Model-based Geostatistics under Spatially Varying Preferential Sampling

André V. Ribeiro Amaral †

andre.ribeiroamaral@kaust.edu.sa

King Abdullah University of Science and Technology

Geospatial Statistics and Health Surveillance Research Group

† Joint work with Elias T. Krainski, Ruiman Zhong, and Paula Moraga.

Introduction

In this work, we propose a **new model** for **geostatistical data** that accounts for **preferential sampling** by including a **spatially varying coefficient** that describes the dependence strength between the process that models the sampling locations and the latent field.

Here, **geostatistics** will refer to the analysis of data sampled from a process $\zeta(x)$ in a spatially continuous domain, say $\mathscr{D}$, at a discrete set of locations $x = (x_1, \cdots, x_n)^\top$, such that $x_i \in \mathscr{D}$, $\forall i$.

Let ξ be the point process that models the locations where the process ζ is observed, then if

$$\pi(\zeta, \xi) \neq \pi(\zeta) \cdot \pi(\xi),$$

where $\pi(x)$ means "distribution of x," we say that we are under a **"preferential sampling"** setting. Otherwise, we are are under a "non-preferential sampling" setting.

Preferential Sampling Model (Diggle et al., 2010)

Suppose that y_i denotes the observed value of a noisy version of the spatial process $\zeta(x_i)$ at a given location x_i, for any i. Then, the following approach is a common choice

$$y_i = \mu + \zeta(x_i) + \epsilon_i, \text{ s.t. } \epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$$

Here, we can assume that $\zeta(x_i)$ has mean zero. In that case, $\mathbb{E}(y_i) = \mu$, $\forall i \in \{1, \cdots, n\}$.

Aiming to allow for the stochastic dependence between ξ and ζ, we will assume the following

1. ζ is a stationary and isotropic Gaussian random process with mean zero, variance σ_ζ^2, and covariance function $r_\zeta(h;\theta)$, where $h = ||x_1 - x_2||$ is the Euclidean distance between x_1 and x_2.
2. $\xi|\zeta(x)$ is a PP with intensity $\lambda(x) = \exp\{\alpha + \gamma \cdot \zeta(x)\}$, such that $x \in \mathscr{D}$, and $\alpha, \gamma \in \mathbb{R}$.
3. Conditional on x and $\zeta(x)$, $y = (y_1, \cdots, y_n)^\top$ is an i.i.d. vector, such that $y_i \sim \text{Normal}(\mu + \zeta(x_i), \sigma_\epsilon^2)$, $\forall i$.

Extended Preferential Sampling Model

To allow for a **spatially varying degree of preferentiality,** instead of assumption "2." as before, we will say that $\xi|\zeta(x)$ is a PP with intensity

$$\lambda(x) = \exp\{\alpha + \gamma(x) \cdot \zeta(x)\}, \tag{1}$$

where $x \in \mathscr{D}$, and $\alpha, \gamma(x) \in \mathbb{R}$. Here, $\gamma(x)$ is a process defined on $\mathscr{D}$ that dictates how the degree of preferentiality must vary over the spatial domain. For now, we will not impose any constraints over $\gamma(x)$.

However, from Equation (1), notice that the multiplicative structure for the preferentiality and latent fields may yield identifiability issues. **This might be a problem!**

Extended Preferential Sampling Model

To alleviate this issue, we will specify $\gamma(x)$ using a (typically small) set of basis functions in the following way

$$\hat{\gamma}(x) = \sum_{k=1}^{K} \beta_k \phi_k(x),$$

where $\beta_k \in \mathbb{R}$, for all $k \in \{1, \cdots, K\}$, are uncorrelated Gaussian distributed coefficients, and $\{\phi_k(x)\}_{k=1}^{K}$ is a set of basis function (examples come next) defined over the same domain $\mathscr{D}$.

Extended Preferential Sampling Model

Thus, the complete model is specified as follows

$$
\begin{aligned}
y_i &= \mu + \zeta(x_i) + \epsilon_i, \text{ s.t. } \epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma_\epsilon^2), \ \forall i \\
\zeta(x) &\sim \text{Gaussian Process}(0, r_\zeta(h;\theta)) \\
\xi|\zeta(x) &\sim \text{Poisson Point Process}(\lambda(x)) \\
\lambda(x) &= \exp\{\alpha + \gamma(x) \cdot \zeta(x)\} \\
\gamma(x) &= \sum_{k=1}^{K} \beta_k \phi_k(x), \text{ s.t. } \beta_k \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma_\beta^2), \ \forall k,
\end{aligned}
\tag{2}
$$

where the covariance function $r_\zeta(h;\theta)$ will be defined based on the Matérn model.

Basis Functions

One might consider different types of **basis functions**—e.g., constant, piecewise constant, (horizontal or vertical) unidirectional triangular, or radial basis (built using a compactly supported Wendland function defined in two dimensions).

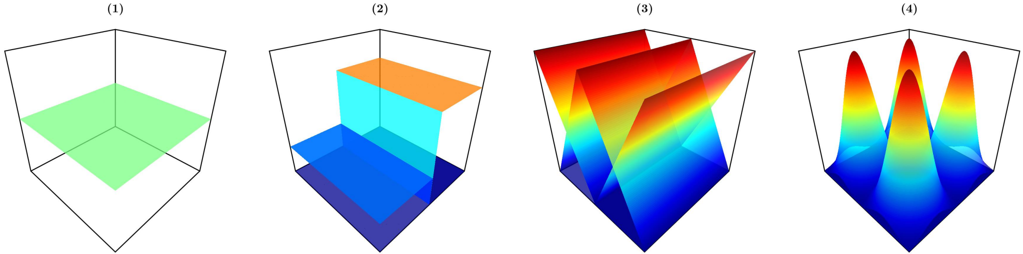

Figure 1: The basis functions were set to (1) constant, (2) piecewise constant, (3) horizontal unidirectional triangular, and (4) radial basis (Wendland).

Example

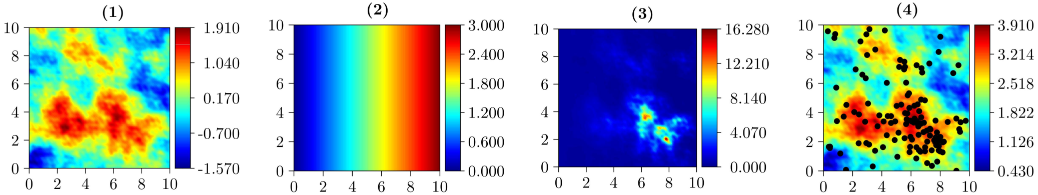

Figure 2: Simulated data in $\mathscr{D} = [0, 10] \times [0, 10]$ for Scenario 04 and $\mathbb{E}(N(\mathscr{D})) = 100$. (1) is a realization of the latent field $\zeta(x)$, (2) is the preferentiality surface $\gamma(x)$ with scale parameter $s = 3$, (3) is the intensity process $\lambda(x)$, and (4) is $\mu + \zeta(x)$ with the observations plotted as points.

Application

We will model air pollution based on the $PM_{2.5}$ levels. In particular, the **$PM_{2.5}$ levels** were collected in the **USA** in 2022 and averaged across the year. In total, we have **942 stations**.

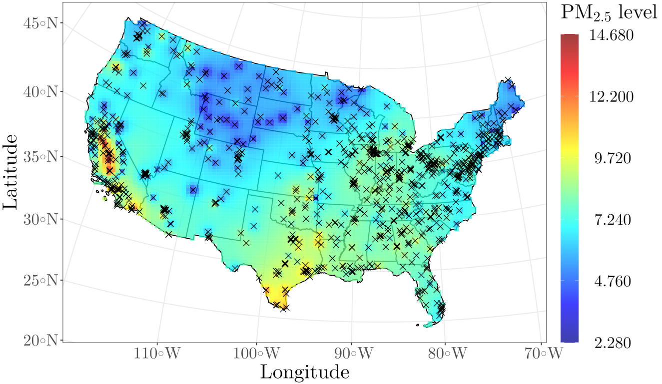

Figure 3: Sampling locations (942 stations) and interpolated values (via IDW) for the $PM_{2.5}$ levels.

Application

We will fit the "radial basis (Wendland)"-based model, such that $K = 15$. The following map show the estimated **PM$_{2.5}$ levels** (based on the posterior mean).

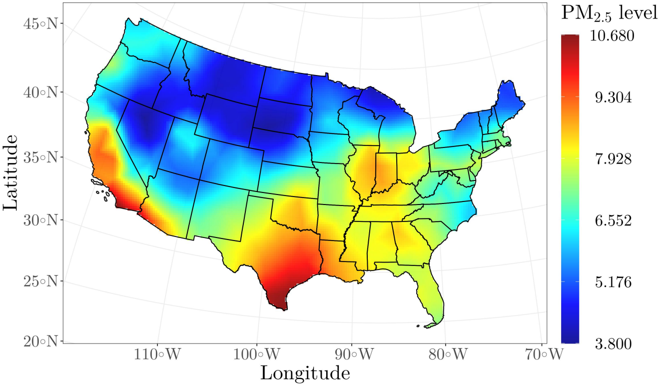

Figure 4: Estimated PM$_{2.5}$ levels (in μg/m^3) in 2022 in the USA territory (excluding "Alaska").

Application

Under the same settings for the fitted model, we can investigate the estimated (based on the posterior mean) **preferentiality surface** $\gamma(x)$.

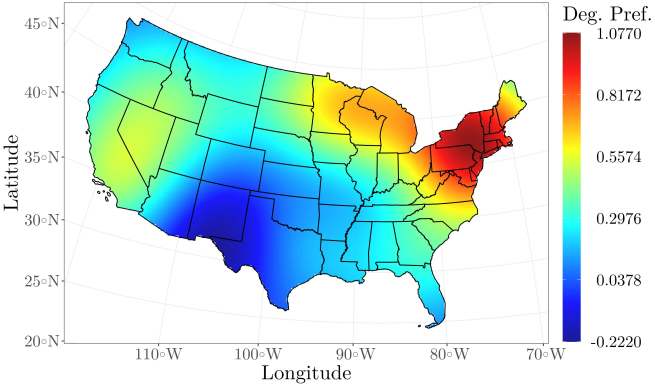

Figure 5: Estimated degree of preferentiality $\hat{\gamma}(x)$, $\forall x \in \mathscr{D} =$ USA, based on $PM_{2.5}$ data.

Application

As a remark, we can also investigate the estimated (based on the posterior mean) the estimated **intensity process** $\lambda(x)$.

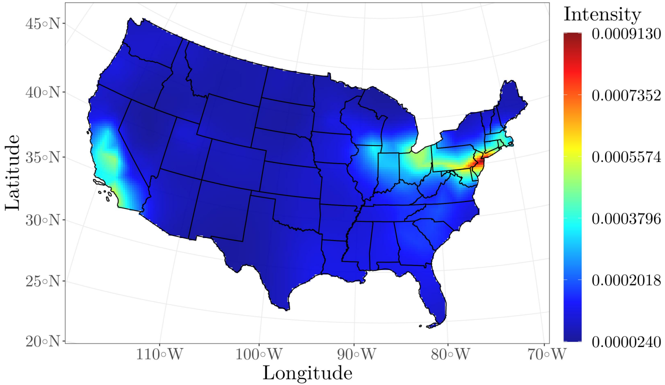

Figure 6: Estimated intensity process $\hat{\lambda}(x)$, $\forall x \in \mathscr{D} = \text{USA}$, based on $PM_{2.5}$ data.

Discussion

We proposed a geostatistical model that accounts for spatially varying **preferential sampling** by allowing the **degree of preferentiality** $\gamma(x)$ **to vary over space**.

To do so, we approximated $\gamma(x)$ by a set of **basis functions** and **unknown coefficients**.

Although I skipped the details, we implemented the model-fitting routines with the **INLA** and **SPDE** approaches which **reduces the computational burden** for parameter estimation and allows fast inference.

We **concluded** that, given enough events, **our model**, along with the implemented inference routine, might **retrieve well** the latent field itself and the spatially varying preferentiality surface, (sometimes) **even under misspecified scenarios**.

As a *final remark*, in the corresponding paper, we offer **guidelines** for the **specification** and **size** of the set of basis functions.