The Third Competition on Spatial Statistics for Large Datasets

Yiping HONG, Yan SONG, Sameh ABDULAH, Ying SUN, Hatem LTAIEF, David E. KEYES, and Marc G. GENTON

Given the computational challenges involved in calculating the maximum likelihood estimates for large spatial datasets, there has been significant interest in the research community regarding approximation methods for estimation and subsequent predictions. However, prior studies examining the evaluation of these methods have primarily focused on scenarios where the data are observed on a regular grid or originate from a uniform distribution of locations. Nevertheless, non-uniformly distributed locations are commonplace in fields like meteorology and ecology. Examples include gridded data with missing observations acquired through remote sensing techniques. To assess the reliability and effectiveness of cutting-edge approximation methods, we have initiated a competition focused on estimation and prediction for large spatial datasets with non-uniformly distributed locations. Participants were invited to employ their preferred methods to generate corresponding confidence and prediction intervals for synthetic datasets of varying sizes and spatial configurations. This competition serves as a valuable opportunity to benchmark and compare different approaches in a controlled setting. We evaluated the submissions from 11 different research teams worldwide. In summary, the Vecchia approximation and the fractional SPDE methods were among the best performers for estimation and prediction. Furthermore, the nearest neighbors Gaussian process and the multi-resolution approximation exhibited excellent performance in predictive tasks. These findings provide valuable guidance for selecting the most appropriate approximation methods based on specific data characteristics.

Supplementary materials accompanying this paper appear online.

Key Words: Confidence interval; Matérn covariance model; Non-uniform distributed locations; Prediction interval; Scoring rule; Spatial statistics.

Y. Hong

Beijing Institute of Technology, No. 5, South Street, Zhongguancun, Haidian District, Beijing, China.

Y. Song · S. Abdulah · Y. Sun · H. Ltaief · D. E. Keyes · M. G. Genton (✉),

King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

(E-mail: *marc.genton@kaust.edu.sa*).

1. INTRODUCTION

The primary and foremost challenge in fitting and analyzing geospatial datasets is often the estimation of spatial models and the ensuing predictions. When dealing with large spatial datasets, modeling becomes computationally complex, particularly when utilizing popular statistical techniques such as maximum likelihood estimation (MLE). The computational challenges of MLE come from the computation of the covariance matrix inverse, a process that demands $O(n^3)$ computational steps and utilizes $O(n^2)$ memory, where n is the number of locations. To mitigate this challenge, extensive literature explored various approximation methods, for instance, composite likelihood (Vecchia 1988), Gaussian predictive process (Finley et al. 2009), covariance tapering (Furrer et al. 2006; Kaufman et al. 2008), and tile low-rank approximations (Abdulah et al. 2018b); see also the reviews in Sun et al. (2012) and Heaton et al. (2019). A significant portion of these studies focused on data observed either on a regular grid or on irregular locations that are evenly distributed, such as when observations are generated from a uniform distribution (Finley et al. 2009; Vecchia 1988; Abdulah et al. 2018b). However, observations in different locations may have varying densities in practical applications. For example, Shi and Cressie (2007) investigated Aerosol Optical Depth (AOD) data collected from the Multi-angle Imaging SpectroRadiometer (MISR) camera, where the observations are shaped like various swaths due to the satellite sampling process. Datta et al. (2016) examined a US biomass dataset, which has two blank squares because the biomass data in Wyoming and New Mexico were not yet released.

In order to ensure the robustness of approximation methods, it is crucial to conduct comprehensive analyses from various aspects, including the distribution of locations in the given spatial domain. This is because some studies have shown that certain approximation methods may not be suitable in certain situations. For example, Stein (2014) demonstrated that low-rank methods might yield poorer performance when neighboring observations exhibit strong correlations compared to the naive independent block approximation method. Consequently, it is important to verify or generalize the approximation methods when dealing with unevenly located datasets. Existing literature primarily emphasizes the generalization of approximation methods, particularly spectral-based methods, for which the estimation originally works for regularly gridded data. For instance, Fuentes (2007) proposed a version of Whittle's likelihood approximation suitable for regular grids with missing data and irregularly spaced data. Another study by Bandyopadhyay et al. (2015) formulated a spatial frequency domain empirical likelihood method for irregularly spaced data. Lu and Tjøstheim (2014) proposed a nonparametric kernel estimator for joint probability distribution functions for irregularly gridded stationary spatial data, which generalizes the nonparametric method for gridded data. Dealing with irregular observation locations with various application backgrounds has also drawn attention in the literature. For example, Huang et al. (2002) proposed an autoregressive tree-structured model to predict satellite data with multiple resolutions. Heaton et al. (2019) reviewed various approximation methods and conducted a competition focusing on lattice data with missing observations. Saas and Gosselin (2014) compared different autoregressive models for count data with irregular gridded observations from real-world data. The motivation of our article is to evaluate the robustness of differ-

ent approximation methods for various spatial data with more cases of unevenly spaced observations, such as clustered observation locations.

In addition to utilizing approximation methods, exact computation of MLE for large datasets becomes feasible with the aid of modern High-Performance Computing (HPC) systems. An example of such capability is demonstrated by the *ExaGeoStat*[1] software, which leverages HPC techniques to enable parallel generation, modeling and prediction of large geospatial datasets using covariance matrices (Abdulah et al. 2018a). By utilizing *ExaGeoStat*, generating large synthetic datasets, such as a dataset with a size of 1 M, with various types of observation locations and covariance structures, becomes feasible.

In 2021 and 2022, we organized two competitions to evaluate the performance of existing methods and tools in estimating and predicting pre-generated spatial and spatio-temporal datasets. These datasets were generated synthetically from specified statistical models, specifically covariance models, using the *ExaGeoStat* software. All the datasets used in these two competitions are now available for download from Huang et al. (2021b) and Abdulah et al. (2022a) as benchmarks for new methods. Furthermore, detailed descriptions of the datasets, the performance of different methods, and discussions of the competition results can be found in Huang et al. (2021a) and Abdulah et al. (2022b).

This year, we organized a third competition with distinct goals and datasets. We extended the data generation capability within *ExaGeoStat* to enable the generation of non-uniform distributed data locations. Unlike previous competitions focused on point estimation and prediction, the 2023 competition revolves around constructing confidence intervals for parameter estimation and prediction intervals. Specifically, the competition comprises four sub-competitions, namely 1a, 1b, 2a, and 2b. Sub-competitions 1a and 1b require participants to provide confidence intervals for parameter estimation, while Sub-competitions 2a and 2b involve providing prediction intervals. The datasets in the competition were generated using stationary Gaussian random fields, employing an isotropic Matérn covariance function, and encompassing diverse designs of irregularly spaced locations. To cater to participants with varying computing resources, we offered two training dataset sizes: 90 K and 900 K, and two testing dataset sizes: 10 K and 100 K. The interval score metric evaluates the accuracy of confidence interval results in estimation and prediction. Subsequently, a ranking strategy was employed to determine the competition winner based on performance. This year's competition garnered submissions from 11 teams, each participating in one or more sub-competitions. Additionally, *ExaGeoStat* participated as a benchmark. For clarity, we refer to the teams by their designated team numbers, as listed in Table S1 in the Supplementary Material with details of team members. The results of the competition can be summarized as follows: Team4 emerged as the winner in Sub-competition 1a (smaller samples) for parameter estimation, while Team2 and Team6 claimed a tied victory in Sub-competition 1b (larger samples). For the prediction aspect, Team2 also secured the first position in Sub-competition 2a (smaller samples), while Team8 triumphed in Sub-competition 2b (larger samples).

Based on the competition results, the performances of the winning teams indicate that the Vecchia approximation and the fractional SPDE methods were among the best performers for both estimation and prediction. Therefore, considering the results of the competition,

---

[1]https://github.com/ecrc/exageostat.

we recommend employing specific approximation methods for different sample sizes and shapes of observation locations. This includes prioritizing using Vecchia approximations and fractional SPDE and considering the nearest neighbors Gaussian process and the multi-resolution approximation for prediction tasks.

The subsequent sections of this article are structured as follows. Section 2 outlines the dataset configurations utilized in our competition. Section 3 presents an overview of the competition objectives and the evaluation methods employed to assess the performance of each submitted result. Section 4 provides a detailed introduction to the methods used in the submitted results. Section 5 presents a comprehensive analysis and comparison of each submitted method, leading to the identification of the best-performing approaches. We provide the conclusions and discussions derived from the competition in Sect. 6.

2. COMPETITION DATASETS

The simulated datasets for the 2023 competition on spatial statistics for large datasets were generated using the stationary isotropic Matérn covariance model, considering various configurations of observed locations. The following subsections provide more detailed information about the datasets corresponding to different observed location settings. In all sub-competitions (Sub-competitions 1a, 1b, 2a, and 2b), we assume that the spatial data $Z(s_i), i = 1, \ldots, n$, follow a zero-mean stationary isotropic Gaussian random field with Matérn covariance:

$$C(\|s_i - s_j\|) := \text{Cov}\{Z(s_i), Z(s_j)\} = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|s_i - s_j\|}{\beta}\right)^\nu \mathcal{K}_\nu \left(\frac{\|s_i - s_j\|}{\beta}\right) + \tau^2 \mathbb{1}_{\{s_i = s_j\}},$$
(1)

where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν, $\Gamma(\cdot)$ is the Gamma function, $\sigma^2, \beta > 0, \nu > 0$, and τ^2 are the variance parameter, range parameter, smoothness parameter, and nugget variance parameter, respectively.

Our competition datasets specifically address scenarios where the observations are irregularly located. The observation locations are confined within the unit square $[0, 1]^2$. There are five different cases considered for the locations of the observations:

1. *Chessboard*: Generated observations on $[0, 1]^2$ with a chessboard shape;

2. *Left-bottom*: Generated observations on $[0, 1]^2$ with locations more concentrated in the left-bottom area;

3. *Satellite*: Scaled observed locations obtained from a satellite remote sensing scenario, with locations missing in some bands;

4. *Clusters*: Generated observations on $[0, 1]^2$ with locations having cluster shapes;

5. *Regular*: Generated observations are uniformly distributed on $[0, 1]^2$.

The first row of Fig. 1 displays five training sets in Sub-competitions 1a and 2a, which are of size 90 K and generated with the above location configurations and parameters shown in Table 1. Except for the Regular case, each training set has two testing sets of size 10 K.
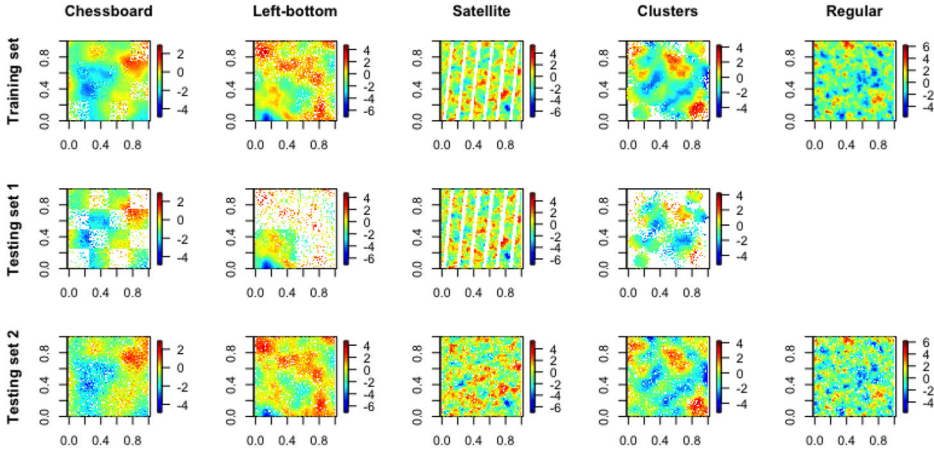
Figure 1. Illustration of competition datasets in Sub-competitions 1a and 2a. Training sets of size 90 K are displayed in the first row. Testing sets of size 10 K with the same distributions as the corresponding training sets are shown in the second row. Testing sets of size 10 K with uniform distributions are shown in the third row.

Table 1. True parameter values in the Matérn model for different types of locations

Type of locations	σ^2	β	Effective range	ν	τ^2
1. Chessboard	1.4	0.164056	0.8	1.6	0.27
2. Left-bottom	1.4	0.052579	0.3	2.3	0.27
3. Satellite	1.7	0.021707	0.1	1.4	0.27
4. Clusters	1.5	0.054710	0.3	2.1	0.27
5. Regular	1.7	0.021707	0.1	1.4	0.27

One has the same distribution as the training set, shown in the second row, but the other has uniform locations, shown in the third row. Similarly, Fig. 2 illustrates the training and testing sets in Sub-competitions 1b and 2b. With the same parameter settings, the numbers of locations in Fig. 2 are ten times larger than those in Fig. 1. All the observation values are generated using the *ExaGeoStat* software (Abdulah et al. 2018a). Moreover, Table 1 presents the true values of σ^2, ν, τ^2, and the effective range for different observations. The effective range is determined by finding the smallest distance h_{eff} at which the ratio $C(h_{\text{eff}})/C(0)$ equals 0.05. It is worth mentioning that the nugget term is disregarded (i.e., τ^2 is set to 0) while determining β based on the predefined effective range.

3. COMPETITION OBJECTIVES AND ASSESSMENT METHODS

Herein, we outline the purpose of our sub-competitions and the criteria we have chosen to evaluate and rank participating teams. The top-ranked teams should demonstrate proficiency in accurately determining estimation confidence intervals or prediction intervals within each sub-competition.
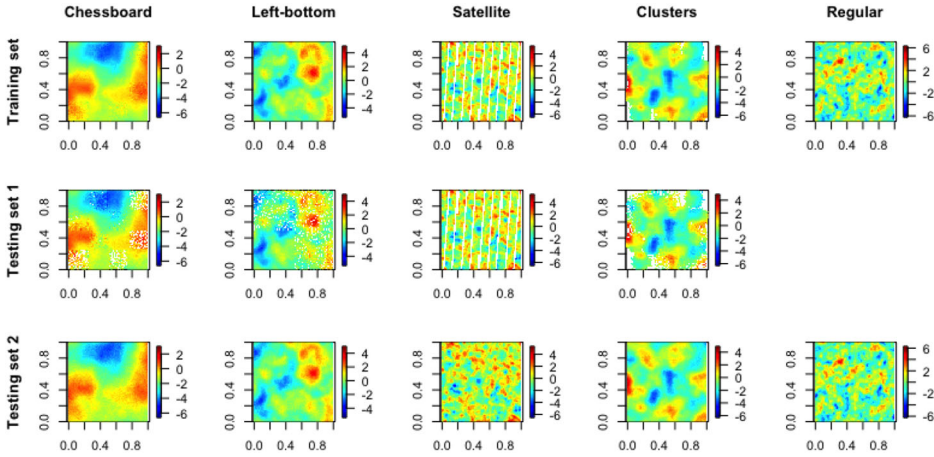
Figure 2. Illustration of competition datasets in Sub-competitions 1b and 2b. Training sets of size 900 K are displayed in the first row. Testing sets of size 100 K with the same distributions as the corresponding training sets are shown in the second row. Testing sets of size 100 K with uniform distributions are shown in the third row.

Table 2. Summary of the four sub-competitions

Sub-competition	Model setting	Target	# of location designs	Training data size	Testing data size
1a	Gaussian Matérn	Estimation (95% confidence interval)	5	90 K	–
1b	Gaussian Matérn	Estimation (95% confidence interval)	5	900 K	–
2a	Gaussian Matérn	Prediction (95% prediction interval)	5	90 K	10 K
2b	Gaussian Matérn	Prediction (95% prediction interval)	5	900 K	100 K

3.1. COMPETITION OBJECTIVES

According to Burt et al. (2020) and Song et al. (2022), location distributions significantly impact the convergence rate of a zero-mean Gaussian process. This motivated us to investigate parameter estimation and prediction under various location settings. Additionally, the existing theory is often developed under the assumption that the testing location distribution is the same as the training location distribution. Therefore, how methods perform when training and testing distributions are different is of interest. Sub-competitions 1a and 1b are dedicated to building confidence intervals for the unknown parameters of the Matérn covariance function in (1) using the provided datasets. Sub-competitions 2a and 2b require participants to generate prediction intervals for the given datasets at specific new locations within the testing datasets. We summarize the configurations of different datasets in Table 2.

For Sub-competitions 1a and 1b, participants were required to present independent 95% confidence intervals for each of the parameters σ^2, β, ν, and τ^2. Simultaneous confidence intervals for these parameters were not considered. In Sub-competitions 2a and 2b, the par-

ticipants were expected to predict the missing measurements at the testing locations using their preferred tools or methods. Then, they had to provide independent 95% prediction intervals for each testing point. While the confidence and prediction intervals were mandatory for the competition, parameter estimation and prediction values for testing points were not required.

3.2. ASSESSMENT METHODS

We rely on the scoring rule defined by Gneiting and Raftery (2007) to evaluate the quality of the confidence and the prediction intervals. Let $[l, u]$ be the $(1 - \alpha)\%$ confidence or prediction interval corresponding to the true value z. The interval score is defined by

$$S_\alpha^{\text{int}}(l, u; z) = (u - l) + \frac{2}{\alpha}(l - z)\mathbb{1}_{\{z < l\}} + \frac{2}{\alpha}(z - u)\mathbb{1}_{\{z > u\}}, \tag{2}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $\alpha = 0.05$.

In the first part, Sub-competitions 1a and 1b, we considered the unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^\top = (\sigma^2, \beta, \nu, \tau^2)^\top$. For each parameter, let $[lc_i, uc_i]$ represents the corresponding 95% confidence interval, where $i = 1, 2, 3, 4$. To evaluate the performance of teams in Sub-competition 1a, assume that $K^{(1a)}$ teams participated. We define $S_{\alpha,k,j}^{\text{int}}(lc_i, uc_i; \theta_i)$ as the interval score of team k for θ_i in dataset j, where $k = 1, \ldots, K^{(1a)}$ and $j = 1, \ldots, 5$. Within each dataset, we sort the interval scores $S_{\alpha,k,j}^{\text{int}}(lc_i, uc_i; \theta_i)$ for each parameter and assign a rank $R_{\alpha,k,j}^{\text{int}}(lc_i, uc_i; \theta_i)$ to each team in ascending order. The rank score for team k and dataset j is defined by

$$R_{\text{est},k,j}^{(1a)} = \frac{1}{4} \sum_{i=1}^{4} R_{\alpha,k,j}^{\text{int}}(lc_i, uc_i; \theta_i).$$

Finally, for each dataset j, we sort $R_{\text{est},k,j}^{(1a)}$, $k = 1, \ldots, K^{(1a)}$ in ascending order and assign rank $Rank_{k,j}^{(1a)}$ to each team. The final score for team k in Sub-competition 1a was calculated as $S_k^{(1a)} = \sum_{j=1}^{5} Rank_{k,j}^{(1a)}$, and the final rank was assigned by sorting $S_k^{(1a)}$ in ascending order. A similar ranking strategy was used for Sub-competition 1b.

In the second part, Sub-competitions 2a and 2b, the prediction interval scoring rule was defined by

$$\text{IS}_{\text{pred}} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} S_\alpha^{\text{int}}(lp_i, up_i; Z_i),$$

where Z_i are the true realization values in the testing datasets, N_{test} is the total number of testing data points, $[lp_i, up_i]$ is the $(1 - \alpha) \times 100\%$ prediction interval of Z_i, and $S_\alpha^{\text{int}}(l, u; z)$ is the interval score defined by (2). Assuming that we have $K^{(2a)}$ participant teams for Sub-competition 2a, let $\text{IS}_{\text{pred},k,t}^{(2a)}$, $k = 1, \ldots, K^{(2a)}$, $t = 1, \ldots, 9$, denote the prediction interval score from team k for testing dataset t. For each testing dataset t, we sort

$\mathrm{IS}_{\mathrm{pred},k,t}^{(2a)}$, $k = 1, \ldots, K^{(2a)}$ in ascending order and assign rank $Rank_{k,t}^{(2a)}$ to each team. The final score for team k in Sub-competition 2a was calculated as $S_k^{(2a)} = \sum_{t=1}^{9} Rank_{k,t}^{(2a)}$, and the final rank was assigned by sorting $S_k^{(2a)}$ in ascending order. A similar ranking strategy was used for Sub-competition 2b.

Following the aforementioned ranking criteria, we present the competition outcomes for all sub-competitions in Tables S2–S5 of the Supplementary Material. Furthermore, we provide the rankings of each team for every dataset, granting teams valuable insights into the performance of their approaches across various scenarios. Within each sub-competition, we highlight the top three teams (or four teams if there is a tie for the third position) as follows:

- Sub-competition 1a: 1) Team4, 2) Team2, 3) Team6 and Team8;

- Sub-competition 1b: 1) Team2 and Team6, 3) Team5;

- Sub-competition 2a: 1) Team2, 2) Team6, 3) Team8;

- Sub-competition 2b: 1) Team8, 2) Team2, 3) Team6.

4. OVERVIEW OF METHODS IN SUBMITTED RESULTS

In this section, we provide an overview of the methods employed by the teams to construct confidence intervals for estimated parameters and prediction intervals. For each sub-competition, we delve into specific information regarding team implementations, with a particular focus on the top-ranked teams.

4.1. SUB-COMPETITIONS 1A AND 1B

For Sub-competitions 1a and 1b, the teams were tasked with assessing confidence intervals for Matérn covariance parameters using datasets of sizes 90 K and 900 K, respectively. The teams employing Bayesian methods can easily obtain confidence intervals through posterior distributions and Markov chain Monte Carlo. The teams utilizing frequentist methods need resampling techniques such as the bootstrap or theoretically deriving and calculating the asymptotic variance for parameter estimators to get confidence intervals. For example, *ExaGeoStat* evaluates the exact MLE for parameters and then computes the asymptotic variance of the MLE by the inverse of the Fisher information matrix, which is obtained by finite difference approximation. The confidence intervals are then evaluated by the asymptotic normality of the MLE. Table 3 summarizes the estimation methods different teams used in Sub-competitions 1a and 1b. Team5 provided three submissions for Sub-competition 1a, which they named (1-1), (1-2), and (2-3).

The Vecchia approximation (Vecchia 1988), classified under composite likelihood methods (Varin et al. 2011; Eidsvik et al. 2014), was the prevailing choice for parameter estimation in our competition. It approximates the joint likelihood with a series of conditional likelihoods of $Z(\mathbf{s}_i)$, where the smaller conditioning set includes its (at most) m nearest

Table 3.   Methods used by different teams in Sub-competitions 1a and 1b

Methods	SPDE	Vecchia approximation	Combinations and others
Teams	Team6	Team1, Team2, Team4, Team5 (1-1 and 1-2), Team7, Team8, Team9	Team3, Team5 (2-3), Team10

neighbors. Efficient implementations of the Vecchia approximation have been developed through various approaches and packages (Guinness 2018, 2021; Guinness et al. 2021). In this competition, Team5 (1-1) employed the method proposed by Guinness (2021). Team5 (1-2) relied on their custom-built R package, named Bootstrap for Rapid Inference on Spatial Covariances (BRISC) (Saha and Datta 2022). Team8 employed the Vecchia approximation (20 nearest neighbors) within the Nearest Neighbors Gaussian Process (NNGP) framework. NNGP is a rigorously-defined spatial process that offers valid finite-dimensional Gaussian distributions characterized by sparse precision matrices, founded on the principle that spatial correlation is most pronounced among neighboring observations (Datta et al. 2016). The range and smoothness parameters were assigned uniform priors, while the variance parameters were assigned inverse gamma priors. Additionally, teams such as Team1, Team2, Team7, and Team9 utilized the R package *GpGp* (Guinness et al. 2021) for their implementations. The main difference among them is the value of m. For example, Team1 and Team 9 employed the default option and $m = 30$, respectively. Team7 used $m = 30$ and 15 nearest neighbors in the procedures of parameter estimation and resampling, respectively. As the sample size n increased to 900 K, Team7 introduced a data reduction technique known as hexagonal binning. In this approach, the domain was divided into numerous hexagons, and the mean value of the data within each hexagon was used to represent its overall value. Subsequently, Team7 applied their Vecchia approximation to the hexagonal-binned data.

We now provide a detailed account of Team4's winning strategy in Sub-competition 1a, where they harnessed the power of the Vecchia approximation combined with a parametric bootstrap technique to construct 95% confidence intervals. Initially, they obtained the MLE of the parameter vector $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}$, using the Vecchia approximation with $m = 1000$. Next, they generated simulated data based on the assumption that $\hat{\boldsymbol{\theta}}$ represents the true values. Then, by applying the Vecchia approximation with $m = 300$ to the simulated data, they obtained another MLE estimate denoted as $\hat{\boldsymbol{\theta}}^{*}$. Finally, they obtained the inference result by repeating the above steps.

Team2, securing the first two places in both Sub-competitions 1a and 1b, also embraced the idea of combining bootstrap techniques with the Vecchia approximation. However, they incorporated some unique aspects into their approach. Team2 developed additional Rcpp-based functions to enhance their estimation, building upon the R package *GpGp*. These custom functions extended the capabilities of the existing fit_model() function in *GpGp* by enabling the estimation of Matérn parameters even when the mean of the spatial process is precisely zero. Team2 also made a distinct choice for the value of m. To elaborate, they first maximized a 10-neighbor approximation, then proceeded to maximize a 30-neighbor

approximation, utilizing the 10-neighbor estimates as initial values. Finally, they maximized a 60-neighbor approximation, using the 30 neighbor estimates as starting values. Additionally, Team2 utilized the fast_Gp_sim_Linv() function for simulating parametric bootstrap samples while obtaining the corresponding L-inverses through the vecchia_Linv() function. Furthermore, they employed the selm() function from the R package *sn* (Azzalini 2022) for each parameter to fit a skew-normal distribution to the Bootstrap estimates. They then calculated the 2.5% and 97.5% quantiles of the distribution using the qsn() function. In Sub-competition 1b, Team2 adjusted the procedure of generating bootstrap samples while keeping other procedures. Specifically, they computed the MLE 1000 times on sub-samples of 10,000 points from the given set of locations. They obtained the samples by giving uneven weights to each location to de-cluster them. First, the weight of a given point was calculated as the inverse of the total number of points in a specific neighborhood of that point. Next, they normalized the weights so that they sum to one. Then, they re-estimated the parameters for each "approximate" bootstrap sample.

In addition to the Vecchia approximation, Team6, who won (tie) the Sub-competition 1b and claimed the third place in Sub-competition 1a, employed the fractional SPDE method, an efficient Bayesian approach involving a sparse approximation for the precision matrix. Team6 exclusively employed the fractional SPDE approach in all sub-competitions, explicitly relying on the covariance-based rational approximations of fractional SPDEs as detailed in Bolin et al. (2023). The approach is based on the concept that the Gaussian field can be expressed as a solution to a fractional-order stochastic partial differential equation. To approximate this equation, the method combines a finite element approximation with a rational approximation of the fractional power. The implementation of the method can be found in the *rSPDE* package, which is accessible on CRAN. The package provides interfaces to both *R-INLA* (Rue et al. 2009) and *inlabru* (Bachl et al. 2019). In the competition, Team6 utilized the *inlabru* interface, resulting in all their outcomes being obtained through a Bayesian approach with hyperparameter priors. For further information on the methodology, please refer to the comprehensive explanation provided in Bolin et al. (2023). The specific choices for the priors and model approximations were as follows. The priors were chosen as rather flat log-normal priors for all parameters except the smoothness parameters, which had a truncated log-normal prior (as described in the vignette on the package homepage). The rSPDE model was the one with the SPDE parameterization. The mesh was generated using inla.mesh.2d() with cutoff, offset, and extension depending on the dataset.

Apart from the above-mentioned methods, Team3 employed a variational inference to approximate posterior densities for GP models. They estimated the spatial autocorrelation for each training dataset through a variogram and then used it to construct a new sampling strategy for cross-validating the parameters. Team5 (2-3) utilized a method described in Zhang et al. (2019), which combines the concepts of covariance tapering (Kaufman et al. 2008) and low-rank approximation (Cressie and Johannesson 2008; Banerjee et al. 2008). Team10 did not provide descriptions of their methods.

Table 4. Methods used by teams for evaluating the kriging predictors and variances in Sub-competitions 2a
and 2b. The low-rank and Vecchia methods were employed as low-rank and Vecchia approximations,
respectively

Methods	LaGP	Low-Rank	SPDE	Vecchia	Combinations and Others
Teams	Team1 Team9 (LaGP)	Team7 Team8	Team6	Team2 Team4 Team9 (Vecchia)	Team3 Team10 Team11

4.2. Sub-competitions 2a and 2b

In Sub-competitions 2a and 2b, teams were tasked with assessing prediction intervals for the provided testing sets, utilizing datasets of sizes 90 K and 900 K, respectively. We are particularly interested in the techniques employed to compute the inverse of a large covariance matrix in both kriging predictors and variances. These methods used by different teams are summarized in Table 4. Team9 provided two submissions in Sub-competition 2a, which were named Team9 (LaGP) and Team9 (Vecchia), respectively, due to their techniques. Additionally, *ExaGeoStat* was applied to Sub-competition 2a as a benchmark. ExaGeoStat (MLE) and (True) represent methods of calculating prediction intervals by MLE and true parameters, respectively.

Low-rank approximation methods are commonly used in the case of large spatial datasets. These methods approximate a process as a linear combination of a smaller number (q) of basis functions with random coefficients (Cressie and Johannesson 2008; Banerjee et al. 2008). The multi-resolution approximation (MRA) is one of these methods. As described by Katzfuss (2017), the MRA approach constructs basis functions of varying resolutions through recursive partitioning of the spatial domain. The basis functions are induced with covariance functions and knots, resulting in an effective approximation technique. Katzfuss and Guinness (2021) integrated the multi-resolution approximation (MRA) into a broader framework of Vecchia approximation. They further developed the R package *GPvecchia*, which encompasses the MRA implementation and was utilized by Team7. Moreover, Team8 relied on a parallelizable implementation of MRA proposed by Huang et al. (2019). The MRA was configured to divide partitions into two at every resolution and to have 49 knots within each partition.

The local approximation Gaussian process (LaGP), proposed as a solution to the prediction problem in large-scale Gaussian processes, offers an effective approach. It predicts values at any given location **s** by training a Gaussian process on a small subset of the data. This subset can either consist of the nearest neighbors of **s** or be selected to minimize the mean squared prediction error at **s**. Team1 and Team9 (LaGP) utilized the R package *laGP* (Gramacy 2016) for their implementations.

While the Vecchia approximation has proven to be successful in Sub-competitions 1a and 1b, it is worth noting that Team2 also employed this method in Sub-competitions 2a and 2b. Therefore, we explore how they utilized Vecchia approximation for these sub-competitions.
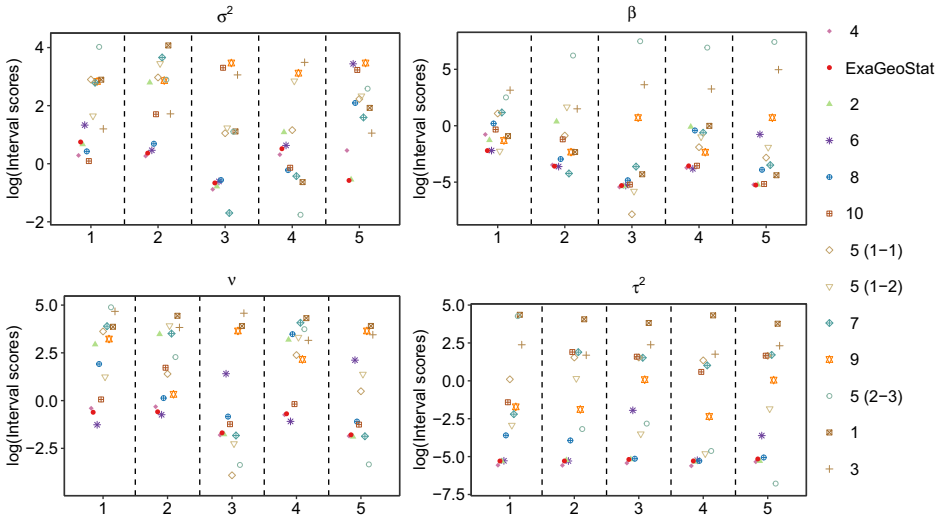
Figure 3. Interval scores (log scale) for confidence intervals of each parameter of the Matérn model submitted by teams in Sub-competition 1a. The order of the color/team legend goes from best to worst performance.

Team2 used their method to estimate parameters in Sub-competition 1a as a first step. They then obtained kriging predictions by employing an adjusted version of the function predictions() in the *GpGp* package. This adjusted version allowed for reordering and was used with a selection of the 200 nearest neighbors ($m = 200$). Next, leveraging the marginal distributions of the predictive samples, Team2 used the function cond_sim() from the *GpGp* package to simulate 1000 predictive samples. Lastly, they derived the predictive variances and computed pointwise prediction intervals with a confidence level of 95%.

Besides the above popular techniques, various methods were used in Sub-competitions 2a and 2b. Team3 evaluated the parameters by fixing $v = 2.5$ for all datasets, and then calculated the prediction intervals based on the estimation. Team11 employed the gradient-boosting decision tree (GBDT) method using package *lightgbm* (Ke et al. 2017). They randomized the training data and repeated the training for 2000 times to get a Monte Carlo mean and variance estimate.

5. COMPETITION RESULT ANALYSIS

In previous sections, we have demonstrated and outlined the approaches employed by the teams, along with presenting the rankings for each sub-competition. Now, we analyze the competition results, specifically the interval scores, to facilitate a thorough comparison and analysis of team performances. It is worth noting that a smaller interval score corresponds to a more favorable estimated interval. Figure 3 shows the values of $IS^{int}_{0.05,k,j}(lc_i, uc_i; \theta_i)$, where $k = 1, \ldots, 13$, $j = 1, \ldots, 5$, and $i = 1, \ldots, 4$, i.e., interval scores for confidence intervals of each parameter submitted by teams in Sub-competition 1a. Although we designed various location distributions to assess parameter evaluation difficulty, it is difficult to determine which dataset posed the greatest challenge.
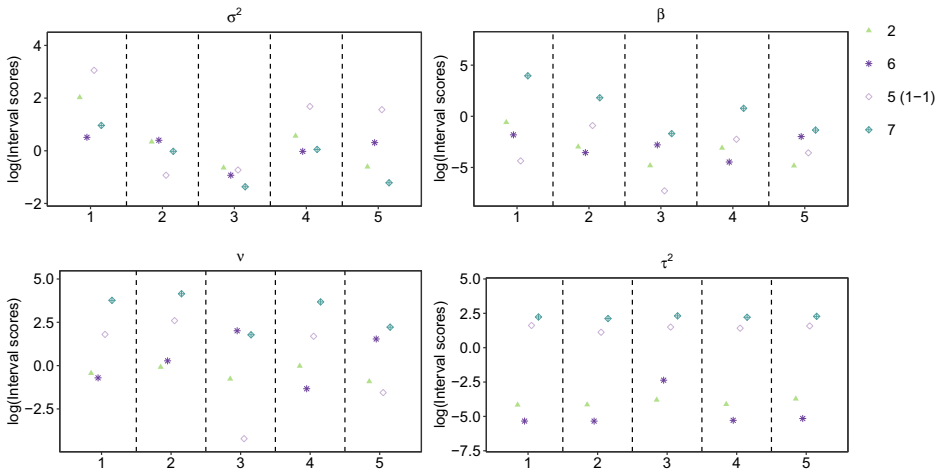
Figure 4.   Interval scores (log scale) for confidence intervals of each parameter of the Matérn model submitted by teams in Sub-competition 1b. The order of the color/team legend goes from best to worst performance.

Table 3 reveals that numerous teams employed the Vecchia approximation method and the R package *GpGp*. However, by examining Fig. 3, significant gaps can be observed when comparing the interval scores of Team2, Team4, and Team8 with those of Team1, Team7, and Team9. The gaps may come from the number of nearest neighbors. For example, Team4 used 1000 and 300 nearest neighbors in parameters evaluation and bootstrap, respectively. Team2 initially used $m = 10$ and 30 for obtaining initial parameters but ultimately settled on $m = 60$. In contrast, Team1, Team7, and Team9 employed a maximum of 30 nearest neighbors throughout the entire procedure. There could be other factors contributing to the observed gaps. For instance, Team1 constructed confidence intervals using the information matrix, while other methods relied on resampling techniques. Also, Team5 (2-3) used the method of Zhang et al. (2019), combining low-rank approximation and covariance tapering to obtain a full-scale approximation. Their poor performance in this competition may be because of our varied location distributions shown in Fig. 1. Albeit using exact MLEs for parameters, *ExaGeoStat* is not quite the winner in Sub-competition 1a, since it calculates confidence intervals based on a derived asymptotic Fisher information matrix of parameters. Some methods may provide over-optimistic confidence intervals compared to the MLE-derived interval. Such intervals can behave better for one case, but cannot be best for all parameters and all sample settings. Note that the y-axis scales of the sub-plots are different. This is why we opted to average the ranking scores across all parameters rather than relying solely on interval scores. The first column in Fig. 3 shows a larger point concentration in the top half region. This indicates that estimating the parameters σ^2 and ν is relatively more challenging.

Figure 4 shows the interval scores in Sub-competition 1b. When comparing Fig. 4 to Fig. 3, it can be observed that Team2 exhibits improved performance as the sample size increases to 900 K. As the sample size increases, the nearest neighbors in a fixed domain tend to become even closer, potentially providing additional information for parameter evaluation. Recall that Team7 employed a data reduction technique known as hexagonal binning.
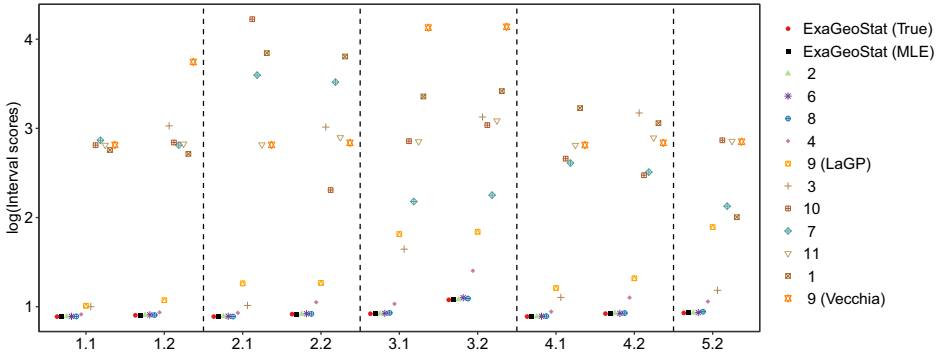
Figure 5. Averaged interval scores (log scale) for prediction intervals among testing points submitted by teams in Sub-competition 2a. In x-axis, the first number ranging from 1 to 5 represents location settings illustrated in Table 1, and the second number 1 or 2 represents testing sets 1 or 2, respectively. The order of the color/team legend goes from best to worst performance.

This step enhanced computational efficiency but also resulted in some loss of information. The consistent high performance of Team6 highlights the advantages of their method. Additionally, it is worth noting that all the data were generated with Matérn covariance, which aligns perfectly with their chosen method.

We provided two testing sets for Sub-competitions 2a and 2b for each training set. The first testing set maintained the same location distributions as the corresponding training set, while the second testing set followed a uniform distribution across the entire domain. Figure 5 demonstrates the values of $IS_{\text{pred},k,t}^{(2a)}$, where $k = 1, \ldots, 13$ and $t = 1, \ldots, 9$, i.e., the averaged interval scores for prediction intervals among testing points in Sub-competition 2a. The values are also provided in Table S6 for better comparison. Based on Fig. 5, it can be observed that most teams exhibit slightly higher interval scores on testing set 2, indicating better performance on testing set 1 compared to testing set 2. This observation is reasonable considering that testing set 2 follows a different distribution from the training set. An interesting aspect to consider is how to evaluate the impact of such distributional differences on interval scores. In almost all case, both ExaGeoStat (MLE) and ExaGeoStat (True) dominate Sub-competition 2a, with exact MLE and true parameters, as well as a closed-form prediction interval expression. In some testing sets 2, i.e., "1.2", "2.2", and "3.2", ExaGeoStat (MLE) performs slightly better than ExaGeoStat (True), which is a consequence of the estimation error for prediction mean squared errors on testing points. Team3 kept $\nu = 2.5$ fixed during parameter estimation and subsequently calculated prediction intervals based on those estimates. By comparing 2.5 with values of ν in Table 1, we can see a worse parameter estimation, which further affects their prediction intervals, especially for testing set 2. Recall that both Team4 and Team2 used the Vecchia approximation, and Team4 behaves better in Sub-competition 1a. However, Team4 is worse than Team2 in this sub-competition. Team4 calculated the theoretical prediction intervals, whereas Team2 used the bootstrap to obtain empirical prediction intervals.
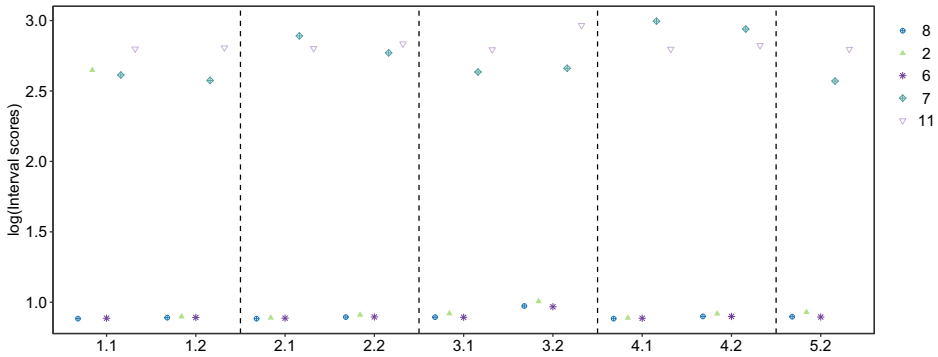
Figure 6. Averaged interval scores (log scale) for prediction intervals among testing points submitted by teams in Sub-competition 2b. In x-axis, the first number ranging from 1 to 5 represents location settings illustrated in Table 1, and the second number 1 or 2 represents testing sets 1 or 2, respectively. The order of the color/team legend goes from best to worst performance.

Figure 6 illustrates the averaged interval scores for Sub-competition 2b. Comparing Fig. 6 with Fig. 5, interval scores slightly decrease because of the increase in data size. Team7 and Team11 might consider exploring additional parameter tuning to enhance their performance.

6. CONCLUSION AND DISCUSSION

We have conducted a competition to explore the estimation and prediction performances of the Matérn covariance model, considering various distributions of irregularly located observations. In this article, we presented concise introductions to the approximation, estimation, and prediction methods employed by each participating team, followed by an evaluation of their performances using interval scores. Based on the results, we observed that the method employed by Team2 is suitable when the observation locations follow a Regular pattern or pertain to remote sensing applications. This particular method demonstrated excellent prediction performance and competitive estimation performance. In these settings, the observation locations are distributed more evenly compared to other types, such as Chessboard, Left-bottom, and Clusters configurations. We observed that Team6's method is suitable for unevenly distributed observation locations. This method demonstrated overall good performance in both estimation and prediction across various sample sizes and observation types. For spatial data with smaller sizes and unevenly distributed locations, we observed the Team4's method is suitable for estimation. This method has proven to be the best performer in Sub-competition 1a and exhibited superior interval scores compared to Team2's method, except for scenarios involving a Regular pattern and remote sensing observations. If the primary focus is solely on prediction performance, the Team8's method is suitable. This method emerged as the winner in Sub-competition 2b, and the difference between its performance and other winning methods in Sub-competition 2a was minimal.

Considering the methods employed by the winners of this competition, the Vecchia approximation and the fractional SPDE method appeared suitable for both estimation and prediction tasks concerning spatial data. Additionally, one can consider the NNGP method

and its corresponding multi-resolution approximation method for prediction purposes. Given that the fractional SPDE method proved to be effective for various observation locations in our competition, and considering its relatively lesser prominence in the literature compared to methods that primarily approximate the likelihood function or covariance matrix, we will further explore the estimation accuracy and prediction efficiency of the fractional SPDE method when dealing with irregularly observed data.

ACKNOWLEDGEMENTS

Declarations

Competing interests No conflicts of interests or competing interests.
Data Availability We have made all the datasets publicly available online (http://hdl.handle.net/10754/694767) for future assessments of other existing or new methods.

REFERENCES

Abdulah S, Alamri F, Ltaief H, Sun Y, Keyes DE, Genton MG (2022a) Data for the second competition on spatial statistics for large datasets. http://hdl.handle.net/10754/680231

Abdulah S, Alamri F, Nag P, Sun Y, Ltaief H, Keyes DE, Genton MG (2022) The second competition on spatial statistics for large datasets. J Data Sci 20:439–460

Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2018) ExaGeoStat: a high performance unified software for geostatistics on manycore systems. IEEE Trans Parallel Distrib Syst 29:2771–2784

Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2018b) Parallel approximation of the maximum likelihood estimation for the prediction of large-scale geostatistics simulations. In: 2018 IEEE international conference on cluster computing (CLUSTER), IEEE, pp. 98–108

Azzalini A (2022) The R package sn: The skew-normal and related distributions such as the skew-*t* and the SUN (version 2.1.0). Università degli Studi di Padova, Italia, home page: http://azzalini.stat.unipd.it/SN/

Bachl FE, Lindgren F, Borchers DL, Illian JB (2019) inlabru: an R package for Bayesian spatial modelling from ecological survey data. Methods Ecol Evol 10:760–766

Bandyopadhyay S, Lahiri SN, Nordman DJ (2015) A frequency domain empirical likelihood method for irregularly spaced spatial data. Ann Stat 43:519–545

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 70:825–848

Bolin D, Simas AB, Xiong Z (2023) Covariance-based rational approximations of fractional SPDEs for computationally efficient Bayesian inference. J Comput Graph Stat. https://doi.org/10.1080/10618600.2023.2231051

Burt DR, Rasmussen CE, Van Der Wilk M (2020) Convergence of sparse variational inference in Gaussian processes regression. J Mach Learn Res 21:5120–5182

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 70:209–226

Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J Am Stat Assoc 111:800–812

Eidsvik J, Shaby B, Reich B, Wheeler M, Niemi J (2014) Estimation and prediction in spatial models With block composite likelihoods. J Comput Graph Stat 23:295–315

Finley AO, Sang H, Banerjee S, Gelfand AE (2009) Improving the performance of predictive process modeling for large datasets. Comput Stat Data Anal 53:2873–2884

Fuentes M (2007) Approximate likelihood for large irregularly spaced spatial data. J Am Stat Assoc 102:321–331

Furrer R, Genton MG, Nychka D (2006) Covariance tapering for interpolation of large spatial datasets. J Comput Graph Stat 15:502–523

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378

Gramacy RB (2016) laGP: large-scale spatial modeling via local approximate Gaussian processes in R. J Stat Softw 72:1–46

Guinness J (2018) Permutation and grouping methods for sharpening Gaussian process approximations. Technometrics 60:415–429

Guinness J (2021) Gaussian process learning via Fisher scoring of Vecchia's approximation. Stat Comput 31:25

Guinness J, Katzfuss M, Fahmy Y (2021) GpGp: fast Gaussian process computation using Vecchia's approximation, R package version 0.4.0

Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, Guhaniyogi R, Gerber F, Gramacy RB, Hammerling D, Katzfuss M et al (2019) A case study competition among methods for analyzing large spatial data. J Agric Biol Environ Stat 24:398–425

Huang H, Abdulah S, Sun Y, Ltaief H, Keyes DE, Genton MG (2021) Competition on spatial statistics for large datasets. J Agric Biol Environ Stat 26:580–595

Huang H, Abdulah S, Sun Y, Ltaief H, Keyes DE, Genton MG (2021b) Data for competition on spatial statistics for large datasets. http://hdl.handle.net/10754/669153

Huang H, Blake LR, Hammerling DM (2019) Pushing the limit: a hybrid parallel implementation of the multi-resolution approximation for massive data. Technical report

Huang H-C, Cressie N, Gabrosek J (2002) Fast, resolution-consistent spatial prediction of global processes from satellite data. J Comput Graph Stat 11:63–88

Katzfuss M (2017) A multi-resolution approximation for massive spatial datasets. J Am Stat Assoc 112:201–214

Katzfuss M, Guinness J (2021) A general framework for Vecchia approximations of Gaussian processes. Stat Sci 36:124–141

Kaufman CG, Schervish MJ, Nychka DW (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. J Am Stat Assoc 103:1545–1555

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st international conference on neural information processing systems, pp. 3149-3157

Lu Z, Tjøstheim D (2014) Nonparametric estimation of probability density functions for irregularly observed spatial data. J Am Stat Assoc 109:1546–1564

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B (Stat Methodol) 71:319–392

Saas Y, Gosselin F (2014) Comparison of regression methods for spatially-autocorrelated count data on regularly- and irregularly-spaced locations. Ecography 37:476–489

Saha A, Datta A (2022) BRISC: fast inference for large spatial datasets using BRISC. R package version 1.0.5

Shi T, Cressie N (2007) Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite. Environmetrics 18:665–680

Song Y, Dai W, Genton MG (2022) Large-scale low-rank Gaussian process prediction with support points. arXiv:2207.12804

Stein ML (2014) Limitations on low rank approximations for covariance matrices of spatial data. Spat Stat 8:1–19

Sun Y, Li B, Genton MG (2012) Geostatistics for large datasets. In: Porcu E, Montero JM, Schlather M (eds) Space-time processes and challenges related to environmental problems, vol 207. Springer, Berlin, pp 55–77

Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Stat Sin 21:5–42

Vecchia AV (1988) Estimation and model identification for continuous spatial processes. J R Stat Soc Ser B (Methodol) 50:297–312

Zhang B, Sang H, Huang J (2019) Smoothed full-scale approximation of Gaussian process models for computation of large spatial datasets. Stat Sin 29:1711–1737