Data fusion for estimating ambient air pollution with spatial disalignment

Bayesian Statistics

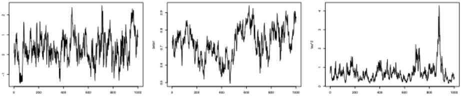
Giulia Barbato
Michele Cosi
Martina Del Basso
Chiara Esposito
Alessandro Frabetti
Marco Rizzo

Mentors: F. Denti[1] and V.J. Berrocal[2]

14th February 2023

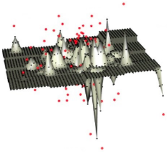
[1] Università Cattolica del Sacro Cuore [2] University of California, Irvine

Table of contents



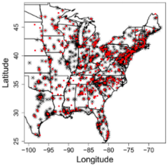
Results & Conclusions

Gibbs Sampler & Kriging prediction



Model and Goal

Full Conditional Distributions



Data Resume

Data

Theoretical data

Two types of spatial data:

(1) Air Quality System (**AQS**) Monitor Measurements

(2) Outputs of Community Multiscale Air Quality (**CMAQ**) Modeling System

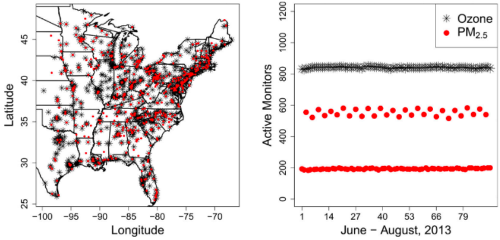


Figure 1: Air quality system active monitor locations (left panel) and time series plots of the number of daily active monitors from June 1–August 31, 2013 (right panel) for ozone and $PM_{2.5}$. [2]

Simulated data

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

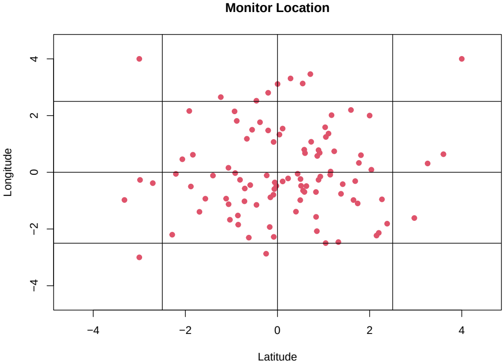


Figure 2: Monitor Locations of simulated data.

Goal

How to combine spatially this 2 type of informations (areal and point wise data) in order to predict the pollutant concentration at locations that lack monitoring measurements?

The Model

Model:
For our project, we will consider the model proposed in [1]:

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

where

- $Y(\mathbf{s})$ is the square root of the observed ozone concentration at a point s
- $x(B)$ is the square root of the numerical model output of the CMAQ[1], daily average ozone value in its 8 hr maximum concentration, over each grid cell B
- $\epsilon(\mathbf{s})$ is a white noise process with nugget variance τ^2

[1] Community Multiscale Air Quality

Spatial alignment via Gaussian Processes

We define:

$$\tilde{\beta}_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$$
$$\tilde{\beta}_1(\mathbf{s}) = \beta_1 + \beta_1(\mathbf{s}) \tag{2}$$

where

- β_0, β_1 are the overall additive and multiplicative components of the CMAQ model
- $\beta_0(\mathbf{s}), \beta_1(\mathbf{s})$ are spatially-varying coefficients that govern the local adjustments to the additive and multiplicative components

$\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are spatial processes: we model them as bivariate mean-zero **Gaussian spatial processes** using coregionalization:

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

where $\mathbf{A}$ is the coregionalization matrix, and $w_0(\mathbf{s})$ and $w_1(\mathbf{s})$ are two mean-zero unit-variance independent Gaussian processes with covariance function

$$\operatorname{cov}\left(w_j(\mathbf{s}), w_j\left(\mathbf{s}'\right)\right) = \exp\left(-\frac{|\mathbf{s}-\mathbf{s}'|^2}{\phi_j}\right). \tag{4}$$

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

Road map and goal

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

1. Derive and implement the full conditionals for the model described before
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

Full conditional distributions

Full conditionals of β, τ^2

As shown last time, we derived the full conditional distributions of $\boldsymbol{\beta}$, τ^2, $\omega_0(\mathbf{s})$ and $\omega_1(\mathbf{s})$.

Full conditional of β[a] :

$$(\beta \mid y, \theta_{-\beta}, X) \sim N_2\left(\left(\frac{X^TX}{\tau^2} + \Sigma_\beta^{-1}\right)^{-1} \cdot \left(\frac{X^TX}{\tau^2}\hat{\beta} + \Sigma_\beta^{-1}\mathbf{b}\right), \left(\frac{X^TX}{\tau^2} + \Sigma_\beta^{-1}\right)^{-1}\right)$$

Full conditional of $\boldsymbol{\tau}^2$[b] :

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

[a] $\beta \sim N_2\left(\mathbf{b}, \Sigma_\beta\right)$, $\quad X = (1_n \quad \mathbf{x}(\mathbf{B}))$, $\quad \hat{\beta} = \left(X^TX\right)^{-1}X^T\mathbf{z}$, $\quad \mathbf{z} = \mathbf{y} - \beta_0(\mathbf{s}) - \beta_1(\mathbf{s})x(B)$

[b] $\tau^2 \sim IG(a, b)$, $\quad \mu = \beta_0 + \beta_0(\mathbf{s}_i) + \beta_1 x(B) + \beta_1(\mathbf{s}_i)x(B)$, $\quad \mathbf{s}_i = (long, lat)$ point i^{th}

Full conditionals of $\omega_0(\mathbf{s})$, and $\omega_1(\mathbf{s})$

Full conditional of $\omega_0(\mathbf{s})^a$:

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

Full conditional of $\omega_1(\mathbf{s})^b$:

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

---

[a] $\omega_0(\mathbf{s}) \sim N_s(\mathbf{b}_{0s}, \Sigma_{0s})$, $\quad c = \mathbf{diag}(a_{00} + a_{10}x(B))$,
$\mathbf{z} = \mathbf{y}$ - $\beta_0\mathbf{1} + \beta_1 x(B) - \mathbf{diag}(a_{11}x(B))\omega_1(\mathbf{s})$, $\quad \hat{\omega_0}(\mathbf{s}) = (c^T c)^{-1}c^T\mathbf{z}$

[b] $\omega_1(\mathbf{s}) \sim N_s(\mathbf{b}_{1s}, \Sigma_{1s})$, $\quad c = \mathbf{diag}(a_{11}x(B))$,
$\mathbf{z} = \mathbf{y}$ - $\beta_0\mathbf{1} + \beta_1 x(B) - \mathbf{diag}(a_{00} + a_{10}x(B))\omega_0(\mathbf{s})$, $\quad \hat{\omega_1}(\mathbf{s}) = (c^T c)^{-1}c^T\mathbf{z}$

Full conditionals of ϕ_0 and ϕ_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 $i = 0, 1$ and $j_i = 1, \ldots, n_i$ where n_i is the cardinality of $\mathcal{S}_i$.

The prior of ϕ_i is a discrete uniform distribution on $\mathcal{S}_i$

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

Considering:

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

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.
We obtain a Categorical distribution on $\mathcal{S}_i$ with probability vector $\boldsymbol{p}^{(i)} = (p_1, \ldots, p_{n_i})$ as full conditional, where $p_j^{(i)} = P(\phi_i = a_i + \frac{j_i - 1}{n_i - 1}(b_i - a_i))$ for $j_i = 1, \ldots, n_i$

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

Gibbs Sampler

Gibbs Sampler

We implement the MCMC Gibbs Sampler and run the algorithm for 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 the parameters of interest:

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

Algorithm 1 Gibbs Sampler

Require: Full conditionals of $\beta_0, \beta_1, \tau^2, \omega_0(\mathbf{s}), \omega_1(\mathbf{s}), \phi_0, \phi_1$
  for $k = 1, \ldots, 2000$ **do**
    Sample in order β, τ^2,$\omega_0(\mathbf{s})$, $\omega_1(\mathbf{s})$ from the Full conditionals
    for $j = 1, \ldots, n_i$, with $i = 0, 1$ **do**
      Compute $p_j^{(i)}$
    end for
    Sample $\phi_i \sim$ Categorical($\mathbf{p}^{(\mathbf{i})}$)
  end for
return $\beta_0, \beta_1, \tau^2, \omega_0(\mathbf{s}), \omega_1(\mathbf{s}), \phi_0, \phi_1$

Prediction with Kriging

Prediction with Kriging

We apply **Kriging** to make spatial prediction on the other points in order to incorporate the knowledge that the training data provides.

Defining s^{new} as new locations where we do not have the data provided by the monitors:

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

where

- $\mathcal{K}$ is the covariance matrix calculated through the Kernel function, using an Euclidean distance.
- $\overline{\tau^2}$ and $\overline{\omega_0(\mathbf{s})}$, $\overline{\omega_1(\mathbf{s})}$ are those obtained from the Gibbs Sampler algorithm.

The mean estimate of the $PM_{2.5}$ concentration in the new points:

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

Data Simulation

Priors

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 sample the values from the following distributions and keep them fixed:

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

Traceplots and frequencies

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

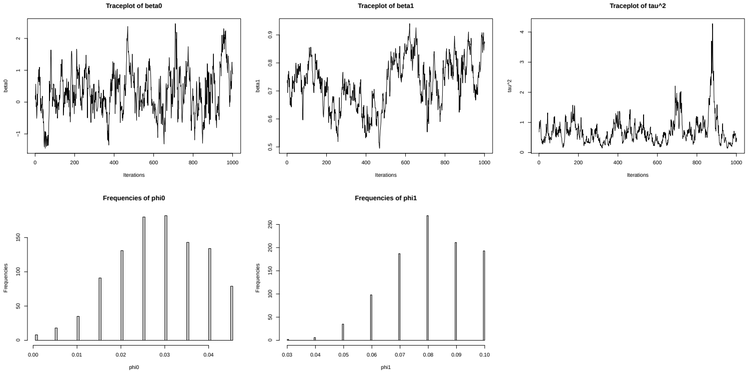


Figure 3: Traceplots for β_0, β_1, τ^2 and frequencies of ϕ_0 and ϕ_1

Results

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 predictions for our space

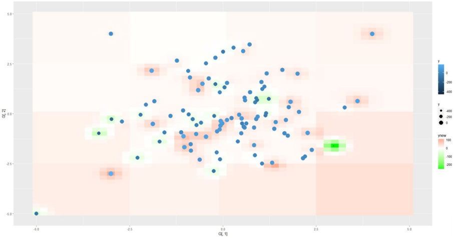


Figure 4: 2d prediction with 100 monitors.
Green (red) areas correspond to smaller (higher) values observed by monitors.

3d-view

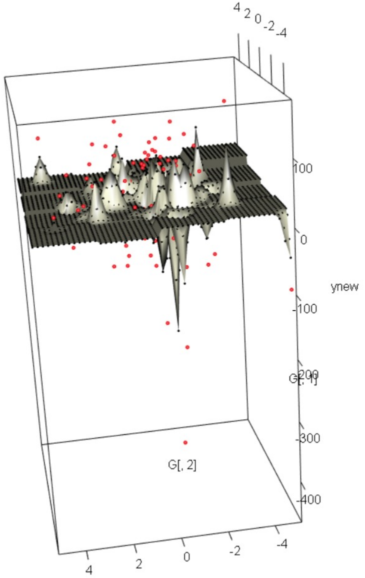


Figure 5: 3d view final result obtained with simulated data.

Application with Real Data

Real Data

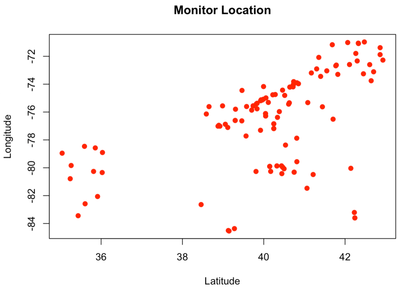


Figure 6: Monitor locations of Real Data.

Figure 7: Satellite view of the region of interest (North-Est America)

Results obtained applying our model to real data

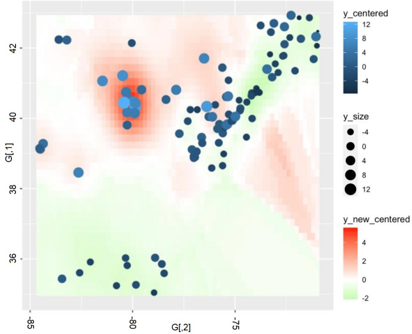


Figure 8: 2d view of the prediction of $PM_{2.5}$ concentration, centered around the mean value.

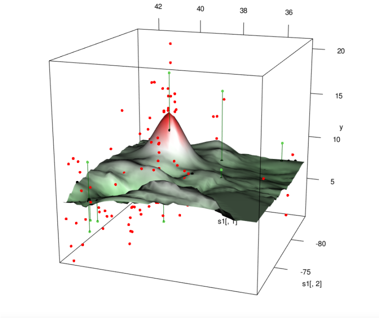


Figure 9: 3d view of the prediction of $PM_{2.5}$ concentration.

Results obtained applying our model to real data with flat ocean

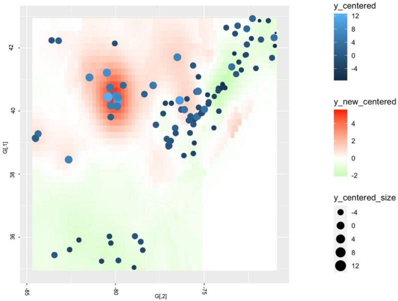


Figure 10: 2d view of the prediction of $PM_{2.5}$ concentration, centered around the mean value.

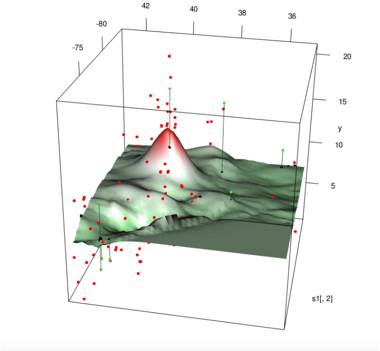


Figure 11: 3d view of the prediction of $PM_{2.5}$ concentration.

References

[1] Veronica J. Berrocal, Alan E. Gelfand, and David M. Holland. "A spatio-temporal downscaler for output from numerical models". In: *Journal of Agricultural, Biological, and Environmental Statistics* 15.2 (2010), pp. 176–197. ISSN: 10857117. DOI: 10.1007/s13253-009-0004-z.

[2] Joshua L. Warren et al. "Spatial distributed lag data fusion for estimating ambient air pollution". In: *Annals of Applied Statistics* 15.1 (2021), pp. 323–342. ISSN: 19417330. DOI: 10.1214/20-AOAS1399.

Thank you for your consideration!