Bayesian meta-learning approach for feasible large spatial analysis

Luca Presicce

Ph.D. student in Statistics

University of Milan-Bicocca, Department of Economics, Management & Statistics

SIS 2024 - 52nd Scientific Meeting of the Italian Statistical Society, June 17-20, 2024, Bari (Italy)

This is a joint work with

Sudipto Banerjee

University of California, Los Angeles

- ☞ Y. Yao, A. Vehtari, D. Simpson and A. Gelman (2018) "Using stacking to average Bayesian predictive distributions", Bayesian analysis, vol. 13.
- ☞ S. Banerjee (2020) "Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework", Spatital Statistics, vol. 37.
- ☞ L. Zhang, W. Tang and S. Banerjee (2023) "Exact Bayesian Geostatistics Using Predictive Stacking", arXiv preprint, arXiv:2304.12414.

Issues in Geostatistical Models

Geostatistical modeling is afflicted by onerous computational effort when the number of locations is vast (the so-called “Big-n” problem).

- ☞ Despite extensive literature, spatial inference remains unfeasible for moderate data sets on modest computing environments.
- ☞ Our efforts fall under “meta-” learning: split a data set into smaller sets, and combined local results to approximate full Bayesian inference.
- ☞ We introduce Bayesian predictive stacking (BPS) in spatial meta-analysis, providing feasible uncertainty quantification on modest computing architectures.

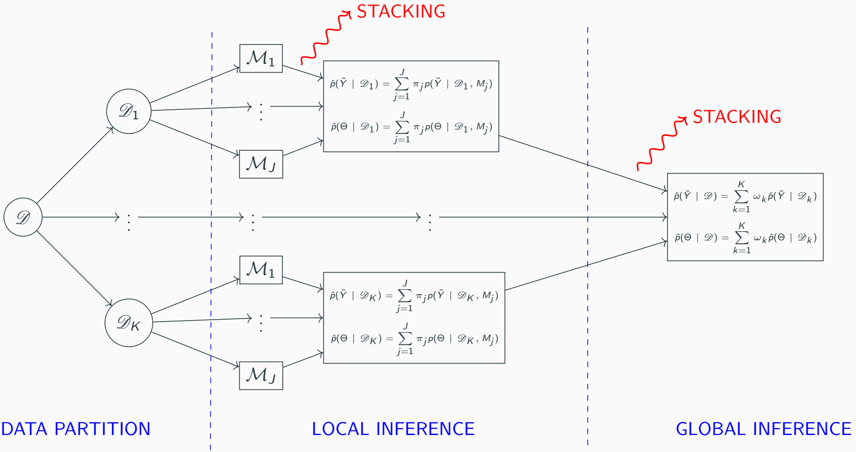

Figure 1: Double Bayesian predictive stacking approach representation

Dimensions:

- 1 million train locations
- 2000 partitions (500 locations each)
- 2500 holdout locations
- 5 computational cores

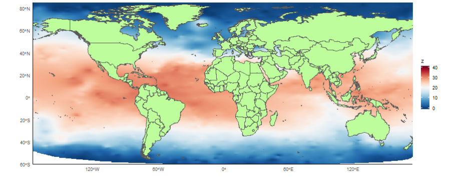

Figure 2: Holdout data surface interpolation for SST data analysis.

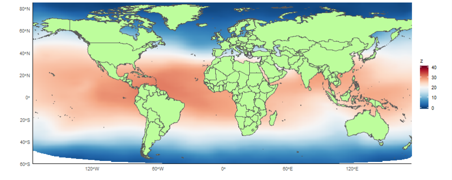

Figure 3: Predicted (MAP) surface interpolation for SST data analysis.

Achievements:

- ≈ 60 minutes (1 hour)
- RMSPE almost seven times lower w.r.t. Bayesian conjugate model

spBPS: Accelerated Spatial Modeling by Bayesian Predictive Stacking

Let me conclude by presenting the novel R package created, named spBPS

- ☞ Introduce BPS framework for univariate, and multivariate, geostatistical modeling.
- ☞ Use Rcpp/C++ -based code, allowing faster and scalable parallel computations.
- ☞ Available @lucapresicce/spBPS on GitHub, (and soon on CRAN).

Check it out on my GitHub!

Thanks for your attention!

Univariate Spatial regression - Latent model

Let consider

- ☞ $\mathcal{S} = \{s_1, \dots, s_n\} \subset \mathcal{D}$ be a set of n locations,
- ☞ $y = [y(s_i)]^\top$ be $n \times 1$ vector (for $i = 1, \dots, n$),
- ☞ $X = [x(s_i)^\top]$ be $n \times p$ matrix full rank p (for $i = 1, \dots, n$).

Such data can be modeled using:

$$\begin{aligned} &y = X\beta + \omega + e_y \ , \quad e_y \sim \mathrm{N}(0, \delta^2\sigma^2\mathbb{I}_n) \ , \quad \omega \mid \sigma^2 \sim \mathrm{N}\left(0, \sigma^2\rho_\phi(\mathcal{S}, \mathcal{S})\right) \ ; \\ &\beta = \mu_\beta + e_\beta \ , \quad e_\beta \sim \mathrm{N}(0, \sigma^2 V_\beta) \ ; \quad \sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma). \end{aligned} \tag{1}$$

Where

- ☞ $\delta^2 := \tau^2/\sigma^2 \in [0, 1]$ is the noise-to-spatial variance ratio,
- ☞ $\omega = [\omega(s_i)]^\top$ is $n \times 1$ latent process (for $i = 1, \dots, n$),
- ☞ $\rho_\phi(\mathcal{S}, \mathcal{S})$ be the $n \times n$ spatial correlation matrix,
- ☞ $\phi \in \mathbb{R}^+$ index spatial correlation function $\rho_\phi(\cdot, \cdot)$.