Assessing the Effect of Model-based Geostatistics Under Preferential Sampling for Spatial Data Analysis

André Victor Ribeiro Amaral [1,⋆], Paula Moraga [1]

[1] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST).
⋆ E-mail: andre.ribeiroamaral@kaust.edu.sa

1. Introduction

For problems from the geostatistics domain, it is usually assumed that the sampling process is independent of the process of interest. However, this may not always be the case, and in situations where this assumption does not hold, we say we have *preferential sampling*, as in [1].

Before, models that account for preferential sampling were fitted by rewriting the likelihood function in a way that it could be seen as an expectation, allowing researchers to approximate it by Monte Carlo methods. However, more recently, a Bayesian approach relying on the Integrated Nested Laplace Approximation (INLA) and Stochastic Partial Differential Equation (SPDE) methods started to be employed.

2. Geostatistical Model

A geostatistical model to predict a spatially continuous process can be defined as follows. Suppose that Y_i denotes the observed value of a noisy version of a spatial process $S(x_i)$ at some given location $x_i \in \mathcal{A} \subseteq \mathbb{R}^2$, for $i \in \{1, \cdots, n\}$, in the following manner

$$Y_i = \mu + S(x_i) + \epsilon_i, \qquad (1)$$

where ϵ_i are independent Gaussian zero-mean random variables with $\text{Var}(\epsilon_i) = \sigma^2_\epsilon$. Also, $S(x_i)$ will be assumed to have zero mean—meaning that $\mathbb{E}(Y_i) = \mu$, $\forall i$.

Additionally, let $x = (x_1, \cdots, x_n)$ be a realization of a random vector $X = (X_1, \cdots, X_n)$ and $S(x) = (S(x_1), \cdots, S(x_n))$ a realization of a random process $S(X) = (S(X_1), \cdots, S(X_n))$ evaluated at X. In this case, although Model (1) is fairly common, it usually assumes that X is stochastically independent from $S(X)$, which is not reasonable in many situations.

To account for this dependence, as in [1], the following additional assumptions for Model (1) are required

1. S is a stationary and isotropic Gaussian random field with mean zero, variance σ^2, and correlation function $r(h; \theta) = \text{corr}(S(x_1), S(x_2))$, for $h \neq 0$, where h is the Euclidean distance between x_1 and x_2.
2. $X | S \sim$ Poisson Process$(\lambda(x))$ with intensity $\lambda(x) = \exp\{\alpha + \beta S(x)\}$, for $\alpha, \beta \in \mathbb{R}$.
3. Conditional on S and X, the $Y = (Y_1, \cdots, Y_n)$ is a vector of independent Gaussian random variables, such that $Y_i \sim \text{Normal}(\mu + S(x_i), \sigma^2_\epsilon)$, $\forall i$.

3. Inference

Since the original paper that introduced the preferential sampling idea was published by [1], people have been working on this class of problems using different approaches. Here, we will present two of them, namely the original idea, and the a method that uses INLA and the SPDE techniques.

Original Approach

Start by recalling that, if we consider Model (1) and if we want to predict the value of the process in, say, x_0, we can use, for instance, the Best Unbiased Linear Predictor (BLUP). And to do this, we have to be able to estimate the parameters of the model. In particular, if $S(x)$ is a Gaussian random field with a covariance structure described by $\Sigma(\theta)$, this can be done through the Maximum Likelihood method.

In that case, if X and $S(X)$ are not independent, then the likelihood function $\mathcal{L}(\theta)$, given the data, is

$$\begin{aligned} \mathcal{L}(\theta) = [X, Y] &= \textstyle\int [X, Y, S] dS \\ &= \textstyle\int [Y | S, X][X | S][S] dS. \qquad (2) \end{aligned}$$

Therefore, to determine θ that maximizes $\mathcal{L}(\theta)$, one has to solve the integral in Equation (2). And for this problem, [1] has proposed a way to approximate $\int [Y|S, X][X|S][S]dS$ using a Monte Carlo method. Then, from the approximated likelihood function, they could do inference by determining θ that maximizes $\mathcal{L}_{\text{Approx.}}(\theta)$.

INLA and SPDE Approach

An alternative approach to estimate the model parameters and make prediction for Model (1) is to use the INLA and SPDE approaches, which can be easily implemented with the `R-INLA` package [4]. In a nutshell, INLA is a method for approximating Bayesian inference in latent Gaussian models [3]. In particular, models are of the form

$$\begin{aligned} y_i | S(x_i), \theta &\sim \pi(y_i | S(x_i), \theta), \text{ for } i \in \{1, \cdots, n\} \\ S(x) | \theta &\sim \text{Normal}(\mu(\theta), Q(\theta)^{-1}) \\ \theta &\sim \pi(\theta), \end{aligned}$$

where $y = (y_1, \ldots, y_n)$ is the vector or observed values, $x = (x_1, \ldots, x_n)$ is a Gaussian random field, and $\theta = (\theta_1, \ldots, \theta_k)$, for some $k \in \mathbb{N}$, is a vector of hyperparameters. $\mu(\theta)$ and $Q(\theta)$ represent the mean vector and the precision matrix, respectively.

From the above formulation, notice that Model (1) can be classified as a latent Gaussian model, and therefore we can use the INLA method. To fit Model (1) model using `R-INLA`, we will take an SPDE approach. As showed in [5], a Gaussian random field with Matérn covariance matrix can be expressed as a solution of

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau S(x)) = \mathcal{W}(x),$$

where Δ is the Laplacian, $\mathcal{W}(s)$ is a Gaussian white-noise random process, α controls the smoothness of the random field, τ controls the variance, and κ is a scale parameter. Based on this, [2] proposed a new approach to represent a Gaussian random field with Matérn covariance as a Gaussian Markov Random Field (GMRF), by representing a solution to the SPDE using the finite element method. This representation implies a sparse precision matrix for the spatial effects, allowing the implementation of fast computational methods to do inference.

4. Simulation and Results

For $n = 100$ sampled locations x_i, we take measurements of the simulated processes S. In preferential sampling scenarios, points are a realization of $X|S \sim$ Poisson Process$(\lambda(x))$, s.t. $\lambda(x) = \exp\{\alpha + \beta S(x)\}$ with $\beta > 0$. However, in non-preferential sampling scenarios, $X \sim$ Poisson Process$(\lambda(x))$ s.t. $\lambda(x) = \exp\{\alpha\}$.

Different scenarios were considered, but here we will present just two; one for preferential sampling and another one for non-preferential sampling. In all cases, we set $\mu = 0$ and $\sigma^2_\epsilon = 1$. Then, after generating data, we fit Model (1) under the assumption that X and $S(X)$ are **independent** (`A1`) and **dependent** (`A2`).

For instance, Figure 1 shows simulated scenario under preferential sampling and fitted models based `A1` and `A2`. Visual inspection suggests better results for model `A2`.

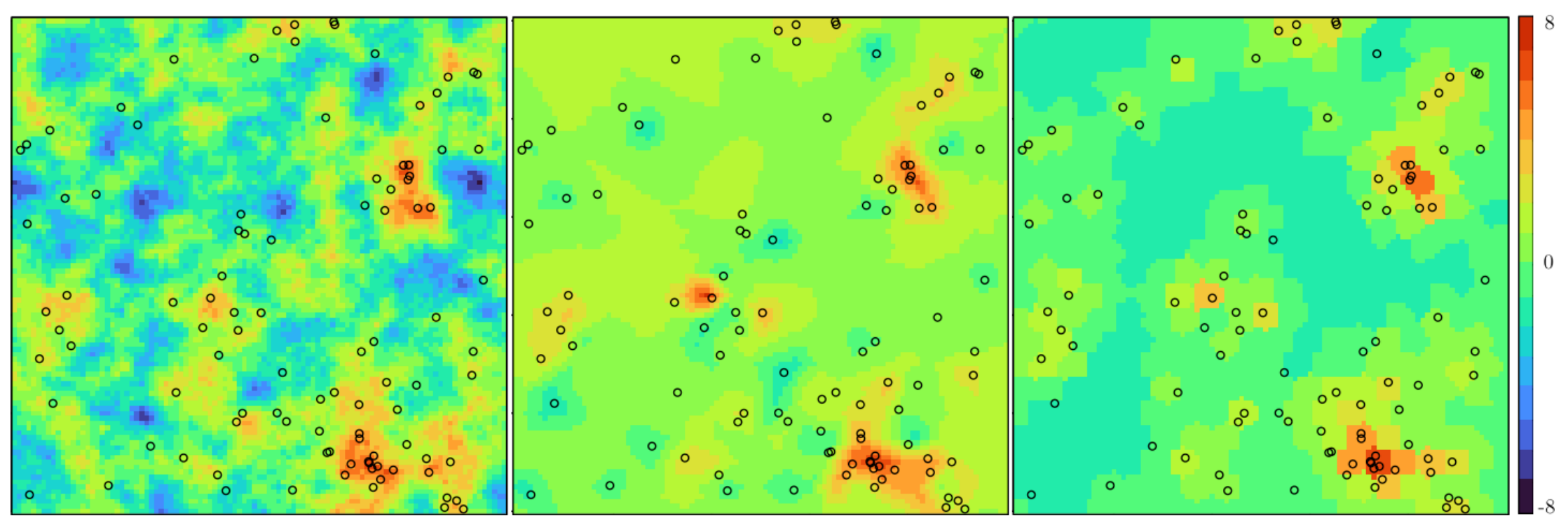

Figure 1: Simulated S and $X|S$ processes (left) with estimation based on models `A1` (center) and `A2` (right).

For the two simulated scenarios and the two fitted models, we assessed the performance of `A1` and `A2` by computing the Mean Squared Error (MSE). We simulated $m = 50$ data sets from each of the two scenarios and computed the mean and quantiles of the MSEs (Table 1).

Scenario	Model	Mean (SD) of MSE	5th—95th perc. of MSE
Non-pref. sampl.	`A1`	3.16 (0.33)	2.70—3.75
Non-pref. sampl.	`A2`	3.59 (0.48)	3.03—4.50
Prefer. sampling	`A1`	5.63 (1.07)	4.19—7.80
Prefer. sampling	`A2`	3.44 (0.59)	2.80—4.80

Table 1: Computed statistics for the MSEs for models `A1` and `A2` in the two scenarios.

From the table, for the scenario in which preferential sampling was **not** considered for the data generation procedure, `A1` performed *slightly* better than `A2`; however, for data generated with preferential sampling, model `A2` performed *much* better than model `A1` (w.r.t. the MSE).

5. Conclusions

Careful consideration of the obtained data is needed in order to determine the most appropriate modeling approach in each situation. Sometimes, we will need to account for preferential sampling if the obtained sample depends on the underlying spatial process. In other situations, though, standard geostatistical models will suffice to obtain valid inferences.

References

[1] Peter J Diggle, Raquel Menezes, and Ting-li Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232, 2010.

[2] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[3] Paula Moraga. *Geospatial health data: Modeling and visualization with R-INLA and Shiny*. CRC Press, 2019.

[4] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[5] Peter Whittle. Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994, 1963.