\beta}} + a_{00}\omega_0^{new} + a_{10}X\omega_0^{new} + a_{11}X\omega_1^{new}$$

that is our mean estimate of the $PM_{2.5}$ concentration in the new points.

8 Simulated Data

We built a grid of points $[-5, 5] \times [-5, 5]$ upon which we took a set of 100 finite points $\mathbf{s}_i = (x, y)$ in order to represent our monitors. After that we created 16 equally spaced cells.

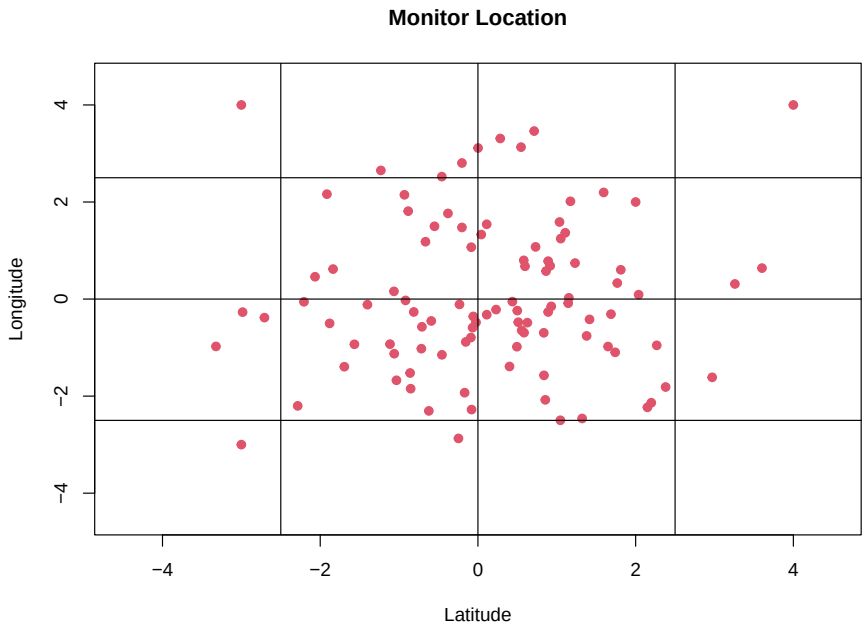


Figure 1: Monitor Location of simulated data.

Chosen priors:

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}\right)$$
$$\tau^2 \sim \mathrm{IG}(4, 0.2)$$
$$\phi_0 \sim \mathrm{UD}([0.0005, 0.05], 9)$$
$$\phi_1 \sim \mathrm{UD}([0.01, 0.1], 9)$$

Instead for the coregionalization matrix we extract the values coming from the following distribution:

$$a_{00}, a_{11} \sim \mathrm{logN}\,(0, 1)$$
$$a_{10} \sim \mathrm{N}(0, 1)$$

After running the Gibbs Sampler on 2000 iterations we obtain the following traceplots and frequencies for our parameters:

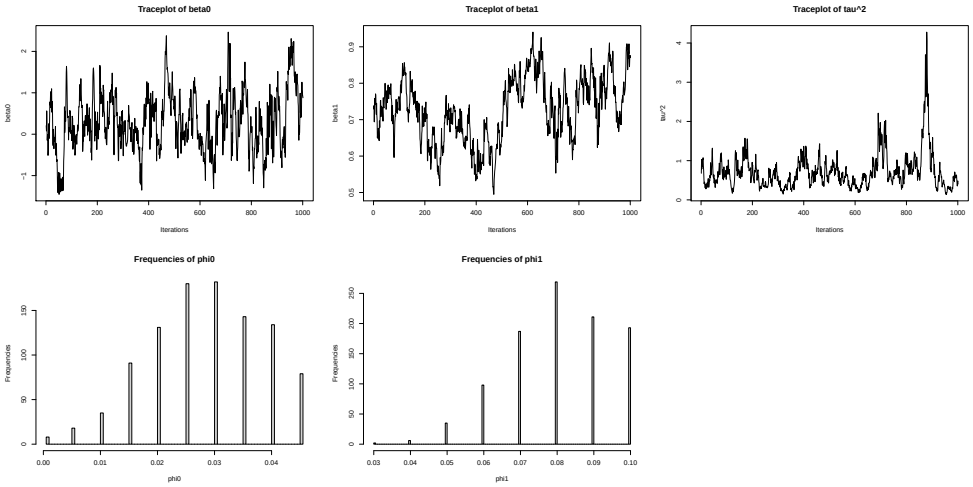


Figure 2: Traceplots for β_0, β_1, τ^2 and frequencies of ϕ_0 and ϕ_1

Averaging the values of the parameters over the last 1000 iterations, we obtain the following estimates:

$$\overline{\beta_0} = 0.3097$$
$$\overline{\beta_1} = 0.7282$$
$$\overline{\tau^2} = 0.7171$$

Finally, we merge the two gaussian processes obtaining the following prediction for our space, in which green zone correspond to healthy locations with clean air while red zone to unhealthy ones.

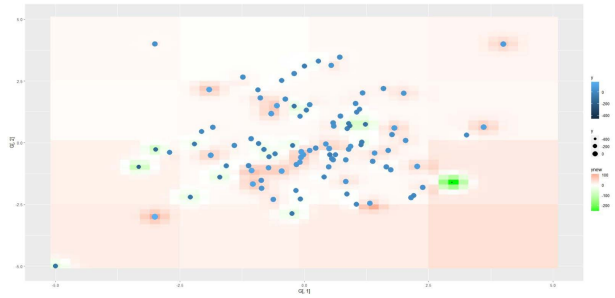


Figure 3: 2d prediction with 100 monitors.
Green (red) area correspond to smaller (higher) values observed by monitors.

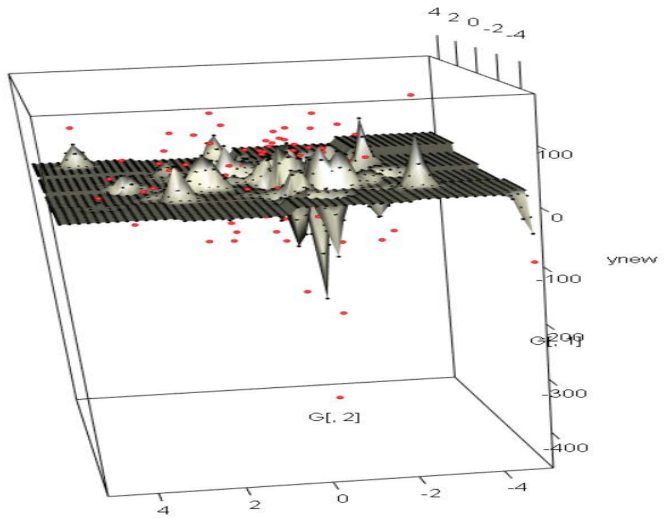


Figure 4: 3d view final result obtained with simulated data.

Looking at the results, we can see that the variance $\overline{\tau^2}$ seems to be consistent w.r.t. the initial value, while different behaviour goes for $\overline{\beta_0}$ and $\overline{\beta_1}$ which seem to increase. On the other hand, the $\overline{\beta}$ values are at least positive which could be expected from the fact that due to air pollution two points in the space necessarily have a positive correlation. Furthermore, the slope $\overline{\beta_1}$ seems to have a major impact w.r.t. the intercept $\overline{\beta_0}$, which was expected due to the high spatial correlation.

9 Real Data

The first dataset (*Obspm2.5concNe2018training*) collects data from $PM_{2.5}$ centralines of the NorthEast Side of the United States of America.

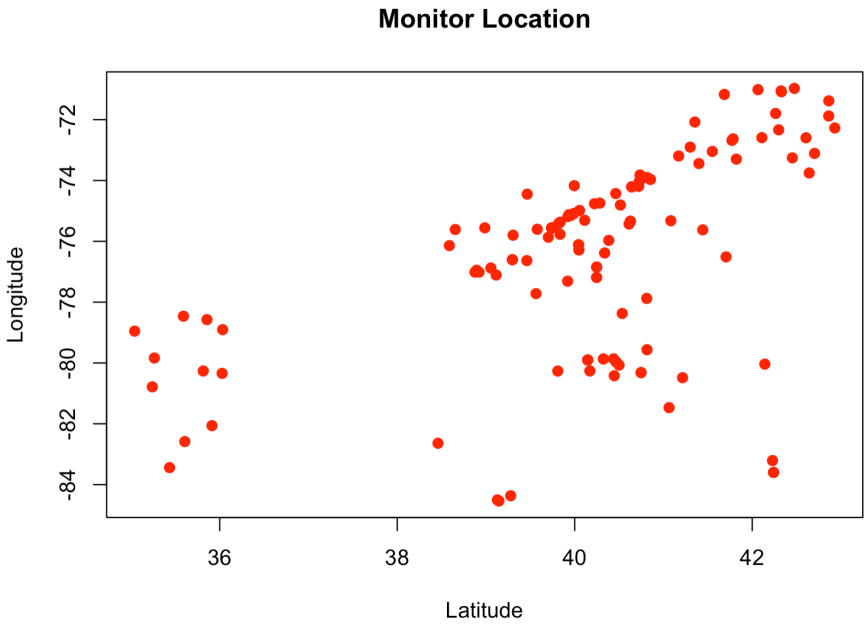


Figure 5: Monitor locations of real data.

The second dataset we consider, instead, contains information about the population density of cities in the region of interest. We obtain it by merging together two different datasets: one with latitude and longitude of the cities in the region of interest (*US City Population Densities*[1]) and the second one with their population densities (*US 2020 Census Cities Populations and Coordinates*[2][3]) .

As a proxy for CMAQ values to use as a covariate of our model, we considered the population density values for each point.The reason behind this choice is that in urban areas the amount of pollutants emitted is very high and it is related to the population density.

We proceed by associating at each point in the space, corresponding to both

[1] https://www.kaggle.com/datasets/mmcgurr/us-city-population-densities
[2] https://www.kaggle.com/datasets/darinhawley/us-2021-census-cities-populations-coordinates
[3] https://worldpopulationreview.com/states/state-abbreviations

monitors and points of the grid, the weighted average of the population density of the closest five city neighbours in terms of euclidean distance. This means that the closer is the city to the point, the higher influence the city has in terms of population density with respect to the others.
Moreover, we consider the logarithm of the population density in order to differentiate more the size of the cities and the number of inhabitants.

We divide the dataset into training and test set. We use the former to fit the model and to predict $PM_{2.5}$ concentration in unknown locations through Kriging, as we did in Section 8.
The following are the traceplots of the parameters and kriging prediction of our model:

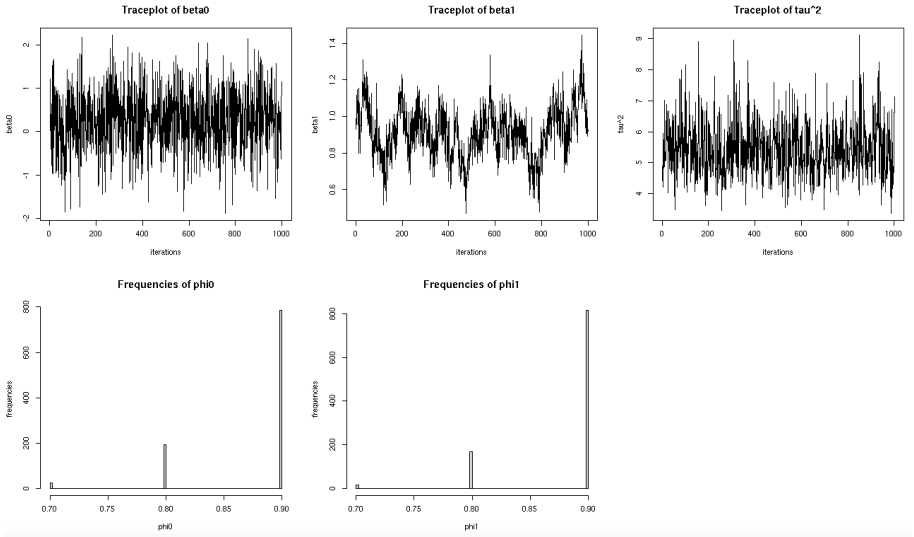
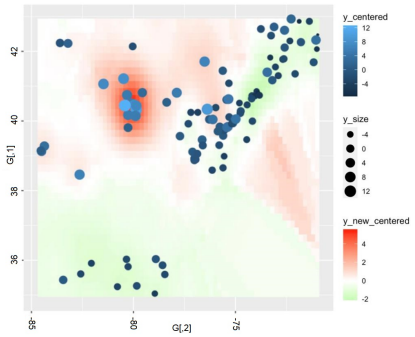
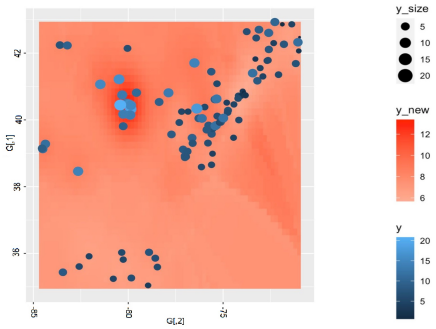


Figure 6: Traceplots of the parameters regarding real data.

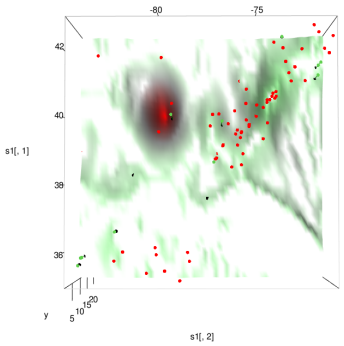
(a) Satellite view of the region of interest (North-Est America).



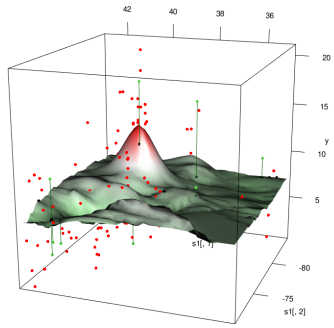
(b) 2d view of $PM_{2.5}$ concentration, centered around the mean value.



(c) 2d view of $PM_{2.5}$ concentration.



(d) 3d view from below.



(e) 3d view of the prediction of $PM_{2.5}$ concentration.

Figure 7: Results obtained applying our model to real data.

For an example, you can see from the plot that the region near Indianapolis is very polluted. This is due to two reasons:

- the manures run off of large animal farms
- the dependence on the automobile and the lack of investment in public transit account for Indianapolis residents to drive more miles per capita than any other large U.S. urban area

We also tried to take into account the geographical constraints as the presence of the ocean, which we supposed to be less pollutated w.r.t. to the city. We assign a priori to those points the mean value of the density population corresponding to the overall region of interest. Using the procedure above, we overcome an important obstacle regarding some part of the ocean where the value of the population density was linked to the big cities of the coast even though those locations were too far.

We use as a term of comparison the MSE (Mean Squared Error) and MAE (Mean Absolute Error).
We computed the value for both the models (with and without flat ocean) and we found identical value. This is due to the fact that our test set was just composed by continental data. So in that sector, our GP process remains the same (MAE=2.90 ;MSE=13.14).

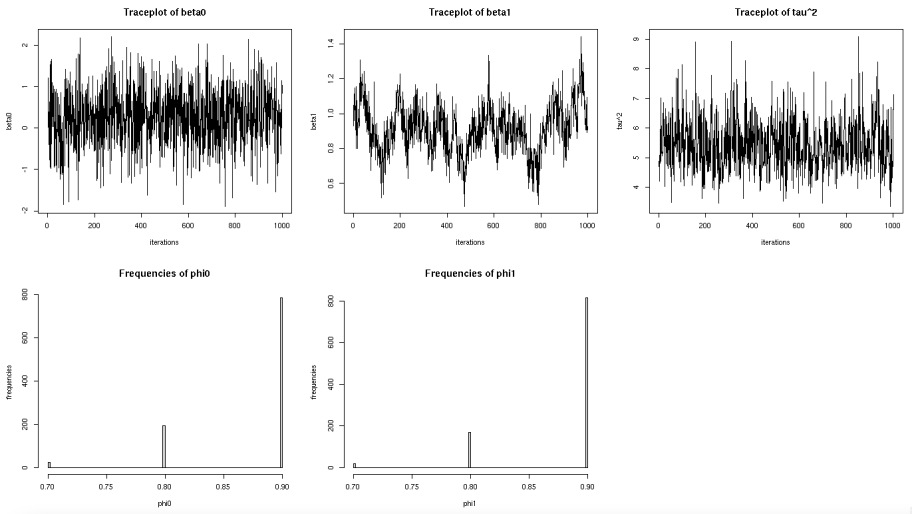
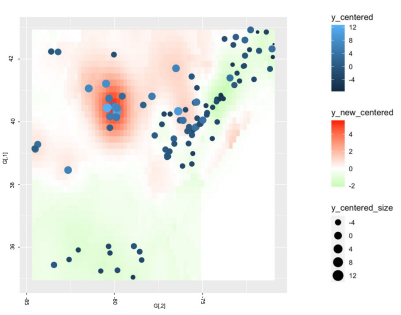
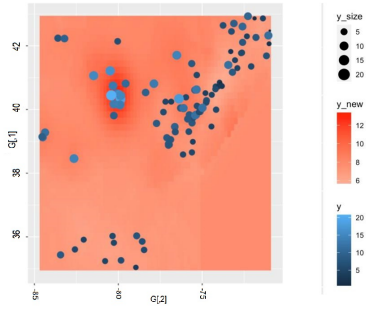


Figure 8: Traceplots of the parameters regarding the real data with flat ocean.

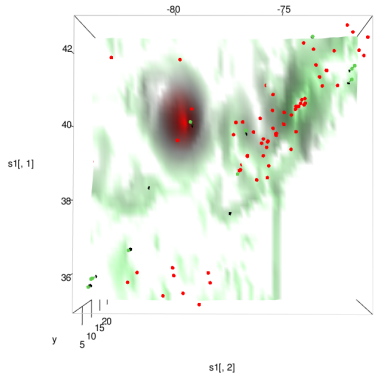
(a) Satellite view of the region of interest(North-Est America).



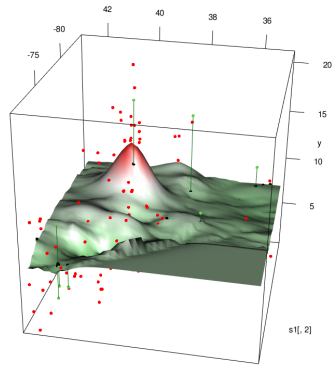
(b) 2d view of $PM_{2.5}$ concentration, centered around the mean value.



(c) 2d view of $PM_{2.5}$ concentration.



(d) 3d view from below.



(e) 3d view of the prediction of $PM_{2.5}$ concentration.

Figure 9: Results obtained applying our model to real data with flat ocean.

10 Conclusion

We succeeded in implementing our model. As a matter of fact, the algorithm seems to work pretty well, predicting values in different space location.

In the field of Geostatistics, bayesian approach is very effective.
Thanks to the usage of two different Gaussian processes linked through the coregionalization matrix, we are able to capture efficiently the value of the pollutant at locations where we do not have a real estimate.

Even so, we have faced some issues. First of all, how to take into consideration the ocean. Indeed, we tried to solve it by taking a fixed value of our covariate (the mean of the population density), but we could considered a more complex model with a Constrained Gaussian Process (Geographical Constrained on the support).
Finally, to overcome the problem of high computational run time, it is possible to implement also other C++ functions.

Further Devolpment

The earlier downscaler model of Berrocal et al. [2010] was later extended: new models developed are intended to address two potential concerns with the model output. One recognizes that there may be useful information in the outputs for grid cells that are neighbors of the one in which the location lies. The second acknowledges potential spatial misalignment between a station and its putatively associated grid cell. In order to do this, several proposals have been developed:

- In Berrocal et al. [2012] was proposed a GMRF smoothed downscaler as a first solution:

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \beta_1 \tilde{V}(B) + \epsilon(\mathbf{s}) \quad \epsilon(\mathbf{s}) \sim N\left(0, \tau^2\right)$$
$$\{V(B) : V(B) = \mu + V(B)\}$$

  where:

$$x(B) = \mu + V(B) + \eta(B) \quad \eta(B) \overset{ind}{\sim} N\left(0, \sigma^2\right)$$

  with μ an overall mean and V(B) a mean-zero Gaussian Markov random field equipped with a conditionally autoregressive structure.

- Still in Berrocal et al. [2012] a second solution was given by a smoothed downscaler using spatially varying random weights:

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \beta_1 \tilde{x}(\mathbf{s}) + \epsilon(\mathbf{s}) \quad \epsilon(\mathbf{s}) \sim N\left(0, \tau^2\right)$$
$$\tilde{x}(\mathbf{s}) = \sum_{k=1}^{g} w_k(\mathbf{s}) x\left(B_k\right)$$

  where: $w_k(\mathbf{s})$ are the weights.

- In Warren et al. [2021], by introducing a spatial distributed lag data fusion model with spatially-lagged predictors, defined by taking the average of grid cell average concentrations in squares surrounding the grid cell containing the pollution monitor. One assumes the larger the distance from the monitor, the lower the regression weight on the lagged average.

$$Y\left(\mathbf{s}_{ij}\right)=\tilde{\beta}_0\left(\mathbf{s}_{ij}\right)+\tilde{\beta}_1\left(\mathbf{s}_{ij}\right)\sum_{l=0}^{L}\bar{x}_{B_i,l}\left(\frac{\pi_{B_i,l}}{\sum_{k=0}^{L}\pi_{B_i,k}}\right)+\epsilon\left(\mathbf{s}_{ij}\right)$$

  The average of the individual CMAQ estimates comprising spatial lag l surrounding CMAQ grid cell B_i is denoted as $\bar{x}_{B_i,l}$. The corresponding model for $\pi_{B_i,l}$ is given as

$$\pi_{B_i,l}=\Phi\left(\mu+\alpha_{B_i}\right)^l, l=0,\ldots,L,$$

  where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, μ represents the global spatial lag structure common to all CMAQ locations and α_{B_i} is the deviation from the global lag structure specific to CMAQ grid cell B_i.

  Moreover, since the PM2.5 data are hourly data (at least for some of the control units), one possible further step, not already developed by these previous articles, could be repeating the same process of the weights, but considering also the time (instead of the space) and take into account also the difference in temporal resolution. Given the real data, since the fine dust are related to the traffic, one could see what are their daily profiles and understand a scheme that could be used for the satellite.

GitHub

This is the link in which you can find all the codes that we have developed:

https://github.com/ale1998bo/Data-fusion-for-estimating-ambient-air-pollution-with-spatial-disalignment.git

Appendix

Appendix A Further details on spatial models

This brief summary is taken form Gelfand et al. [2003]. It is a way of explaining better all the coefficients involved in our model.

Recall the Gaussian stationary spatial process model as in

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \epsilon(\mathbf{s})$$

where $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\beta$ and $\epsilon(\mathbf{s})$ is a white noise process, that is,

$$\mathrm{E}(\epsilon(\mathbf{s})) = 0, \mathrm{Var}(\epsilon(\mathbf{s})) = \tau^2, \mathrm{Cov}\left(\epsilon(\mathbf{s}), \epsilon\left(\mathbf{s}'\right)\right) = 0$$

and $W(\mathbf{s})$ is a second-order stationary mean 0 process independent of the white noise process; that is,

$$\mathrm{E}(W(\mathbf{s})) = 0, \mathrm{Var}(W(\mathbf{s})) = \sigma^2, \mathrm{Cov}\left(W(\mathbf{s}), W\left(\mathbf{s}'\right)\right) = \sigma^2\rho\left(\mathbf{s}, \mathbf{s}'; \phi\right)$$

where ρ is a valid two-dimensional correlation function.

The $W(\mathbf{s})$ are viewed as spatial random effects, and the first formula implicitly defines a hierarchical model. Letting $\mu(\mathbf{s}) = \beta_0 + \beta_1 x(\mathbf{s})$, write $W(\mathbf{s}) = \beta_0(\mathbf{s})$ and define $\tilde{\beta}_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$.
Then $\beta_0(\mathbf{s})$ can be interpreted as a random spatial adjustment at location s to the overall intercept β_0. Equivalently, $\tilde{\beta}_0(\mathbf{s})$ can be viewed as a random intercept process. For an observed set of locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ given $\beta_0, \beta_1, \{\beta_0\left(\mathbf{s}_i\right)\}$ and τ^2, the $Y\left(\mathbf{s}_i\right) = \beta_0 + \beta_1 x\left(\mathbf{s_i}\right) + \beta_0\left(\mathbf{s_i}\right) + \epsilon\left(\mathbf{s_i}\right), i = 1, \ldots, n$, are conditionally independent.

The first-stage likelihood $L\left(\beta_0, \beta_1, \beta_0\left(\mathbf{s}_i\right), \tau^2 \mid \mathbf{y}\right)$ is equal to

$$\left(\tau^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\tau^2}\sum\left(Y\left(\mathbf{s}_i\right) - \left(\beta_0 + \beta_1 x\left(\mathbf{s}_i\right) + \beta_0\left(\mathbf{s}_i\right)\right)^2\right)\right\}$$

The distribution of $\beta_0 = \left(\beta_0\left(\mathbf{s}_1\right), \ldots, \beta_0\left(\mathbf{s}_n\right)\right)^T$ is:

$$f\left(\boldsymbol{\beta}_0 \mid \sigma_0^2, \phi_0\right) = N\left(\mathbf{0}, \sigma_0^2 H_0\left(\phi_0\right)\right),$$

where $\left(H_0\left(\phi_0\right)\right)_{ij} = \rho_0\left(\mathbf{s}_i - \mathbf{s}_j; \phi_0\right)$ and $\phi = (\gamma, \nu)$, where γ is a decay parameter and ν is a smoothness parameter. With a prior on $\beta_0, \beta_1, \tau^2, \sigma_0^2$, and ϕ_0, specification of the Bayesian hierarchical model is completed. It is possible to integrate over $\boldsymbol{\beta}_0$, obtaining the marginal likelihood:

$$L\left(\beta_0, \beta_1, \tau^2, \sigma_0^2, \phi_0; \mathbf{y}\right) = \left|\sigma_0^2 H_0\left(\phi_0\right) + \tau^2 I\right|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \beta_0\mathbf{1} - \beta_1\mathbf{x}\right)^T\left(\sigma_0^2 H_0\left(\phi_0\right) + \tau^2 I\right)^{-1}\left(\mathbf{y} - \beta_0\mathbf{1} - \beta_1\mathbf{x}\right)\right\}$$

where $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n))^T$.

The samples from $\beta_0 \mid \mathbf{y}$ can be obtained one-for-one because

$$f(\boldsymbol{\beta}_0 \mid \mathbf{y}) = \int f\left(\boldsymbol{\beta}_0 \mid \beta_0, \beta_1, \tau^2, \sigma_0^2, \phi_0, \mathbf{y}\right) f\left(\beta_0, \beta_1, \tau^2, \sigma_0^2, \phi_0 \mid \mathbf{y}\right)$$

defining $f\left(\beta_0 \mid \beta_0, \beta_1, \tau^2, \sigma_0^2, \phi_0, \mathbf{y}\right)$ as

$$N\left(\left(\frac{1}{\tau^2} I + \frac{1}{\sigma_0^2} H_0^{-1}(\phi_0)\right)^{-1} \frac{1}{\tau^2}(\mathbf{y} - \beta_0 \mathbf{1} - \beta_1 \mathbf{x}),\ \left(\frac{1}{\tau^2} I + \frac{1}{\sigma_0^2} H_0^{-1}(\phi_0)\right)^{-1}\right)$$

It's possible also to obtain samples from the posterior of the $\beta_0(\mathbf{s})$ process at a new location, say $\mathbf{s}_{\text{new}}$, to provide interpolation for the $\beta_0(\mathbf{s})$ surface. Specifically,

$$f(\beta_0(\mathbf{s}_{\text{new}}) \mid \mathbf{y}) = \int f\left(\beta_0(\mathbf{s}_{new}) \mid \boldsymbol{\beta}_0, \sigma_0^2, \phi_0\right) f\left(\boldsymbol{\beta}_0, \sigma_0^2, \phi_0 \mid \mathbf{y}\right).$$

The first density under the integral is a univariate normal that can be written down directly from the specification of $\beta_0(\mathbf{s})$. For the prediction of $y(\mathbf{s}_{\text{new}})$ given $\mathbf{y}$, it requires

$$\begin{aligned} f(y(\mathbf{s}_{\text{new}}) \mid \mathbf{y}) = \int & f\left(y(\mathbf{s}_{\text{new}}) \mid \beta_0, \beta_1, \beta_0(\mathbf{s}_{\text{new}}), \tau^2\right) \\ & \cdot f\left(\beta_0(\mathbf{s}_{\text{new}}) \mid \beta_0, \sigma_0^2, \phi_0\right) \\ & \cdot f\left(\boldsymbol{\beta}_0, \beta_0, \beta_1, \tau^2, \sigma_0^2, \phi_0 \mid \mathbf{y}\right) \end{aligned}$$

The first term under the integral sign is a normal density. Again, it is straightforward to obtain samples from this predictive distribution.

The following models will also be analyzed with the same approach. The last one is "our" model which is in the article Berrocal et al. [2012].

$$Y(\mathbf{s}) = \beta_0 + \beta_1 x(\mathbf{s}) + \beta_1(\mathbf{s}) x(\mathbf{s}) + \epsilon(\mathbf{s}).$$
$$Y(\mathbf{s}) = \beta_0 + \beta_0 x(\mathbf{s}) + \beta_1 x(\mathbf{s}) + \beta_1(\mathbf{s}) x(\mathbf{s}) + \epsilon(\mathbf{s})$$

.

References

Veronica J. Berrocal, Alan E. Gelfand, and David M. Holland. A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(2):176–197, 2010. ISSN 10857117. doi: 10.1007/s13253-009-0004-z.

Veronica J. Berrocal, Alan E. Gelfand, and David M. Holland. Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality. *Biometrics*, 68(3):837–848, 2012. ISSN 0006341X. doi: 10.1111/j.1541-0420.2011.01725.x.

Christopher K.I. Williams Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. the MIT Press, 2006.

Sandrah P Eckel, Myles Cockburn, Yu-Hsiang Shu, Huiyu Deng, Frederick W Lurmann, Lihua Liu, and Frank D Gilliland. Air pollution affects lung cancer survival. *Thorax*, 71(10):891–898, 2016.

Alan E. Gelfand, Hyon Jung Kim, C. F. Sirmans, and Sudipto Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003. ISSN 01621459. doi: 10.1198/016214503000170.

Joshua L. Warren, Marie Lynn Miranda, Joshua L. Tootoo, Claire E. Osgood, and Michelle L. Bell. Spatial distributed lag data fusion for estimating ambient air pollution. *Annals of Applied Statistics*, 15(1):323–342, 2021. ISSN 19417330. doi: 10.1214/20-AOAS1399.