ExaGeoStatR: Harnessing HPC Capabilities for Large Scale Geospatial Modeling Using R

Sameh Abdulah, Yuxiao Li, Jian Cao, Hatem Ltaief, David E. Keyes, Marc G. Genton, and Ying Sun
Extreme Computing Research Center, King Abdullah University of Science and Technology

ExaGeoStatR

Motivation

- Gaussian random fields (GRFs) are valuable mathematical tools to fit spatial datasets for applications dealing with measurements regularly or irregularly located across geographical regions.
- The Exascale GeoStatistics project (ExaGeoStat) aims at optimizing the likelihood function for spatial data on massively parallel systems.
- The likelihood evaluation may require generating a covariance matrix Σ and performing the Cholesky factorization of Σ, which necessitates $O(n^3)$ operations and $O(n^2)$ memory. This implies that some of these standard methods for GRFs are not capable to cope with large datasets.
- The objective of ExaGeoStat is to enable statisticians to tackle computationally challenging scientific problems at large scale, while abstracting the hardware complexity.
- ExaGeoStatR further raises the game by providing wrappers in R, i.e., the *de facto* programming environment for statisticians, to the ExaGeoStat main functions. This renders large scale simulations feasible on massively parallel systems, while keeping the simplicity of the R environment.

Problem Statement

- Suppose Z is a stationary GRF and the main domain $D \subset \mathbb{R}^d$ at n locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$. The random vector $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}^T$ follows a multivariate Gaussian distribution:

$$\forall \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D, \quad \{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}^{\mathrm{T}} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma), \tag{1}$$

where $\boldsymbol{\mu}$ and Σ are the mean vector and the covariance matrix of the n-dimensional multivariate normal distribution.
- Given $\boldsymbol{\mu}$ and Σ, the likelihood of observing $\mathbf{z} = \{z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)\}^{\mathrm{T}}$ at the n locations is

$$L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\} \tag{2}$$

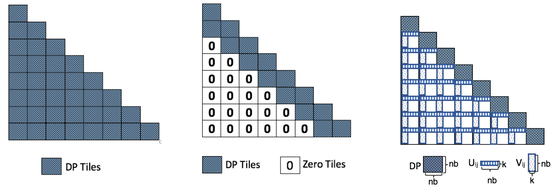
- The (i, j)-th element of Σ is $\Sigma_{ij} = C(\mathbf{s}_i, \mathbf{s}_j)$, where the covariance function $C(\mathbf{s}_i, \mathbf{s}_j)$ is assumed to have a parametric form with unknown vector of parameters $\boldsymbol{\theta}$.
- We choose the isotropic Matérn covariance kernel,

$$\Sigma_{ij} = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\beta}\right)^{\nu} \mathcal{K}_\nu \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\beta}\right), \tag{3}$$

where $\|\mathbf{s}_i - \mathbf{s}_j\|$ is the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, and the variance, spatial range, and smoothness parameters, i.e., σ^2, β, and ν, respectively, determine the properties of the GRF.

ExaGeoStatR Features

- ExaGeoStatR variant computation techniques:

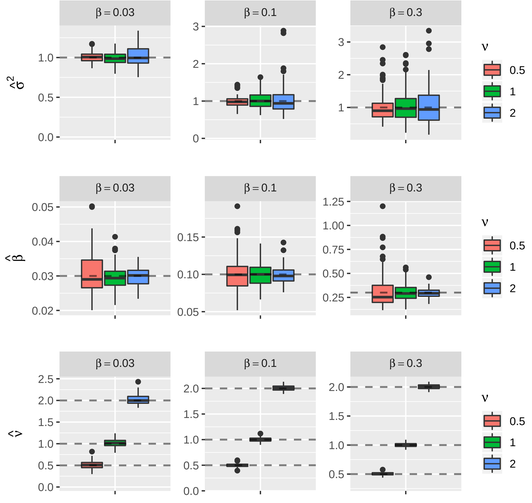


Exact | Diagonal Super Tile (DST) | Tile Low Rank (TLR)

- Overview of ExaGeoStatR functions:

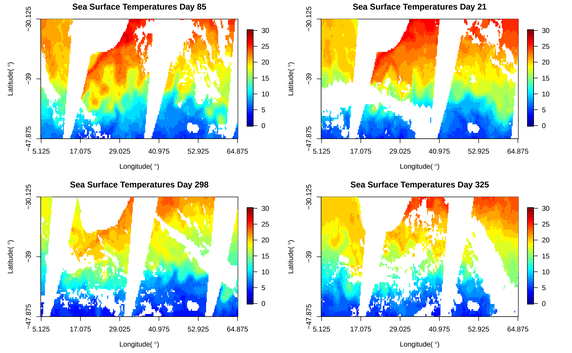
Function Name	Description
exageostat_init	Initiate ExaGeoStat instance.
simulate_data_exact	Generate **Z** measurements vector.
simulate_obs_exact	Generate **Z** measurements vector on n given 2D locations.
exact_mle	Exact parameter vector evaluation.
dst_mle	DST approximation parameter vector evaluation.
tlr_mle	TLR approximation parameter vector evaluation.
exageostat_finalize	Finalize active ExaGeoStat instance.

Synthetic Data Accuracy Estimation

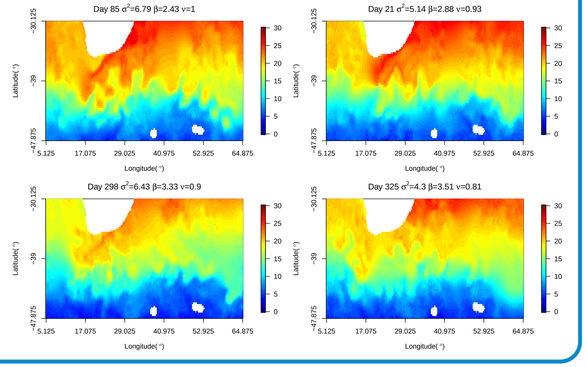


Sea Surface Temperature Data

- The sea surface temperature collected by satellite for the Agulhas and surrounding areas off the shore of South Africa.
- The region is abstracted into 72 x 240 regular grid. The spatial resolution is approximately 25 kilometers.
- The data covers 331 days, from Jan. 1st to Nov. 26th, 2004. We selected four days, day 21, day 85, day 298, and day 325.
- The original sea surface temperatures in Celsius (°C) where the locations with unavailable values are colored in white.

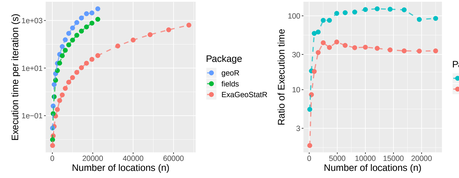


- The predicted sea surface temperatures based on the linear mean structure and the kriging results where the land area is not predicted and colored in white.

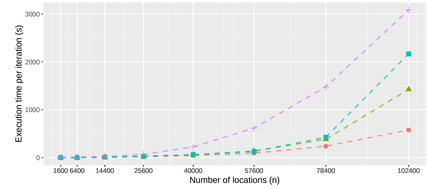


Performance Results

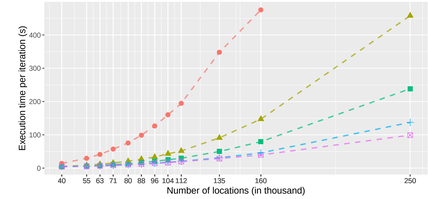
- Performance on a dual-socket 8-core Intel Sandy Bridge processor running at 2.0 GHz:



- Performance on a dual-socket 14-core Intel Broadwell processor running at 2.4 GHz equipped with 8x NVIDIA K80s:



- Performance on Cray XC40 system, each node is a dual-socket 16-core Intel Haswell processor running at 2.3 GHz:



References

- S. Abdulah, H. Ltaief, Y. Sun, M. G. Genton, and D. E. Keyes. "ExaGeoStat: A high performance unified software for geostatistics on manycore systems." IEEE Transactions on Parallel and Distributed Systems 29, no. 12 (2018): 2771-2784.
- D. Nychka, R. Furrer, and S. Sain. "fields: Tools for spatial data." National Center for Atmospheric Research (2005).
- P. J. Ribeiro and P. J. Diggle. "geoR: Analysis of geostatistical data." R package version 1.5.2 (2016).